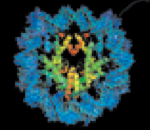


PACIFIC SYMPOSIUM ON BIOCOMPUTING 2013

The Pacific Symposium on Biocomputing (PSB) 2013 is an international, multidisciplinary conference for the presentation and discussion of current research in the theory and application of computational methods in problems of biological significance. Presentations are rigorously peer reviewed and are published in an archival proceedings volume. PSB 2013 will be held on January 3 – 7, 2013 in Kohala Coast, Hawaii. Tutorials and workshops will be offered prior to the start of the conference.

PSB 2013 will bring together top researchers from the US, the Asian Pacific nations, and around the world to exchange research results and address open issues in all aspects of computational biology. It is a forum for the presentation of work in databases, algorithms, interfaces, visualization, modeling, and other computational methods, as applied to biological problems, with emphasis on applications in data-rich areas of molecular biology.

The PSB has been designed to be responsive to the need for critical mass in sub-disciplines within biocomputing. For that reason, it is the only meeting whose sessions are defined dynamically each year in response to specific proposals. PSB sessions are organized by leaders of research in biocomputing's "hot topics." In this way, the meeting provides an early forum for serious examination of emerging methods and approaches in this rapidly changing field.

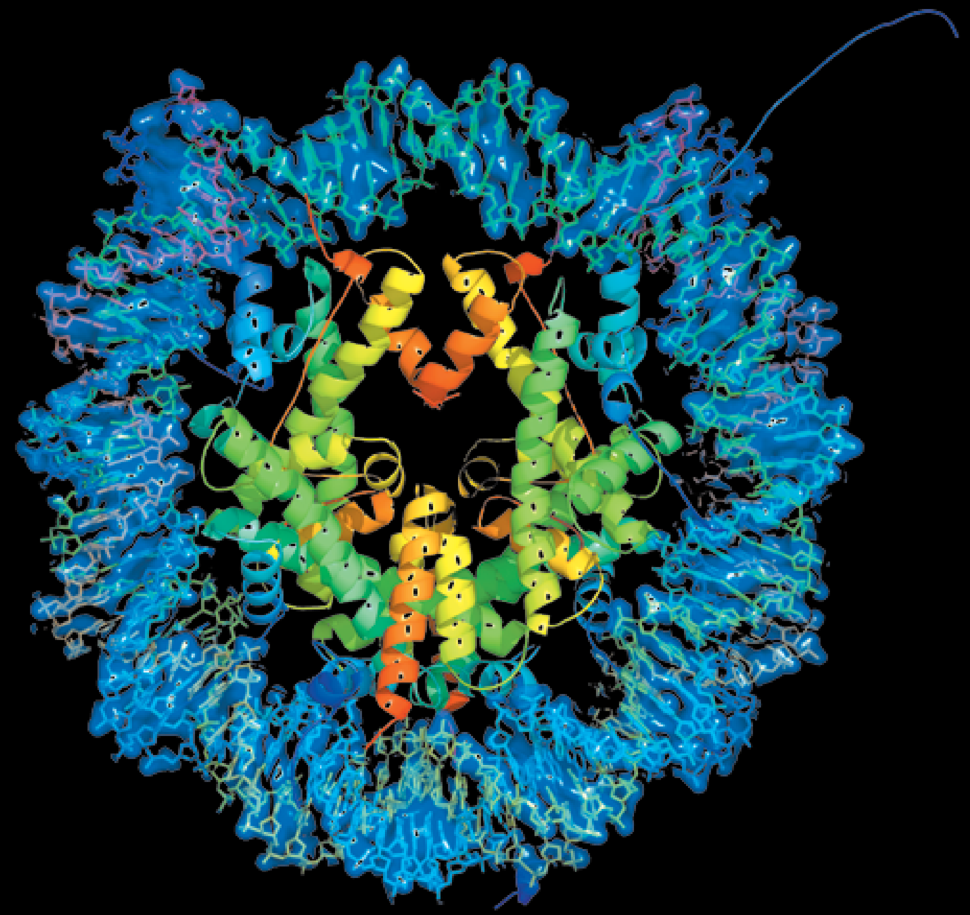
World Scientific
www.worldscientific.com
8382 eb

ISBN 978-981-4447-97-3 (eb)
9 789814 447973

R. B. Altman
A. K. Dunker
L. Hunter
T. Murray
T. E. Klein

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2013

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2013



Edited by

Russ B. Altman, A. Keith Dunker,
Lawrence Hunter, Tiffany Murray & Teri E. Klein

Cover image:
This image depicts a molecular model of the Nucleosome (PDB ID: 1aoi, Luger et al. (1997) Nature 389, 251–260) — The nucleosome is the organising principle behind higher ordered chromatin structure. The histone core of the nucleosome exemplifies the many molecular mechanisms that have evolved to regulate access to the DNA in chromatin.

Image by D. Rey Banatao,
Pacific Symposium on Biocomputing.

Preface v

COMPUTATIONAL DRUG REPOSITIONING

Session Introduction..1
Zhiyong Lu, Pankaj Agarwal, and Atul Butte

Evaluation of Analytical Methods for Connectivity Map Data..5
Jie Cheng, Qing Xie, Vinod Kumar, Mark Hurle, Johannes M. Freudenberg, Lun Yang, Pankaj Agarwal

Predictive Systems Biology Approach to Broad-Spectrum, Host-Directed Drug Target Discovery in Infectious Diseases...17
Ramon M. Felciano, Sina Bavari, Daniel R. Richards, Jean-Noel Billaud, Travis Warren, Rekha Panchal, Andreas Krämer

A Novel Multi-Modal Drug Repurposing Approach for Identification of Potent ACK1 Inhibitors..29
Sharangdhar S. Phatak, Shuxing Zhang

Protein-chemical Interaction Prediction via Kernelized Sparse Learning SVM..41
Yi Shi, Xinhua Zhang, Xiaoping Liao, Guohui Lin, Dale Schuurmans

Drug Target Predictions Based on Heterogeneous Graph Inference...53
Wenhui Wang, Sen Yang, Jing Li

EPIGENOMICS

Session Introduction..65
Alexander J. Hartemink, Manolis Kellis, William Stafford Noble, Zhiping Weng

A Powerful Statistical Method for Identifying Differentially Methylated Markers in Complex Diseases...69
Surin Ahn, Tao Wang

Using DNase Digestion Data to Accurately Identify Transcription Factor Binding Sites..........................80
Kaixuan Luo, Alexander J. Hartemink

Epigenomic Model of Cardiac Enhancers with Application to Genome Wide Association Studies..92
Avinash Das Sahu, Radhouane Aniba, Yen-Pei Christy Chang, Sridhar Hannenhalli

IDENTIFICATION OF ABERRANT PATHWAY AND NETWORK ACTIVITY FROM HIGH-THROUGHPUT DATA

Session Introduction..103
Rachel Karchin, Michael F. Ochs, Joshua M. Stuart, Trey Ideker, Joel S. Bader

*From Uncertain Protein Interaction Networks to Signaling Pathways Through Intensive Color Coding*111
Haitham Gabr, Alin Dobra, Tamer Kahveci

Identifying Master Regulators of Cancer and Their Downstream Targets by Integrating Genomic and Epigenomic Features 123
Olivier Gevaert, Sylvia Plevritis

Module Cover – A New Approach to Genotype-Phenotype Studies 135
Yoo-Ah Kim, Raheleh Salari, Stefan Wuchty, Teresa M. Przytycka

Next-Generation Analysis of Cataracts: Determining Knowledge Driven Gene-Gene Interactions Using Biofilter, and Gene-Environment Interactions Using the PhenX Toolkit 147
Sarah A. Pendergrass, Shefali S. Verma, Emily R. Holzinger, Carrie B. Moore, John Wallace, Scott M. Dudek, Wayne Huggins, Terrie Kitchner, Carol Waudby, Richard Berg, Catherine A. McCarty, Marylyn D. Ritchie

Interpreting Personal Transcriptomes: Personalized Mechanism-Scale Profiling of RNA-seq Data 159
Alan Perez-Rathke, Haiquan Li, Yves A. Lussier

PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS THERAPY

Session Introduction....................171
Oliver Stegle, Steven E. Brenner, Quaid Morris, Jennifer Listgarten

AMP: Assembly Matching Pursuit....................175
Surojit Biswas, Vladimir Jojic

Characterization of the Metabochip in Diverse Populations from the International HapMap Project in the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Project....................188
Dana C. Crawford, Robert Goodloe, Kristin Brown-Gentry, Sarah Wilson, Jamie Roberson, Niloufar B. Gillani, Marylyn D. Ritchie, Holli H. Dilks, William S. Bush

Insights into Diseases of Human Telomerase from Dynamical Modeling....................200
Samuel Coulbourn Flores, Georgeta Zemora, Christina Waldsichh

Spectral Clustering Strategies for Heterogeneous Disease Data....................212
Grace T. Huang, Kathryn I. Cunningham, Panayiotis V. Benos, Chakra S. Chennubhotla

Systematic Identification of Risk Factors for Alzheimer's Disease Through Shared Genetic Architecture and Electronic Medical Records....................224
Li Li, David Ruau, Rong Chen, Susan Weber, Atul J. Butte

A Correlated Meta-Analysis Strategy for Data Mining "Omic" Scans....................236
Michael A. Province, Ingrid B. Borecki

PHYLOGENOMICS AND POPULATION GENOMICS: MODELS, ALGORITHMS, AND ANALYTICAL TOOLS

Session Introduction....................247
Luay Nakhleh, Noah Rosenberg, Tandy Warnow

Inferring Optimal Species Trees Under Gene Duplication and Loss 250

M. S. Bayzid, Siavash Mirarab, Tandy Warnow

Evaluating Variations on the STAR Algorithm for Relative Efficiency and Sample Sizes Needed to Reconstruct Species Trees............262
James H. Degnan

The Behavior of Admixed Populations in Neighbor-Joining Inference of Population Trees............273
Naama M. Kopelman, Lewi Stone, Olivier Gascuel, Noah A. Rosenberg

Maximum Likelihood Phylogenetic Reconstruction from High-Resolution Whole-Genome Data and a Tree of 68 Eukaryotes............285
Yu Lin, Fei Hu, Jijun Tang, Bernard M.E. Moret

An Analytical Comparison of Multilocus Methods Under the Multispecies Coalescent: The Three-Taxon Case............297
Sebastien Roch

POST-NGS: INTERPRETATION AND ANALYSIS OF NEXT GENERATION SEQUENCING DATA FOR BASIC AND TRANSLATIONAL SCIENCE

Session Introduction............307
Gurkan Bebek, Mehmet Koyuturk, Thomas LaFramboise, Benjamin J. Raphael, Mark R. Chance

LSHPlace: Fast Phylogenetic Placement Using Locality-Sensitive Hashing............310
Daniel G. Brown, Jakub Truszkowski

*ChIPModule: Systematic Discovery of Transcription Factors and Their Cofactors from ChIP-seq Data*320
Jun Ding, Xiaohui Cai, Ying Wang, Haiyan Hu, Xiaoman Li

Using BioBin to Explore Rare Variant Population Stratification............332
Carrie B. Moore, John R. Wallace, Alex T. Frase, Sarah A. Pendergrass, Marylyn D. Ritchie

Detecting Highly Differentiated Copy-Number Variants from Pooled Population Sequencing............344
Daniel R. Schrider, David J. Begun, Matthew W. Hahn

MetaSeq: Privacy Preserving Meta-Analysis of Sequencing-Based Association Studies............356
Angad Pal Singh, Samreen Zafer, Itsik Pe'er

TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY

Session Introduction............368
Graciela H. Gonzalez, Kevin Bretonnel Cohen, Casey Greene, Udo Hahn, Maricel G. Kann, Robert Leaman, Nigam Shah, Jieping Ye

Enabling High-Throughput Genotype-Phenotype Associations in the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) Project as Part of the Population Architecture Using Genomics and Epidemiology (PAGE) Study............373
William S. Bush, Jonathan Boston, Sarah A. Pendergrass, Logan Dumitrescu, Robert Goodloe, Kristin Brown-Gentry, Sarah Wilson, Bob McClellan Jr., Eric Torstenson, Melissa A. Basford, Kylee L. Spencer, Marylyn D. Ritchie, Dana C. Crawford

ATHENA: A Tool for Meta-Dimensional Analysis Applied to Genotypes and Gene Expression Data to Predict HDL Cholesterol Levels 385
Emily R Holzinger, Scott M. Dudek, Alex T. Frase, Ronald M. Krauss, Marisa W. Medina, Marylyn D. Ritchie

Statistical Epistasis Networks Reduce the Computational Complexity of Searching Three-Locus Genetic Models 397
Ting Hu, Angeline S. Andrew, Margaret R. Karagas, Jason H. Moore

Evaluation of Linear Classifiers on Articles Containing Pharmacokinetic Evidence of Drug-Drug Interactions 409
Artemy Kolchinsky, Analia Lourenco, Lang Li, Luis M. Rocha

Incorporating Expert Terminology and Disease Risk Factors into Consumer Health Vocabularies 421
Michael Seedorff, Kevin J. Peterson, Laurie A. Nelsen, Cristian Cocos, Jennifer B. McCormick, Christopher G. Chute, Jyotishman Pathak

Detection of Protein Catalytic Sites in the Biomedical Literature 433
Karin Verspoor, Andrew MacKinlay, Judith D. Cohn, Michael E. Wall

WORKSHOPS

Modeling cell heterogeneity: from single-cell variations to mixed cells populations 445
Eric Batchelor, Maricel G. Kann, Teresa M. Przytycka, Benjamin J. Raphael, and Damian Wojtowicz

Computational Biology in the Cloud: Methods and New Insights from Computing at Scale 451
Peter M. Kasson

Computational Challenges of Mass Phenotyping 454
Lawrence Hunter

The Future of Genome-Based Medicine 456
Quaid Morris, Steven E. Brenner, Jennifer Listgarten, Oliver Stegle

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2013

2013 marks the 18th Pacific Symposium on Biocomputing. In addition to being published by World Scientific and indexed in PubMED, the proceedings from all previous meetings are available online at http://psb.stanford.edu/psb-online/. PSB provides sessions focusing on emerging areas in biomedical computation. These sessions are typically conceived at the previous PSB meeting as people discuss the opportunities for new sessions. Once again, we have a very exciting set of areas that build on previous sessions and introduce new topics. The excitement about "Big Data" that has received general attention in all fields is particularly relevant to biomedicine. Many of our sessions are based on the premise that there are now amazing data sets available for analysis and integration. These are offering new models for discovery. The data sets range from molecular level to cellular, organism and even population levels. In many cases, the best uses of them require broad biomedical computation skills in data integration, data mining, machine learning, and modeling. The efforts of a dedicated group of leaders has produced an outstanding set of sessions, with associated introductory tutorials. These organizers provide the scientific core of PSB and their sessions are as follows:

Computational Drug Repositioning
Zhiyong Lu, Pankaj Agarwal, and Atul Butte

Epigenomics
Alexander J. Hartemink, Manolis Kellis, William Stafford Noble, and Zhiping Weng

Identification of Aberrant Pathway and Network Activity from High-Throughput Data
Rachel Karchin, Michael F. Ochs, Joshua M. Stuart, Trey Ideker, and Joel S. Bader

Personalized Medicine: From Genotypes and Molecular Phenotypes Towards Therapy
Oliver Stegle, Steven E. Brenner, Quaid Morris, and Jennifer Listgarten

Phylogenomics and Population Genomics: Models, Algorithms, and Analytical Tools
Luay Nakhleh, Noah Rosenberg, and Tandy Warnow

Post-NGS: Interpretation and Analysis of Next Generation Sequencing Data for Basic and Translational Science
Gurkan Bebek, Mehmet Koyuturk, Thomas LaFramboise, Benjamin J. Raphael, and Mark R. Chance

Text and Data Mining for Biomedical Discovery
Graciela H. Gonzalez, Kevin Bretonnel Cohen, Casey Greene, Udo Hahn, Maricel G. Kann, Robert Leaman, Nigam Shah, Jieping Yie

We are also pleased to present four workshops in which investigators with a common interest come together to exchange results and new ideas in a format that is more informal than the peer-reviewed sessions. For this year, the workshops and their organizers are:

Modeling cell heterogeneity: from single-cell variations to mixed cells populations
Eric Batchelor, Maricel G. Kann, Teresa M. Przytycka, Benjamin J. Raphael, and Damian Wojtowicz

Computational Biology in the Cloud: Methods and New Insights from Computing at Scale
Peter M. Kasson

Computational Challenges of Mass Phenotyping
Lawrence Hunter

The Future of Genome-Based Medicine
Quaid Morris, Steven E. Brenner, Jennifer Listgarten, and Oliver Stegle

We thank our keynote speakers, Dan Roden (Science Keynote) and David Ewing Duncan (Ethical, Legal and Social Implications Keynote).

Tiffany Murray manages the peer review process and assembly of the proceedings, and also plays a key role in many other aspects of the meeting. We also thank the National Institutes of Health and the International Society for Computational Biology (ISCB) for travel grant support. We are particularly grateful to the onsite PSB staff Al Conde, Brant Hansen, Georgia Hansen, BJ McKay-Morrison, Jackson Miller, Kasey Miller, and Paul Murray for their assistance. We also acknowledge the many busy researchers who reviewed the submitted manuscripts on a very tight schedule. The partial list following this preface does not include many who wished to remain anonymous, and of course we apologize to any who may have been left out by mistake.

We look forward to a great meeting once again.

Aloha!

Pacific Symposium on Biocomputing Co-Chairs,
October 5, 2012

Russ B. Altman
Departments of Bioengineering, Genetics & Medicine, Stanford University

A. Keith Dunker
Department of Biochemistry and Molecular Biology, Indiana University School of Medicine

Lawrence Hunter
Department of Pharmacology, University of Colorado Health Sciences Center

Teri E. Klein
Department of Genetics, Stanford University

Thanks to the reviewers…

Finally, we wish to thank the scores of reviewers. PSB aims for every paper in this volume be reviewed by three independent referees. Since there is a large volume of submitted papers, paper reviews require a great deal of work from many people. We are grateful to all of you listed below and to anyone whose name we may have accidentally omitted or who wished to remain anonymous.

Uri David Akavia
Elizabeth Allman
Julia Antonio
Lars Arvestad
Mukul Bansal
Gill Bejerano
Andreas Beyer
Olivier Bodenreider
Richard Bonneau
Karsten Borgwardt
Pascal Braun
Steven Brenner
Yana Bromberg
Soren Brunak
David Bryant
Michael Cantor
Ren-Hua Chung
Melissa Cline
Adrien Coulet
Edna Cukierman
Phuong Dao
Z. John Daye
Michael DeGiorgio
James Degnan
Joel Dudley
Daniel Eisinger
Guy Fernald
Juliane Fluck
Iddo Friedberg
Nicolò Fusi
Julien Gagneur
Don Geman
Assaf Gottlieb
Danielle Greenwalt
Mark Holder
Antti Honkela
Fereydoun Hormozdiari
Rui Hu
Carol Huang
Hailiang Huang
Daniel Hyduke
Phil Jaeger
Ethan Jewett
Nebojsa Jojic
Purvesh Khatri
Ju Han Kim
Carl Kingsford
Jacob Koehler
Laura Kubatko
Jens Lagergren
Mathieu Lajoie
Christina Leslie
Xin Li
Yun Li
Tien-Ho Lin
George Linderman
Christoph Lippert
Tianyun Liu
Yu Liu
Yves Lussier
Jian Ma
Avi Ma'ayan
John Marioni
Erick Matsen
Sean Maxwell
Purvi Mehta
Sean Mooney
Alex Morgan
Sara Mostafavi
Joanna Mountain
Leopold Parts
Chirag Patel
Evan Paull
Aris Persidis
Laila Poisson
Gerald Quon
Matthew Rasmussen
Gunnar Rätsch
Magnus Rattray
Susanna Repo
Fabio Rinaldi
Tom Rindflesch
Phoebe Roberts
Sebastien Roch
Eitan Rubin
Matthew Ruffalo
Andrey Rzhetsky
Julio Saez-Rodriguez
David Sankoff
Philippe Sanseau
Neil Sarkar
Michale Schroeder
Roded Sharan
Ronglai Shen
Ilya Shmulevich
Artem Sokolov
Mike Steel
Gary Stormo
Yun Tang
Cuong Than
William Thompson
Caroline Uhler
Fabio Vandin
Bjarni Vilhjalmsson
Peter Waltman
John Wilbur
Yufeng Wu
Gokhan Yavas
Roni Zeiger
Jin Zhang
Louxin Zhang
Derrick Zwickl

COMPUTATIONAL DRUG REPOSITIONING

ZHIYONG LU[†]
National Center for Biotechnology Information (NCBI)
Bethesda, MD, 20894 USA
Email: zhiyong.lu@nih.gov

PANKAJ AGARWAL
Systematic Drug Repositioning, Computational Biology, GlaxoSmithKline Pharmaceuticals R&D
King of Prussia, PA 19406, USA
Email: pankaj.agarwal@gsk.com

ATUL J. BUTTE[¶]
Division of Systems Medicine, Stanford University School of Medicine
Stanford, CA 94305, USA
Email: abutte@stanford.edu

Despite increasing investments in pharmaceutical R&D, there is a continuing paucity of new drug approvals. Drug discovery continues to be a lengthy and resource-consuming process in spite of all the advances in genomics, life sciences, and technology. Indeed, it is estimated that about 90% of the drugs fail during development in phase 1 clinical trials[1] and that it takes billions of dollars in investment and an average of 15 years to bring a new drug to the market[2].

Meanwhile, there is an ever-growing effort to apply computational power to improve the effectiveness and efficiency of drug discovery. Traditional computational methods in drug discovery were focused on understanding which proteins could make good drug targets, sequence analysis, modeling drugs binding to proteins, and the analysis of biological data. With the attention on translational research in recent years, a new set of computational methods are being developed which examine drug-target associations and drug off-target effects through system and network approaches. These new approaches take advantage of the unprecedented large-scale high-throughput measurements, such as drug chemical structures and screens[3, 4], side effect profiles[5, 6], transcriptional responses after drug treatment[7, 8], genome wide association studies[9], and combined knowledge[10, 11]. More importantly there are increasing reports of these findings being validated in experimental models[5, 7, 12], thus clarifying the value proposition for computational drug discovery. As a result, now is an exciting time for computational scientists to gain evidence for reusing an existing drug for a different use or generate testable hypotheses for further screening[13].

Despite the progress, there is clearly room for technical improvement with regard to computational repurposing approaches. Furthermore, to materialize the true potential and impact of these

[†] Work supported by NIH Intramural Research Program, National Library of Medicine
[¶] Work supported by National Institute for General Medical Science (R01 GM079719).

methods, much work is needed to show that they can be successfully adopted into practical applications. Hence, the aim of our session is to provide a forum to bring together the research community for a serious examination of these important issues. The five papers accepted to the session represent the breadth of research interests in the field: a graph-based inference method for predicting drug targets, a machine-learning algorithm for predicting protein-chemical interaction, an integrated method for identifying drug candidates against a novel cancer target, a knowledge-based method for target identification against infectious agents, and a systematic evaluation of similarity measures in the use of connectivity map data for drug repurposing.

Wang et al. propose a novel computational method for target prediction, known as heterogeneous graph based inference (HGBI). HGBI integrates drug-drug similarities, target-target similarities, and drug-target interactions into a heterogeneous graph. They model the drug-target interactions as the stabilized information flow problem across the heterogeneous graph. Cross-validation results show that HGBI significantly outperforms the state of the art in predicting novel targets for drugs. Furthermore, using a case study, the authors show that in practice HGBI can be used to rank candidate drug targets and that top-ranked results may be worth further experimental screening.

Shi et al. present a different approach for predicting target-drug interactions, where target-target similarities are often first obtained using the primary amino acid sequences. In order to do so, unlike the existing methods that generally rely on measuring the maximum local similarity between two protein sequences, the authors propose a novel sparse learning method that considers sets of key short peptides shared by proteins interacting with the same drug. Their method integrates feature selection, multi-instance learning, and Gaussian kernelization into an L1 norm support vector machine classifier. According to their experimental results, their approach can not only outperform the previous methods, but also reveals an optimal subset of potential binding regions.

Phatak and Zhang propose a computational pipeline for identifying novel drug candidates through integrating separate results from structure-based virtual screening, chemical-genomic similarity search, and graph-based similarity search. To demonstrate its feasibility in practical use, the authors report the repurposing of existing drugs against a novel cancer target ACK1, which is significantly overexpressed in breast cancer and prostate cancer patients. They screened 1,447 marketed drugs, and merged complementary hits from different methods to select ten drugs for experimental testing. They found four of these drugs to be potent ACK1 inhibitors. Interestingly, Dasatinib, one of the final four drugs they discovered computationally was also recently found effective on inhibiting ACK1-related prostate cancer progression in a separate experimental study.

Felciano et al. identified novel drug targets against six different pathogens such as Ebola and Marburg virus. Using knowledge of the immune system and host-pathogen pathways, their method automatically generates a list of potential target proteins that may have a beneficial therapeutic effect against at least two of the six pathogens. Then, the candidate targets in the list are reviewed and prioritized for being further validated in vitro and in vivo experiments. Next, experimental results are normalized such that target validation could be compared across targets and pathogens examined in their study. Finally, based on their analysis, 34% of their predicted targets are shown to be promising in mouse models. Their work demonstrates the potential for knowledge-based methods in host-directed drug target discovery.

Cheng et al. present a systematic evaluation on different similarity measures used in methods that aim to identify related transcriptional profiles based on connectivity map data. Using the drug compounds with shared Anatomical Therapeutic Chemical (ATC) classification as the gold standard, they compare four different measures for the identification of similar drug pairs and find that their proposed Xtreme cosine similarity score achieves the highest accuracy. Moreover, their benchmark experiments show that smaller gene signatures outperform larger ones. They also find that good transcriptional response to drug treatment is necessary but not sufficient to achieve high AUCs.

This is the first year Computational Drug Repositioning has been offered as a track at the Pacific Symposium on Biocomputing, and we are pleased with the results of our call for participation. Given the interest seen here, new meetings being proposed just focused on drug repositioning. The National Center for Advancing Translational Sciences (NCATS) is partnering with pharmaceutical companies to offer funding for repositioning. The future seems quite bright for investigators conducting research in this field.

Acknowledgments

The session co-chairs are grateful to the numerous reviewers for their help in selecting the best papers among many excellent submissions.

References

1. A. Krantz, *Nat Biotechnol*, **13**, 1294, (1998)
2. C. P. Adams and V. V. Brantner, *Health Aff (Millwood)*, **2**, 420, (2006)
3. M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth, *Nature*, **7270**, 175, (2009)
4. S. J. Swamidass, *Brief Bioinform*, **4**, 327, (2011)
5. M. Campillos, M. Kuhn, A. C. Gavin, L. J. Jensen and P. Bork, *Science*, **5886**, 263, (2008)
6. L. Yang and P. Agarwal, *PLoS One*, **12**, e28025, (2011)
7. M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage and A. J. Butte, *Sci Transl Med*, **96**, 96ra77, (2011)
8. G. Hu and P. Agarwal, *PLoS One*, **8**, e6536, (2009)

9. P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon and V. Mooser, *Nat Biotechnol*, **4**, 317, (2012)
10. J. Li and Z. Lu, *Proceedings (IEEE Int Conf Bioinformatics Biomed)*, (2012)
11. A. Gottlieb, G. Y. Stein, E. Ruppin and R. Sharan, *Mol Syst Biol*, 496, (2011)
12. J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha and A. J. Butte, *Sci Transl Med*, **96**, 96ra76, (2011)
13. P. Sanseau and J. Koehler, *Brief Bioinform*, **4**, 301, (2011)

EVALUATION OF ANALYTICAL METHODS FOR CONNECTIVITY MAP DATA

JIE CHENG[1], QING XIE[2], VINOD KUMAR[2], MARK HURLE[2], JOHANNES M. FREUDENBERG[3], LUN YANG[2], PANKAJ AGARWAL[2]

[1]Statistical and Platform Technologies, GlaxoSmithKline R&D, UP4335, 1250 S Collegeville Rd, Collegeville, PA 19426; [2]Computational Biology, GlaxoSmithKline R&D, UW2230, 709 Swedeland Road, King of Prussia, PA 19406, USA; [3]Computational Biology, GlaxoSmithKline R&D, 3.2085A, 5 Moore Drive, Durham, NC, 27709, USA

Email: Jie.Cheng@gsk.com, Qing.Xie@gsk.com, Vinod.D.Kumar@gsk.com, Mark.R.Hurle@gsk.com, Johannes.M.Freudenberg@gsk.com, Lun.Y.Yang@gsk.com, Pankaj.Agarwal@gsk.com

Connectivity map data and associated methodologies have become a valuable tool in understanding drug mechanism of action (MOA) and discovering new indications for drugs. However, few systematic evaluations have been done to assess the accuracy of these methodologies. One of the difficulties has been the lack of benchmarking data sets. Iskar et al. (PLoS. Comput. Biol. **6**, 2010) predicted the Anatomical Therapeutic Chemical (ATC) drug classification based on drug-induced gene expression profile similarity (DIPS), and quantified the accuracy of their method by computing the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curve. We adopt the same data and extend the methodology, by using a simpler eXtreme cosine (XCos) method, and find it does better in this limited setting than the Kolmogorov-Smirnov (KS) statistic. In fact, for partial AUC (a more relevant statistic for actual application to repositioning) XCos does 17% better than the DIPS method (p=1.2e-7). We also observe that smaller gene signatures (with 100 probes) do better than larger ones (with 500 probes), and that DMSO controls from within the same batch obviate the need for mean centering. As expected there is heterogeneity in the prediction accuracy amongst the various ATC codes. We find that good transcriptional response to drug treatment appears necessary but not sufficient to achieve high AUCs. Certain ATC codes, such as those corresponding to corticosteroids, had much higher AUCs possibly due to strong transcriptional responses and consistency in MOA.

1. Introduction

Identifying the correct disease indication for a drug is an important problem and several computational methods have been described [1]. The problem for any practitioner, however, is to assess the precision of these methods. The desired method should provide relatively high confidence that the first few indications that are predicted for a drug contain at least one that will be validated in clinical trials and make a positive impact on patients. One of the most important techniques in the space of drug repositioning is connectivity map (CMAP) [2].

A key contribution of CMAP has been the establishment of a database of cellular expression profiles in response to drug treatment in cell lines such as MCF7. This has enabled both the discovery of drug MOA and new indications [2,3]. Several CMAP hypotheses suggesting potential therapeutic compounds for new disease indications have been experimentally validated [4-8].

However, despite numerous impressive anecdotal validations, it has proven challenging to quantitatively estimate the accuracy of this technique. A gold standard data set still eludes us in terms of drugs that impact a disease positively. Thus research has turned to the cleaner benchmarking data sets to predict drug relationships. This is with the implicit hope that methods that better predict drug classes will also do better at predicting disease indications for drugs. A useful classification is the Anatomical Therapeutic Chemical (ATC) system, which codes divides drugs into different groups in a hierarchical fashion according to the organ or system on which they act and their therapeutic and chemical characteristics. The ATC level 4 is mostly based on common MOA, and thus has proven useful as a benchmark for comparing similar drugs.

The initial CMAP approach utilized a nonparametric, rank-based Kolmogorov-Smirnov (KS) statistic to connect disease gene expression signatures to drug expression profiles. KS scores are generated based on the location of the genes in the signature (i.e. up and down lists) within the entire ordered list of gene expression changes in response to compound treatment. The disease signatures often come from public repositories of expression profiles, such as Gene Expression Omnibus (GEO) [9]. Compounds from the reference dataset can also be connected with each other using the same type of computation to evaluate the similarity between them.

Iskar et al [10] provided one of the first quantitative evaluations of CMAP methods. They applied a centered mean normalization approach to preprocess the intensity data in order to eliminate batch-specific effects. The pair-wise **d**rug-**i**nduced gene expression **p**rofile **s**imilarity (DIPS) scores between each pair of drugs in CMAP were then calculated using a method similar to inverse total enrichment score (TES) by Iorio et al [11]. (TES itself is modification of KS.) They used compounds with high chemical similarities, and compounds with shared ATC classification as true positives for their benchmarking. They computed the area under the receiver operator characteristic (ROC) curve (AUC0.1) to measure differences at a low false positive rate (FPR=0.1). This emphasizes early retrieval which is important because for repositioning we are willing to sacrifice some true positives to keep false positives low. The performance of DIPS was shown to be superior to the compound vs. biological control comparison method described by Iorio et al.

In addition to modifications of CMAP data processing workflows, many groups have investigated alternatives to the KS statistic. More recently researchers have extended methods based on Spearman’s correlation (EPSA) [12], Fisher’s Exact test (EXALT) [13], Wilcoxon rank-sum test (openSESAME) [14], weighted Pearson correlation[15], logistic regression (LRpath) [16], probabilistic categorization (ProbCD) [17], empirical background p-values[18], random set statistic (GRS) [19]. and partially ranked data[20]. In this paper, we explore an eXtreme Cosine method that truncates the middle of the two expression profiles being compared. This focuses attention on true outliers in both treatments. The cosine is an inner product of two vectors much like Pearson correlation, which has been shown to be superior to GSEA [18].

In this study, we use the ATC classification as the benchmark to compare the eXtreme cosine method (XCos) to other CMAP scoring methods, data processing methods, and signature sizes. Insights from these comparisons will clarify parameter choices, which can then be used in drug repositioning where gold standard benchmarking datasets are more complicated. We score each method using AUC in the early (FPR=0.1 and FPR=0.01) discovery phase. This allows us to determine which compound classes contain robust expression profiles in CMAP data, and which analytical approaches are more accurate at least in this evaluation.

2. Methods

2.1. *Data sources and data processing*

Small-molecule perturbed genome-wide transcriptional response data were downloaded from the Connectivity Map (CMAP, build 02, http://www.broadinstitute.org/CMAP/). These data comprises of 6,100 gene expression instances (treatment vs. vehicle control pairs) from primarily three human cultured cell lines (MCF7, PC3, and HL60) treated with 1,309 bioactive small chemical molecules at varying concentrations. Each instance denotes a treatment and control pair for one small molecule. Each instance has attributes such as perturbagen name, concentration, cell line and batch etc.

Two methods of pre-processing probe level intensities are considered in this paper:

a) **MC**: **M**ean **C**entering CMAP data was obtained directly from P. Bork [personal communication]. The data was generated using the method described by Iskar et al.[10]. Briefly, each compound treatment arrays were grouped based on the cell line and normalized separately using RMA [21]. Vehicle controls from CMAP were discarded and for each batch individual probes for each treatment were mean centered to calculate the average difference values within the batch. The final data consists of 4,849 treatment instances from three cell lines corresponding to 1,144 small molecules.

b) **B**atch **D**MSO **C**ontrol (**BDC**): Using controls from within the batch was proposed in the original CMAP paper [2], and we wanted to directly compare MC to it. Probe level data (CEL files) from CMAP was processed using Array Studio (Omicsoft Corporation, Research Triangle Park, NC, USA). Briefly, microarray datasets were grouped based on the cell line. For each microarray dataset, the probe set intensities were normalized using RMA. Next, all scaled probe sets with values less than primary threshold values (set to 64) for all treatments and control samples was set to that threshold value. The intensity values for each probe set are then $\log_2$ transformed. Finally, the $\log_2$ intensities of each probe set from all vehicle control samples within the same batch and cell line are averaged and subtracted from the treatment sample to generate the corresponding treatment-to-control values and this is termed BDC. We filtered the 6,100 instances to the same 4,849 for MC to make results comparable.

We averaged multiple instances for each compound within a cell line and then across cell lines.

The ATC codes were obtained from Iorio et al. [11] and then supplemented by additional annotation.

2.2. *Pair-wise similarity scores*

We used four methods: KS, TES, DIPS, and XCos to compute similarities between drug pairs.

KS: The initial CMap approach utilized a nonparametric, rank-based Kolmogorov-Smirnov (KS) statistic [2].

TES (inverse total enrichment score) is a measure based on the KS statistic as described in Iorio et al.[11]. A key difference is that this does not require the up and down signature to have consistent direction of scores compared to KS.

DIPS: Uses TES on mean centered (MC) data and we used the data as provided (personal communication, P. Bork).

XCos: The Xtreme cosine similarity score is calculated by retaining only the Xtreme probes for each instance after sorting by decreasing fold-change, i.e., only keeping the top N and bottom N probe sets and setting all other probe sets to zero. The cosine similarity between two Xtreme instances can then be calculated as a dot product of the two vectors. This is a variation of a described method [22]. Cosine similarity is much like Pearson correlation except that the vectors are not centered around their individual means. Unlike Euclidian distance, both cosine similarity and Pearson correlation are scale independent and should be more robust for our purpose.

Pair-wise similarity scores of compounds for each of the three cell lines are generated separately and then combined. Similarities between instances of the same compound are excluded and not included in any of the plots.

2.3. *Method nomenclature*

Eight of the nine methods described in this paper follow the SIM_PROC_SIZE nomenclature. SIM describes the similarity method which is one of KS, TES, or XCos (see section 2.2); PROC describes the data processing method which is either MC or BDC (see section 2.1) and the SIZE is the size of the signature which is either 100 or 500. The KS and TES methods were only evaluated with MC (and not with BDC), thus we have 8 total methods. DIPS is the ninth method as described in Iskar et al. [10]. DIPS is most closely related to TES_MC_500 though DIPS uses a sort order based on detection calls, while our implementation of MC uses a sort order based on fold changes. Moreover, DIPS used only a single ATC for each drug while we used all ATC codes for a drug.

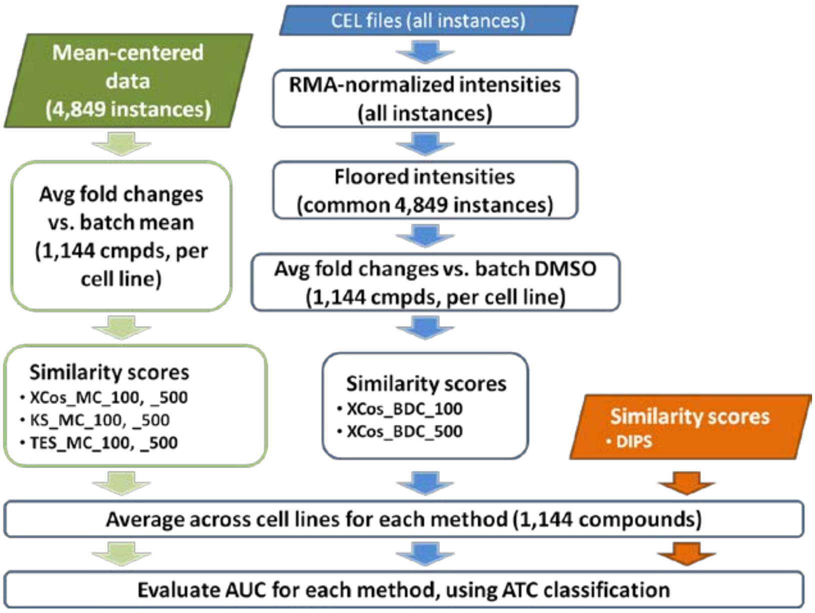

CEL files (all instances)
Mean-centered data (4,849 instances)
RMA-normalized intensities (all instances)
Floored intensities (common 4,849 instances)
Avg fold changes vs. batch mean (1,144 cmpds, per cell line)
Avg fold changes vs. batch DMSO (1,144 cmpds, per cell line)
Similarity scores
• XCos_MC_100, _500
• KS_MC_100, _500
• TES_MC_100, _500
Similarity scores
• XCos_BDC_100
• XCos_BDC_500
Similarity scores
• DIPS
Average across cell lines for each method (1,144 compounds)
Evaluate AUC for each method, using ATC classification

Figure 1. A schematic of the analytical workflow used to generate the AUC. Parallelograms indicate data acquired. The nine measures of similarity scores listed in the three similarity score rectangles were evaluated on the ATC codes.

2.4. *AUCs and p-values*

Pair-wise similarity scores are evaluated using individual ATC codes at different levels as well as using ATC levels from 1 to 4 for each of the nine methods as listed in Figure 1.

For calculating AUC of a particular ATC level, the positive cases are distinct compound pairs that share any ATC code at this level. All other pairs are considered negative cases. These criteria are effective in handling drugs that have multiple ATC codes. The ROCs in Figure 2 and Table 1 use this method as they count matches across ATC level 4 as positives.

For calculating AUC for a specific ATC code, the only relevant pairs are those have at least one compound with this ATC code. The positive cases are defined as distinct compound pairs that both share this ATC code. The negative cases are the compound pairs with only one compound belonging to this ATC code. Thus, if neither compound of a pair share this ATC code, the pair is excluded from the AUC calculation for this ATC code. Figure 3 uses this as the standard as AUCs are shown for individual ATC codes.

The p value calculation for comparing "paired" partial AUC is based on a bootstrap test [23]. $\boldsymbol{Z}$ is defined as $\boldsymbol{(pAUC_1\text{-}pAUC_2)/sd(pAUC_1\text{-}pAUC_2)}$, where $\boldsymbol{pAUC_1}$ and $\boldsymbol{pAUC_2}$ are the two paired partial AUCs to be compared and the ***sd(pAUC$_1$-pAUC$_2$)*** is the standard deviation of the difference between ***pAUC$_1$*** and ***pAUC$_2$***. The standard deviation of the difference between the two AUCs is estimated from the 1,000 bootstrap runs.

2.5. *Expression signal strength*

The expression signal strength (ESS) is defined as the sum of the absolute values of the $\log_2$ of the fold changes of the top and bottom N features (or probesets) of a gene expression profile. We first calculated the ESS of every compound expression profile. The ESS values of the same compound were then averaged within a cell line, and then these were averaged across the three cell lines to generate one ESS value per compound. The ESS for a particular ATC code is calculated by averaging all ESS values of the compounds that belongs to this ATC code. Figure 3 plots ESS on the x-axis with N=50.

3. Results

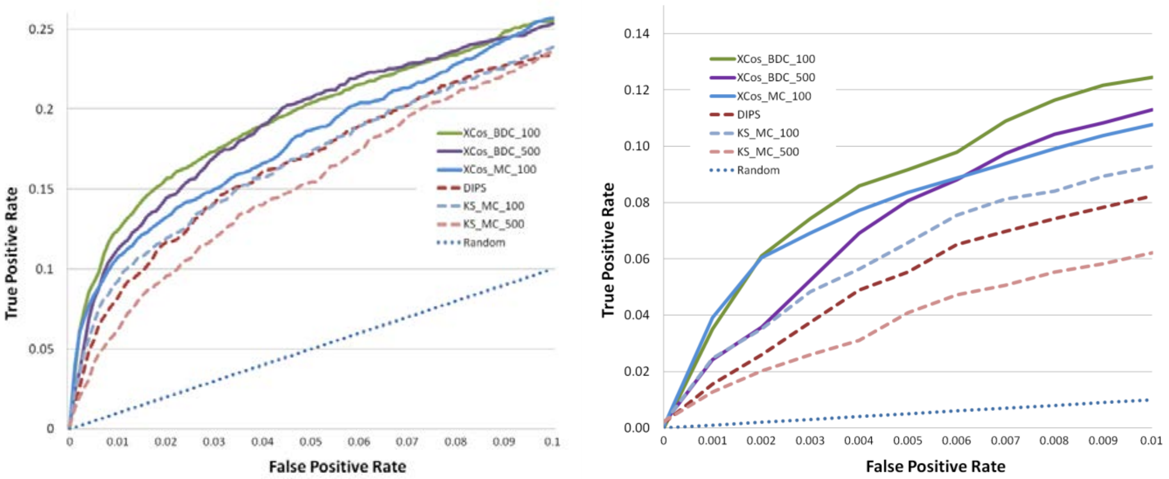
Assessment of methods on 4th level ATC codes

An earlier study showed that DIPS method leads to fewer false positives when compared using a partial AUC value at FPR=0.1 (AUC0.1) counting every pair of drugs which had at least one matching ATC 4th level code as a true positive [10]. Also the AUC0.1 was higher with mean centering (MC) than without mean centering. In this study, we systematically evaluated multiple scoring methods using the same data processing method and AUC measurement. We also suggest and evaluate the performance of the XCos similarity for the expression vectors of pairs of drugs using the top and bottom differentially expressed probes.

XCos_BDC_100 performed best in terms of AUC at FDR=0.1 (see Figure 2 and Table 1). The AUC was 0.***0193*** and significantly better than the DIPS AUC of 0.016 (two tailed p = 1.8e-7). The difference between XCos and DIPS is even larger and more significant at FPR=0.01 (p<1e-13). This may suggest that for early discovery consistent with drug repositioning the XCos with smaller signatures might indeed be better. There are three obvious differences between these two (XCos_BDC_100 and DIPS) methods: *A*) the batch DMSO control (BDC) vs. mean centering (MC), *B*) the size of the signature: 100 vs. 500, and *C*) the method itself: XCos vs. TES. To understand this further, we isolated these three differences.

A. XCos_BDC_100 had higher AUC0.1 compared with XCos_MC_100 (p=5e-4), thus at least for the XCos method, the batch-based DMSO controls are better than mean centering.
B. The AUC for XCos_BDC_100 is higher than for XCos_BDC_500, but not significant statistically (p=0.26). However, the AUC difference for KS_MC_100 compared to KS_MC_500 is significant (p=6e-6), thus at least for KS_MC the smaller signatures are better in this comparison.

C. In terms of method itself, XCos outperformed KS (p=0.008) with mean centering and 100 probe signatures.



(a) (b)

Figure 2: Comparison of the different scoring, data processing methods and signature sizes. Drugs with at least one matching ATC 4th level code are counted as true positives. The two TES scores track KS quite closely so are not shown for clarity. **a**) AUC0.1: Partial ROC curve at the FPR = 0.1**. b**) AUC0.01: Partial ROC curve at the FPR = 0.01**.**

Table 1: Partial AUCs from multiple scoring methods. Drugs with at least one matching ATC 4th level code are counted as true positives.

Method	**AUC0.1: Partial AUC @FPR=0.1**	**AUC0.01: Partial AUC @FPR=0.01**
KS_MC_100	0.01655	6.06e-4
KS_MC_500	0.01503	3.79e-4
TES_MC_100	0.01663	6.12e-4
TES_MC_500	0.01484	3.82e-4
XCos_MC_100	0.01789	7.73e-4
XCos_MC_500	0.01738	6.84e-4
XCos_BDC_100	**0.01926**	**8.56e-4**
XCos_BDC_500	0.01898	7.20e-4
DIPS	0.01642	5.14e-4

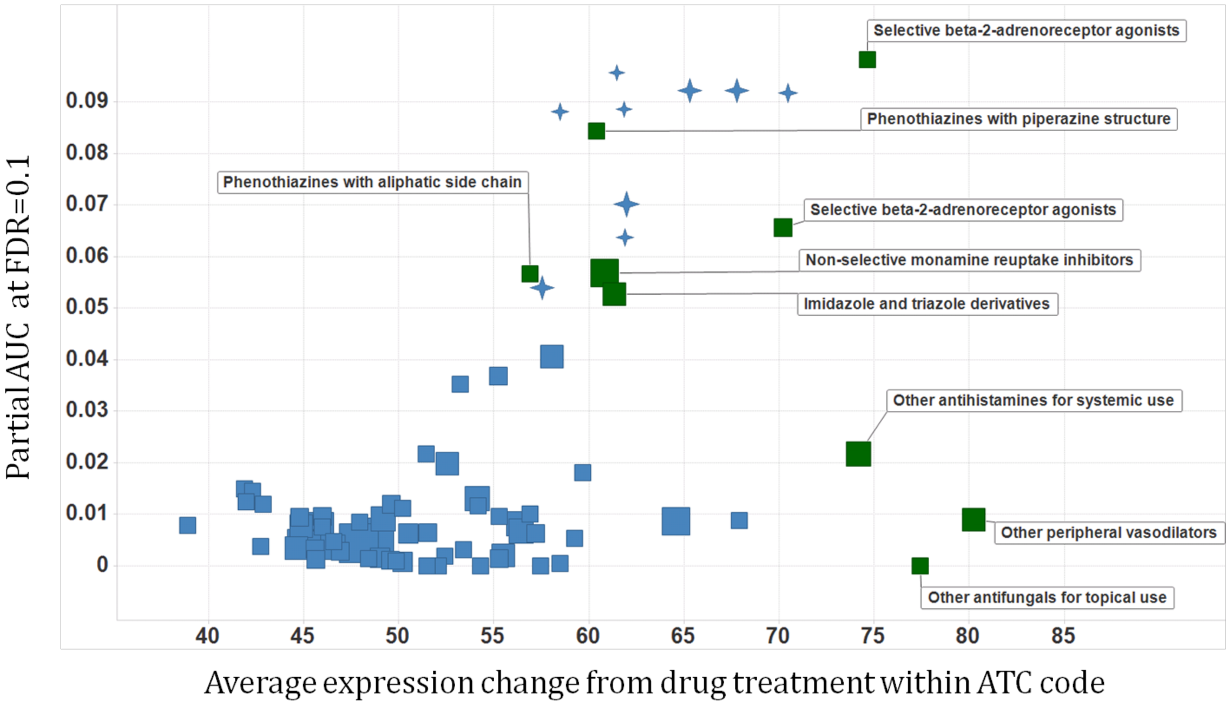


Figure 3. Relationship between AUC0.1 (for XCos_BDC_100) and the average expression change from drug treatment within an ATC level code. ATC codes which primarily describe corticosteroids are indicated by crosses, all other ATC codes are shown as rectangles. Descriptions are provided for ATC codes of interest shown in green rectangles. Points are sized by the number of compounds in the ATC code. All ATC level 4 codes with at least 5 compounds are shown. If all 100 probesets had a uniform absolute fold change of 1.414, it would correspond to an expression level of 50 on the x-axis.

All the p-values were computed as described in the methods. In fact, from Figure 2a and Table 1 the trends mentioned above are quite apparent as well and the AUC0.1 for XCos_BDC_100 is statistically significantly different from the AUC0.1 for all the MC methods in Figure 2a. The above trend in terms of AUC0.1 comparisons on different methods could not be observed on the overall AUCs (data not shown). For overall AUCs, we observed that mean centering outperforms batch-based DMSO controls at least for the XCos method. We also noticed that TES is quite similar to KS and thus not shown in Figure 2 for readability.

Differences amongst ATC codes

The specific ATC codes at level 4 compared to the generic ATC level 1 codes provide more accurate classifiers; in fact, the classification at ATC level 1 is quite close to random (data not shown). Figure 3 displays the heterogeneity in the AUC measures for ATC level 4 codes using XCos_BDC_100. The ATC codes with the strongest signal are dominated by corticosteroids, β2-adrenoreceptor agonists, and phenothiazines. We note a large number of related corticosteroid-related ATC codes with high AUC0.1. On investigation, these are compounds with same MOA but grouped into different ATC codes based on strength, anatomy, and formulation (inhaled, oral or topical).

This figure also shows the dependence of the AUC0.1 on the average change in expression due to compound treatment for a given ATC class. It seems intuitively obvious that if the expression change is low, the analytical methods cannot detect similarity. In addition, we observed that the poorly detected ATC codes with high expression changes (those labeled as starting with “Other”) are often collections of miscellaneous compounds that are unlikely to have common MOAs.

4. Discussion

Numerous methods have been proposed to identify related transcriptional profiles for CMAP readouts. They differ mostly by the underlying similarity measure, some of which are quite simple and have been known for decades, while other, more complex methods rely on powerful computing. Surprisingly, the XCos similarity score, which simply measures the cosine of two signatures, outperforms the standard, Kolmogorov-Smirnov (KS)-based CMAP method (Figure 2). Furthermore, the similarity between related signatures appears to be driven by the genes that change the most between treatment and control. Both XCos and KS scoring methods based on the top 100 features more accurately predicted ATC codes than the ones based on the top 500 features. Of course, both these signature sizes are arbitrary and the optimal signature size should be further explored. Flexible signature sizes, however, have also been explored recently [24]. Finally, the preprocessing method used to compute the signatures plays a significant role as well. We find that mean centering does not improve the similarity scores in comparison to batch based DMSO controls – at least for the XCos method. This contrasts with the earlier results[10], and the reason is not evident; however, possible explanations include our not using probeset detection calls and DIPS comparison not using batch-matched DMSO controls. Moreover, we did not restrict a drug to have exactly one ATC code as required by DIPS [10].

It should be noted that these conclusions should be considered preliminary as they are limited by the use of ATC codes as a “gold standard”. Multiple ATC codes per compound can lead to errors and redundant ATC codes may inflate AUCs. Furthermore, many ATC codes do not properly characterize MOAs (e.g. “other peripheral vasodilators”, Figure 3).

Another limitation may be that the averaging over multiple cell lines averages biological variation for compounds that may have differential responses in the three cell lines. On the other hand, using all available data may lead to more “stable” compound-specific signatures.

Future work should explore additional accuracy measures, as even AUC0.1 and AUC0.01 have too many false positives to be useful in terms of number of hypotheses that can be experimentally validated. It should also compare more methods and isolate the impact of each parameter completely across multiple methods. As indicated in Figure 3, some ATC codes lead to high AUC numbers regardless of the method used i.e. some drug classes are really easy to find with expression profiles. To ensure that such high performing ATC codes do not skew the overall comparison, future work should include a comparison of methodologies focusing only on the more “difficult” ATC codes.

A key challenge for drug repositioning is to develop a gold standard benchmarking data set that will not necessitate the extrapolation of results from drug MOA. With some expert curatorial effort FDA approved indications could be mapped to a disease ontology. However, it is not evident as to what constitutes matching disease signatures as we would also need to determine which of those drugs are disease modifying as opposed to those providing symptomatic relief and not expected to match as true positives. We believe that quantitative assessment of repositioning methodologies is a must, if computational biology is to make a more compelling case for its utility in this field.

5. Acknowledgments

We would like to thank Drs Aaron Spivak, Hui Huang, Kwan Lee, and Philippe Sanseau for feedback and comments.

References

1. J. T. Dudley, T. Deshpande, and A. J. Butte, "Exploiting drug-disease relationships for computational drug repositioning," Brief. Bioinform. **12**, 303-311 (2011).

2. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," Science **313**, 1929-1935 (2006).

3. X.A.Qu and Rajpal D.K. Applications of Connectivity Map in Drug Discovery and Development. Drug Discov Today . 2012.

4. M. Chang, S. Smith, A. Thorpe, M. J. Barratt, and F. Karim, "Evaluation of phenoxybenzamine in the CFA model of pain following gene expression studies and connectivity mapping," Mol. Pain **6**, 56 (2010).

5. S. Claerhout, J. Y. Lim, W. Choi, Y. Y. Park, K. Kim, S. B. Kim, J. S. Lee, G. B. Mills, and J. Y. Cho, "Gene expression signature analysis identifies vorinostat as a candidate therapy for gastric cancer," PLoS. One. **6**, e24662 (2011).

6. J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha, and A. J. Butte, "Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease," Sci. Transl. Med. **3**, 96ra76 (2011).

7. Y. Ishimatsu-Tsuji, T. Soma, and J. Kishimoto, "Identification of novel hair-growth inducers by means of connectivity mapping," FASEB J. **24**, 1489-1496 (2010).

8. S. D. Kunkel, M. Suneja, S. M. Ebert, K. S. Bongers, D. K. Fox, S. E. Malmberg, F. Alipour, R. K. Shields, and C. M. Adams, "mRNA expression signatures of human skeletal muscle atrophy identify a natural compound that increases muscle mass," Cell Metab **13**, 627-638 (2011).

9. T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets--10 years on," Nucleic Acids Res. **39**, D1005-D1010 (2011).

10. M. Iskar, M. Campillos, M. Kuhn, L. J. Jensen, V. van Noort, and P. Bork, "Drug-induced regulation of target expression," PLoS. Comput. Biol. **6**, (2010).

11. F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, and D. di Bernardo, "Discovery of drug mode of action and drug repositioning from transcriptional responses," Proc. Natl. Acad. Sci. U. S. A **107**, 14621-14626 (2010).

12. J. D. Tenenbaum, M. G. Walker, P. J. Utz, and A. J. Butte, "Expression-based Pathway Signature Analysis (EPSA): mining publicly available microarray data for insight into human disease," BMC. Med. Genomics **1**, 51 (2008).

13. Y. Yi, C. Li, C. Miller, and A. L. George, Jr., "Strategy for encoding and comparison of gene expression signatures," Genome Biol. **8**, R133 (2007).

14. A. C. Gower, A. Spira, and M. E. Lenburg, "Discovering biological connections between experimental conditions based on common patterns of differential gene expression," BMC. Bioinformatics. **12**, 381 (2011).

15. J. M. Engreitz, R. Chen, A. A. Morgan, J. T. Dudley, R. Mallelwar, and A. J. Butte, "ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression," Bioinformatics. **27**, 3317-3318 (2011).

16. M. A. Sartor, G. D. Leikauf, and M. Medvedovic, "LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data," Bioinformatics. **25**, 211-217 (2009).

17. R. Z. Vencio and I. Shmulevich, "ProbCD: enrichment analysis accounting for categorization uncertainty," BMC. Bioinformatics. **8**, 383 (2007).

18. S. W. Tanner and P. Agarwal, "Gene Vector Analysis (Geneva): a unified method to detect differentially-regulated gene sets and similar microarray experiments," BMC. Bioinformatics. **9**, 348 (2008).

19. J. M. Freudenberg, S. Sivaganesan, M. Phatak, K. Shinde, and M. Medvedovic, "Generalized random set framework for functional enrichment analysis using primary genomics datasets," Bioinformatics. **27**, 70-77 (2011).

20. M. R. Segal, H. Xiong, H. Bengtsson, R. Bourgon, and R. Gentleman, "Querying genomic databases: refining the connectivity map," Stat. Appl. Genet. Mol. Biol. **11**, (2012).

21. R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," Biostatistics. **4**, 249-264 (2003).

22. A. Ben-Hur and I. Guyon, "Detecting stable clusters using principal component analysis," Methods Mol. Biol. **224**, 159-182 (2003).

23. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, and M. Muller, "pROC: an open-source package for R and S+ to analyze and compare ROC curves," BMC. Bioinformatics. **12**, 77 (2011).

24. D. Shigemizu, Z. Hu, J. H. Hung, C. L. Huang, Y. Wang, and C. DeLisi, "Using functional signatures to identify repositioned drugs for breast, myelogenous leukemia and prostate cancer," PLoS. Comput. Biol. **8**, e1002347 (2012).

PREDICTIVE SYSTEMS BIOLOGY APPROACH TO BROAD-SPECTRUM, HOST-DIRECTED DRUG TARGET DISCOVERY IN INFECTIOUS DISEASES

RAMON M. FELCIANO[†*], SINA BAVARI[‡], DANIEL R. RICHARDS[†], JEAN-NOEL BILLAUD[†],
TRAVIS WARREN[‡], REKHA PANCHAL[‡], ANDREAS KRÄMER[†]

[†]Ingenuity Systems, Inc
1700 Seaport Blvd, 3rd Floor
Redwood City, CA 94107
email: felciano@ingenuity.com
** Corresponding author*

[‡]USAMRIID
1425 Porter Street
Fort Detrick, MD 21702
email: sina.bavari@us.army.mil

Knowledge of immune system and host-pathogen pathways can inform development of targeted therapies and molecular diagnostics based on a mechanistic understanding of disease pathogenesis and the host response. We investigated the feasibility of rapid target discovery for novel broad-spectrum molecular therapeutics through comprehensive systems biology modeling and analysis of pathogen and host-response pathways and mechanisms. We developed a system to identify and prioritize candidate host targets based on strength of mechanistic evidence characterizing the role of the target in pathogenesis and tractability desiderata that include optimal delivery of new indications through potential repurposing of existing compounds or therapeutics. Empirical validation of predicted targets in cellular and mouse model systems documented an effective target prediction rate of 34%, suggesting that such computational discovery approaches should be part of target discovery efforts in operational clinical or biodefense research initiatives. We describe our target discovery methodology, technical implementation, and experimental results. Our work demonstrates the potential for *in silico* pathway models to enable rapid, systematic identification and prioritization of novel targets against existing or emerging biological threats, thus accelerating drug discovery and medical countermeasures research.

1. Background

New and reemerging infectious diseases pose a growing global health risk across public health concerns and potential bioterrorism threats. Pandemic viruses, resistant bacteria, and technology improvements in bioengineering point to a need for accelerated drug discovery[1]. One approach to this challenge is to use computational techniques to efficiently identify drug targets that may effectively mount a defense against one or more biothreats[2]. Biologically diverse pathogens share common or similar mechanism of infection and pathogenesis, and the host has similarly conserved immune response biology[3–5].

We have previously demonstrated the broad applicability of systems biology analyses to drug discovery and development focused on mammalian disease biology[8–10]. We hypothesize that similar computational characterization of pathogen biology, pathogenesis and host-response genomic pathways across multiple infectious agents can enable systematic identification of targets of intervention that will impact multiple pathogens in a similar manner, and thus serve as *broad-spectrum drug targets* that can be modulated by novel or repurposed therapeutic modalities[6,7]. To test this hypothesis, we extended our approach to identify and predict host-based pathway mechanisms that, once validated, would have a beneficial therapeutic effect against a given pathogen. Validated host pathways and targets can then form the basis of drug repurposing studies, for example to identify compounds previously approved for other disease indications but that share a host mechanism leveraged by a pathogen of interest. We developed computational drug target identification extensions to Ingenuity's pre-existing systems biology platform, and performed a pilot study to experimentally validate predicted targets against six representative "pilot pathogens": Ebola virus, Marburg virus, Lassa virus, *Yersinia pestis, Francisella tularensis*, and *Bacillus anthracis*.

2. Methods

2.1. *Overview of our drug target discovery approach*

Our approach (Figure 1) centers on computer-based modeling of disease pathways using semantic technology, scientific knowledge bases (KBs) of mammalian biochemistry, and

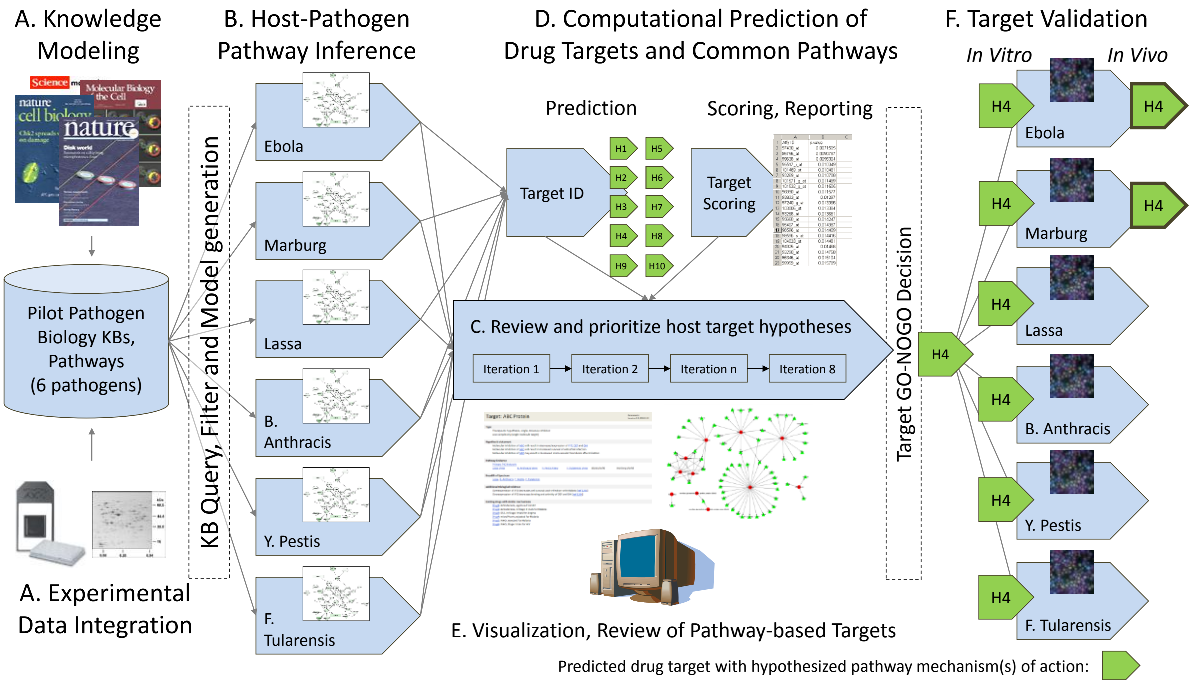

A. Knowledge Modeling
B. Host-Pathogen Pathway Inference
D. Computational Prediction of Drug Targets and Common Pathways
F. Target Validation
In Vitro
In Vivo
Prediction
Scoring, Reporting
Pilot Pathogen Biology KBs, Pathways (6 pathogens)
KB Query, Filter and Model generation
Ebola
Marburg
Lassa
B. Anthracis
Y. Pestis
F. Tularensis
Target ID
H1
H2
H3
H4
H9
H5
H6
H7
H8
H10
Target Scoring
C. Review and prioritize host target hypotheses
Iteration 1
Iteration 2
Iteration n
Iteration 8
Target GO-NOGO Decision
H4
Ebola
Marburg
Lassa
B. Anthracis
Y. Pestis
F. Tularensis
A. Experimental Data Integration
E. Visualization, Review of Pathway-based Targets
Predicted drug target with hypothesized pathway mechanism(s) of action: 🟩

Figure 1. Overview of Ingenuity-USAMRIID predictive systems biology pilot, including knowledge base (KB) construction (A) and host-pathogen pathway model inference (B) for 6 pilot pathogens; multiple rounds ("iterations") of *in silico* target prediction (C) based on suite of target ID algorithms (D); expert review and prioritization of targets using our system prototype (E); and final target selections for *in vitro* and *in* vivo validation at USAMRIID (F). KBs are updated between each iteration. PIC = pathway intervention candidate, i.e. a proposed target centered around the perturbation of a specific pathway of interest.

bioinformatics tools developed by Ingenuity for drug discovery and development and extended herein[10]. We extended existing pathway models of disease biology to bacterial and viral pathogenesis, and developing large-scale, semantically-integrated, knowledge-based models of six pathogens (Ebola virus, Marburg virus, Lassa virus, *Yersinia pestis, Francisella tularensis*, and *Bacillus anthracis*). Specific technology extensions include extending host biomedical ontologies and knowledge models to pathogen biochemistry, pathogenesis staging, and infectious disease; curation and modeling of pathogen-specific pathway content; developing several broad-spectrum target prediction algorithms and target evaluation protocols; and augmenting IPA[11] pathway visualization, filtering and scientific workflows to enable collaborative, team-based broad-spectrum target identification and validation. These extensions, collectively referred to as Pathogen-IPA (P-IPA), were developed as proof-of-concept to demonstrate the feasibility of using computer-based pathway models to accelerate drug target discovery.

2.2. *Knowledge models for target hypothesis generation*

Central to our approach is the notion of *computational hypothesis generation*[12,13], yielding one or more formally-defined "target hypotheses" that relate (1) a host gene or protein and (2) a particular positive or negative impact a drug may have on that target (i.e. "activate" or "inhibit"), and (3) a positive therapeutic effect on one or more clinically-relevant endpoint in hosts infected by each of at least two pathogens. An example of a target hypothesis, rendered computationally to English, is "We hypothesize that inhibition of LAMP2 will counteract the effects of *B. anthracis* and *F. tularensis* (as measured by bacterial uptake studies)". We used P-IPA to computationally characterize the pathogen biology, mechanisms of pathogenesis, and host-response pathways for our 6 pilot pathogens, and use these models to identify and validate one or more such host targets hypotheses.

Table 1. Examples of contextualized pathway findings in our causal reasoning networks, rendered to English syntax through the use of Natural Language Generation algorithms.

Example context	Example of host-pathogen finding(s) in P-IPA causal networks
Attenuated	• Attenuated live *F. tularensis* increases proliferation of human lymphocytes in culture 10-11 months post-treatment.
Virulent	• Decrease of mouse CD45 increases survival of murine-adapted mouse after infection by virulent Ebola virus.
Virulent	• A mutant protein fragment (1-254) (H86K with its Zinc finger domain mutated) from human ZAP protein in Rat2 embryo cells decreases viral replication of Sudan ebolavirus.
Killed or inactivated	• In human neutrophils, killed Marburg virus increases upregulation of human Tlr protein(s) 1 hour post-treatment
Therapeutic (includes vaccine, antiviral, antibacterial	• Oral administration of Salmonella typhimurium-based vector vaccine composed of *Y. pestis* F1 [caf1] protein and of *Y. pestis* V antigen protein increases (by 83 percent) survival of mouse that involves subcutaneous injection of *Y. pestis*.

To generate target hypotheses, we built a global network of causal pathway relationships derived from the Ingenuity Knowledge Base (IKB), a large-scale, manually-curated, semantically-structured ontology-based knowledge base of disease biology research findings[14]. A "finding" is single biochemical insight derived from an original experiment, as supported by primary research or review articles, and tied to a specific biomedical investigation and experimental context. The underlying knowledge representation has semantics based on RDFS[15,16], with pathway models similar to BioPAX Level 3 and SBGN[17], and extensions for modeling drugs, vaccines, biomarkers and clinical phenotypes. We extended IKB with 535,599 new findings curated from primary research, focused on host-pathogen interactions for our 6 pathogens, that increasing the IKB size by 5.1% (Table 1).

Updates to IKB findings and pathway models are ongoing. On a weekly basis a series of knowledge transformations post-process IKB findings to generate (infer) causal networks and other data structures optimized for specific algorithmic approaches (Figure 2), similar to [18] but using semantic rather than linguistic dependency graphs. We infer a causal network where nodes represent form-, species- and state-specific molecules: DNA, RNA, protein, complexes, or pathogen particles, including strain-specific forms. Directional edges represent causal dependencies between the biological activity of linked nodes. These cause-effect relationships include gene regulation, activation / inhibition, chemical modification and other interactions, as supported by one or more experimentally-demonstrated findings from IKB. Such findings are classified by implied direction of change (DOC) of the associated

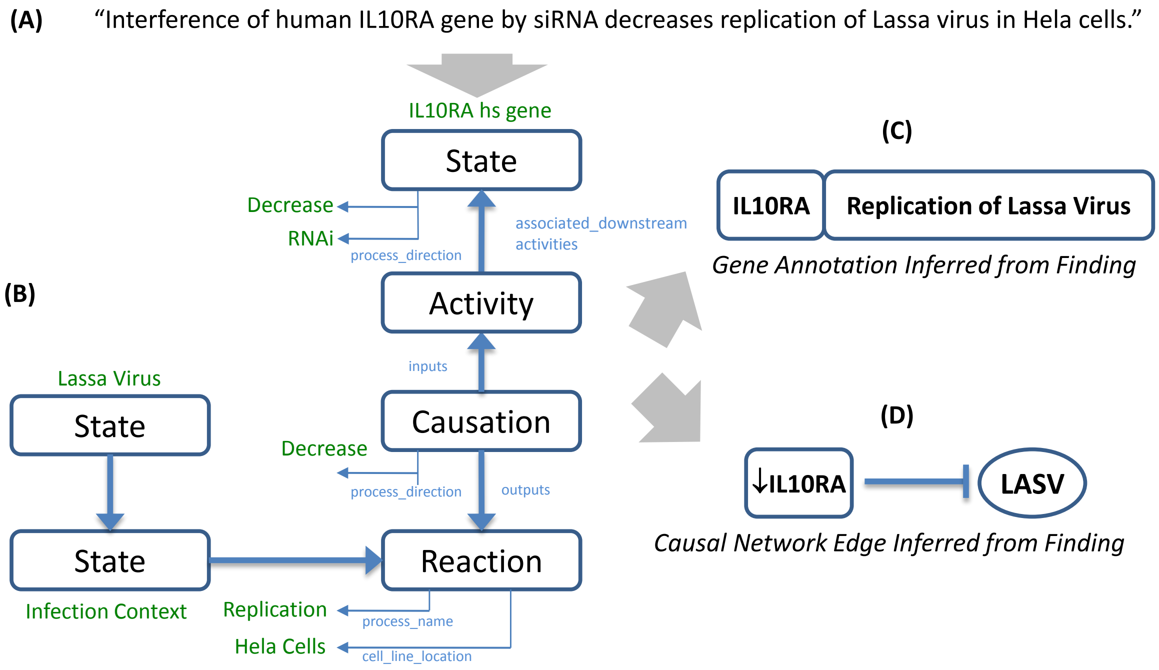

(A)
"Interference of human IL10RA gene by siRNA decreases replication of Lassa virus in Hela cells."
IL10RA hs gene
State
Decrease
RNAi
process_direction
associated_downstream activities
Activity
inputs
Causation
Decrease
process_direction
outputs
(B)
Lassa Virus
State
State
Infection Context
Reaction
Replication
process_name
Hela Cells
cell_line_location
(C)
IL10RA
Replication of Lassa Virus
Gene Annotation Inferred from Finding
(D)
↓IL10RA
LASV
Causal Network Edge Inferred from Finding

Figure 2. Example causal finding used in our predictive analytics. This example illustrates how a single experimental observation (A) is modeled as a semantic network of interrelated concepts (B), which can then be further transformed into a number of secondary data structures useful for computation, such as gene annotations (C) and causal network relationships (D).

causal effect (*increase*, *decrease*, *affects* or *no-effect*). For example, the finding “In human neutrophils, killed Marburg virus increases upregulation of human Tlr protein(s) 1 hour after initial treatment” (see Table 1) would result in a positive causal regulatory relationship between the pathogen (Marburg virus) and host (Tlr protein). Conflicts are resolved by preferentially assigning a DOC if >85% of findings support it, or a non-directional *affects* annotation that must be manually inspected to resolve the conflict.

2.3. *Predictive algorithms for drug target identification*

We identified several *target identification strategies*, each motivated by a specific aspect of pathogenesis that could form basis of a therapeutic strategy and formalized algorithmically to explore the associated hypothesis space using models of host pathobiology pathways. Based on this analysis we developed a general framework for hypothesis generation algorithms, and implemented two complementary approaches for identifying candidate broad-spectrum therapeutic targets, as described in [19] (see supporting materials).

First, we observed that individual host proteins may be regulated in similar ways by multiple pathogens, suggesting an important shared regulatory influence by the pathogen on host proteins. Reversing this regulatory effect may thus therapeutically benefit the host. Our *Commonalities algorithm* seeks to reverse the polarity of multiple pathogens’ similar, direct regulatory effect on a single common host protein, hopefully countering the associated pathogenic impact.

We further observed that multiple host proteins may be similarly regulated by a given pathogen. Rather than pursue a complex “drug cocktail” to target multiple components of this genomic signature, we hypothesize that such panels of host markers may share common upstream regulatory partners. Our second *Upstream Regulators algorithm* thus seeks to identify optimal targets that are upstream of directly affected host molecules, and can serve as a single target more easily modulated by a novel or repurposed drug.

Every target hypothesis generated by these algorithms is supported by a (proposed) pathway mechanism that aggregates immunological evidence and a logical rationale for selecting the target. Hypotheses were further cross-referenced and annotated existing drugs that are either FDA-approved or in various stages of clinical trials for other indications[14,20,21]. Availability of compounds against a protein target was not used to generate hypotheses, but served as a “tie breaker” between otherwise biologically compelling targets when prioritizing our final target list for experimental validation.

2.4. *Experimental design for target validation studies*

To assess the effectiveness of our approach, we performed two-phase *in vitro* and *in vivo* validation studies against our predicted host targets. All validation studies were performed by the Bavari lab at the USAMRIID research facilities, using established protocols for working with our pilot pathogens.

For viral *in vitro* studies, Hela cells were selected as a well-established infection model. Two main experimental approaches were used for validating targets against Ebola, Marburg and Lassa: high content image (HCI) analysis and quantitative real time-PCR (qRT-PCR). Both these assays measure viral replication as the relevant biological endpoint. Inhibition or activation of each targets were achieved by transfection of specific siRNA or transfection of cDNA specific to that target, respectively. For bacterial studies we used three specific types of assays: (1) phagocytosis/bacterial uptake, a HCI assay that measures phagocytosis/ bacterial uptake by the macrophages; (2) fluorescent antibodies specific to pathogen protein(s) used to detect the pathogen that has attached to (and thus phagocytosis by) the host cell; and (3) a Live/Dead assay that measures cytotoxicity.

In vivo studies were designed to further validate inhibition-based targets at the USAMRIID research facilities, based on protocols previously designed in the Bavari lab. To knock down target expression, we used antisense phosphomorpholino oligonucleotides (PMO) inhibition technology (GeneTools, LLC. , Philomath, Oregon). Groups of 10 mice were used: one group per target received target-specific PMOs, and a control group receiving either standard non-specific PMOs, or phosphate buffered saline. All animals received PMOs intraperitoneal (i.p.) or intranasal (i.n.) 4 times (-24h, -4h, 24h, 48h) at 100 to 150 μg per injection per mouse. Mice were challenged i.p. at day 0 with the corresponding lethal dose. For one set of *F. tularensis* experiments, the bacterial challenge was performed using intranasal administration to evaluate survival/protection using a different route of infection.

3. Results

3.1. *Target prediction and prioritization*

We used P-IPA to generate a target pipeline of 490 host proteins whose activation or inhibition was predicted to have a beneficial therapeutic impact against at least two of our 6 pilot pathogens. Through iterative review and filtering using the P-IPA tool suite we prioritized this pipeline to identify the most promising targets and select them for target validation. Target hypotheses were reviewed and prioritized in P-IPA based on:

(a) *Broad-spectrum potential.* Selected host targets must be predicted to impair at least 2 of the 6 pilot pathogens.

(b) *Contextual consistency of pathway evidence.* Targets must be supported by a pathway mechanism consistent with existing research data as well as the clinically relevant disease context (e.g. virulent rather than attenuated pathogen strains)

(c) *De novo experimental evidence.* As special case of (b), we re-integrated our *in vitro* experimental results into IKB as "new but unpublished findings" to facilitate *in vivo* target prioritization,.

(d) *Availability of animal models.* Targets must be testable in a mouse system used by a reference animal model for 5 of our pathogens (Ebola, Marburg, *B. anthracis, F. tularensis,* and *Y. pestis*). To the best of our knowledge, there are no well-validated mouse models for Lassa virus.

(e) *Clinically-relevant endpoints.* Target validity should be confirmed against clinically-relevant endpoints (e.g. improved host survival, reduced viral load, etc).

(f) *Operational tractability.* Host targets were tested using of antisense-based intervention across all experiments evaluating loss-of-function or inhibition-based targets, as permitted by schedule and budget constraints that determined the total number of targets we could test.

We selected 28 target hypotheses (16 inhibition-based targets and 14 activation targets) for Phase 1 *in vitro* validation. In Phase 2, 12 targets were selected for *in vivo* testing, including 8 inhibition targets validated *in vitro* (DUSP1, HSP90B1, LAMP1, SERPIN5, SERPINE2, SMAD3, AP3D1, IL10RA), and 4 new targets selected based on new curated findings highlighted in updated prediction runs (BTRC, HGS, PDCD6IP, PPARA).

3.2. *Example broad-spectrum pathway hypothesis and host drug target*

By way of illustrating our approach, we describe one target prediction in detail (Figure 3). Pathogens may similarly activate or inhibit the function several host proteins. Rather than target these commonly-regulated host proteins individually, the *Upstream Regulator algorithm* treats them as protein signature, and tries to identify a single, additional, host protein that could counter or reverse the impact of the pathogen's effect on this signature. In this example, Ebola and Marburg viruses have been reported to inhibit a number of common host proteins (F2, PROC, PLAU, KLKB1, and C1S). IKB findings (and their underlying research publications) further indicate that SERPINE2 represses the activity of the same proteins. Thus, the algorithmically-generated hypothesis is that both viruses build upon the naturally-occurring suppressive effect of SERPINE2 in the host, and that by removing this effect, we may effectively "pull the rug out" from these viruses and potentially slow pathogenesis by making them work harder. Significantly, our hypothesis re-uses findings from cancer and cardiovascular molecular studies that characterize SERPINE2's effect on the other host proteins include results, as SERPINE2 was previously unassociated with viral hemorrhagic fever infection.

3.3. *Classification of broad-spectrum target validation results*

We formalized our performance evaluation developing a classification framework for target validity that partitioning targets based on whether our experimentation demonstrated a desired effect or lack of effect, and whether that effect was deemed to be clearly demonstrated or whether additional studies were needed to confirm the effect. We used a 5 category scale: *clearly-validated*, *possibly-validated*, *not-tested*, *possibly-not-validated*, and *clearly-not-validated*. For *in vitro* assays, we use 30% reduction in viral load or bacterial uptake as a baseline threshold for a *clearly-validated* classification, adjusted to pathogen-specific thresholds if they exist for a specific virus or bacterium. For *in vivo* assays, target validity was defined as a minimum level of protection conveyed to infected mice, consistent with screening practices. Our baseline threshold was >40% survival in mice after 9 to 22 days (depending on the pathogen) and twice (2x) the standard control survival rate, replicated twice with 10+ mice per experiment. Two other target categories—*possibly-validated*, *possibly-not-validated*—demonstrated lesser phenotypic effect or were not

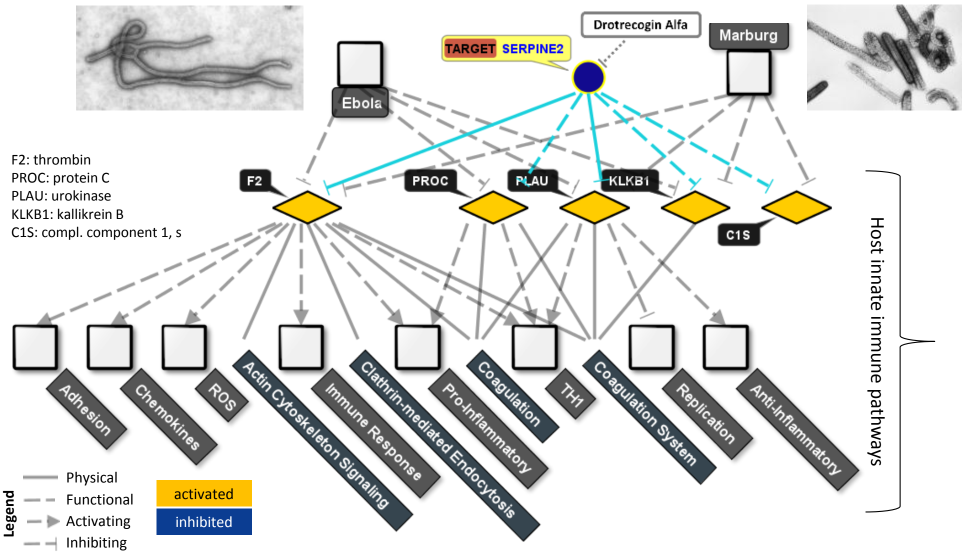
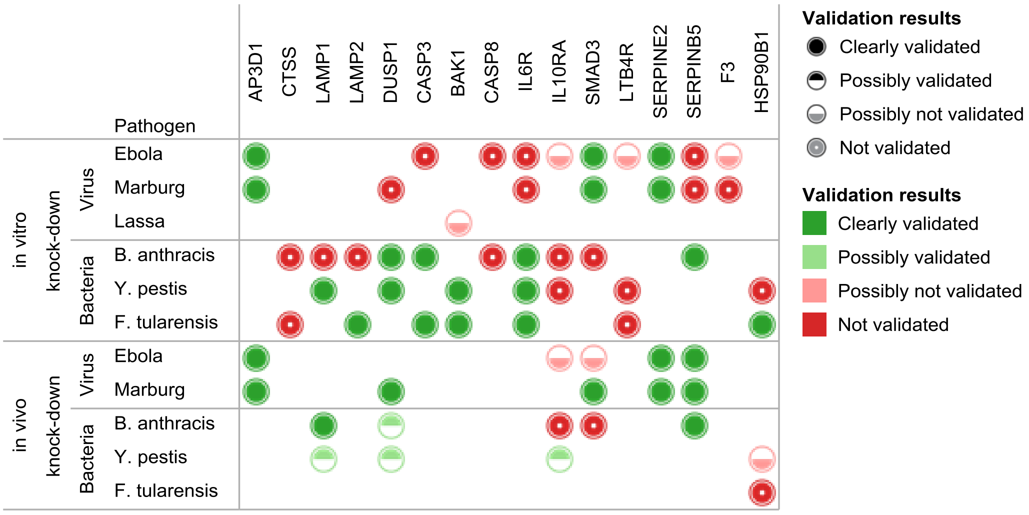


Figure 3. Example of target (SERPINE2, blue node) hypothesis identified by the upstream regulators algorithm as playing a common role in pathogenesis of Ebola virus and Marburg virus, and a drug (Drotrecogin Alfa / Xigris™, Elli Lily)) that may be repositioning for this indication. This drug target hypothesis is grounded in signature of host proteins (yellow nodes) that are commonly downregulated by Ebola and Marburg infection. SERPINE2 is further linked to relevant immune functions, including ones found in viral hemorrhagic fever infection (e.g. coagulation pathways). SERPINE2 was validated *in vitro* and *in vivo* to have the predicted effect on systems infected by the Ebola and Marburg viruses.



			Pathogen	AP3D1	CTSS	LAMP1	LAMP2	DUSP1	CASP3	BAK1	CASP8	IL6R	IL10RA	SMAD3	LTB4R	SERPINE2	SERPINB5	F3	HSP90B1
in vitro	knock-down	Virus	Ebola	🟩●					🟥◉		🟥◉	🟥◉	◒	🟩●	◒	🟩●	🟥◉	◒	
			Marburg	🟩●				🟥◉				🟥◉		🟩●		🟩●	🟥◉	🟥◉	
			Lassa							◒									
		Bacteria	B. anthracis		🟥◉	🟥◉	🟥◉	🟩●	🟩●		🟥◉	🟩●	🟥◉	🟥◉			🟩●		
			Y. pestis			🟩●		🟩●		🟩●		🟩●	🟥◉		🟥◉				🟥◉
			F. tularensis		🟥◉		🟩●		🟩●	🟩●		🟩●			🟥◉				🟩●
in vivo	knock-down	Virus	Ebola	🟩●									◒	◒		🟩●	🟩●		
			Marburg	🟩●				🟩●						🟩●		🟩●	🟩●		
		Bacteria	B. anthracis			🟩●		◓					🟥◉	🟥◉			🟩●		
			Y. pestis			◓		◓					◓						◒
			F. tularensis																🟥◉

Validation results
● Clearly validated
◓ Possibly validated
◒ Possibly not validated
◉ Not validated

Validation results
🟩 Clearly validated
🟩 Possibly validated
🟥 Possibly not validated
🟥 Not validated

Figure 4. Pipeline of prioritized inhibition-based target hypotheses, with our 16 initially-selected inhibition-based targets. *In vitro* and *in vivo* validation results color code the hypothesis arrows based on success or failure classification. For example, the top-left target is AP3D1, which was predicted to have a beneficial effect under Ebola and Marburg infection if the target was inhibited, shown as two down arrows. Knock-down *in vitro* screens and *in vivo* studies confirmed these predictions (filled circles, green). Pipeline visualization is interactive and updated dynamically as new target hypotheses and validation results are integrated.

replicated across multiple experiments, thus requiring additional study to conclusively rule them in or out as drug targets. Commercial availability or maturity of a given compound through the FDA approval process was presented but not used as a validation criterion.

3.4. *Validation of drug target predictions*

We analyzed the performance of our method using both *in vitro* and *in vivo* experimental data by aggregating, discretizing and classifying this hypothesis-specific target validation data into the classifications, described in section 3.3. Briefly, in vitro validation experiments in Phase I demonstrated that 24 of 28 predicted targets resulted in hits against at least one pilot pathogen, Moreover, 22 hits are broad-spectrum (2 + pathogens) target candidates. For example, SERPINB5 showed clear or partial impact against 4 of our 6 pathogens. From this panel of prioritized targets, 11 of these 12 tested targets showed effect against at least 1 pathogen in mice, and 5 clearly inhibit 2+ pathogens (broad-spectrum). Additional targets showed promise, but require additional work to confirm. Inhibition-based targets in Figure 4 have the greatest potential for drug repurposing with compound inhibitors.

4. DISCUSSION

Based on this analysis, 34% directly predicted targets we tested were validated in mouse models, which we believe to a very promising yield. This lower bound (34% for *in vivo*) is a conservative performance assessment, treating only *clearly-valid* results as successes. Performance increases if one includes targets that showed some promising effect but not sufficient to meet our threshold, although this requires additional experimentation to confirm. Table 2 summarizes our findings as predictive success rate, across activation- and inhibition-hypotheses and *in vitro* and *in vivo* results. SERPINB5 is our strongest validated target, clearly validated against *B. Anthracis*, Ebola virus and Marburg virus, and may further show impact against *F. Tularensis* and *Y. Pestis*, although further studies may be required to optimize dosing to confirm this. As our top-ranking target, we believe SERPINB5 is worthy of further investigation to assess mechanism of action.

Table 2. Topline performance of computational target predictions based on *in vitro* and *in vivo* experimental results, across all prioritized, tested hypotheses

Success rate	N (# tested target hypotheses)	Lower bound (*clearly-validated*)	Upper bound (*clearly-validated* + *possibly-validated*)
In vitro	81	27%	46%
In vivo	32	34%	50%

The measured endpoint across these experiments was percentage survival post-infection and treatment. Specifically, we measure the number of mice (out of a total of 10 per group) that survived following PMO treatment and challenge with the corresponding pathogen. For example, 50% survival rate indicates that 5 of 10 mice survived after treatment. In addition to percentage survival, we factored in the number of independent experiments performed, the number of replicates for a sample test, the difference relative to baseline threshold from the standard control, and non-measurable expert evaluation for a given sample. In some cases we were not able to perform identical replicate experiments for a given pathogen.

Interestingly, *in vivo* results out-performed *in vitro* (34% vs. 27%), which may be attributable the limited applicability of cellular assays for modeling host immune biology, as well as the overall lower number of tests run in animal studies relative to our in vitro studies. In addition, the kinetics of each *in vivo* experiment is dependent on each pathogen, and we occasionally observed off-target effects with scrambled PMOs that enabled some increased survival on its own and which we could not control for. This suggests the need for additional research into effective, low-cost alternatives to animal and clinical studies for drug target validation studies[22].

4.1. *Contributions*

We have demonstrated the use of causal network analysis to effectively identify valid drug target hypotheses for a complex disease indication, with a good success rate as demonstrated experimentally through animal studies. To the best of our knowledge, such predictive causal analytics have not been validated to this extent in a host-directed infectious disease context or across multiple viral and bacterial agents. Further, our novel *upstream regulators algorithm* successfully identified previously unassociated valid protein targets based on the predicted propagation of net regulatory effects on the host-pathogen interface. We propose that causal network analysis can extend to previous target identification approaches[7,23] by identifying valid, functionally important targets not identifiable through study of direct host-pathogen interactions alone.

We attribute part of our success the *accuracy and contextual detail of the underlying causal network*, which in turn is based on semantically-normalized IKB content. In particular, IKB findings are (a) manually modeled by experts to ensure accurate representation of the underlying biology[24]; (b) always supported by experimental evidence (no predicted or inferred data); and (c) annotated in sufficient biological and experimental detail to allow finding inclusion or exclusion based on contextual fit to the pathogen in question. We suggest that such normalized, contextualized, experimentally-grounded network datasets can improve the quality of any causal network analyses by driving the algorithm directly (as is our case), or by serving as a high-quality training set for learning-based approaches[25].

Finally, we developed a *framework for rapid, team-based, computational target discovery* to run multiple target ID algorithms in parallel, formalize their predictive outputs and supporting evidence as hypothesized mechanism of action for a novel drug target, and review and prioritize the targets using interactive, collaborative pathway tools. In addition to supporting rapid, evidence-based generation of target lists for medical countermeasures, we believe this model can be extended to include targets identified experimentally e.g. via screening approaches, as well as expert suggested hypotheses[26], thus potentially helping unify computational and experimental target identification approaches.

Our methodology can be applied to any disease where a body of host pathway knowledge has been experimentally characterized and can be modeled as causal, regulatory network relationships. For novel or emerging pathogens that are as of yet unstudied, evolutionary mapping using next-generation sequencing would allow a similar approach using host-pathogen pathway knowledge from closely-related evolutionary neighbors, although some loss of performance should be expected. Finally, a drug repurposing use case could be directly supported by automatically filtering or prioritizing hypotheses anchored by a specific drug or drug class. This would, in turn, highlight candidate compounds for use in target validation studies.

5. Conclusion and future work

Our scientific objective was to identify broad spectrum countermeasures to viral and intracellular biothreats. We have described and evaluated a novel target discovery methodology that is: *host-directed* and *broad-spectrum* in biological focus; *unbiased* in its consideration of prior target association with the disease of interest; *computationally-enabled* by formal models of *disease pathways* and *host-pathogen mechanisms*; and delivers *testable, evidence-based target hypotheses* suitable for experimental validation in rapid response scenario. Our empirical results validate this approach and, more generally, for the use of causal analysis for the discovery of novel drug targets. While our "pathogen and mechanism first" approach focuses primarily on broad-spectrum therapeutics, we believe this approach is readily adaptable to single-spectrum (i.e. against only one pathogen) target identification scenarios as well as other disease areas. We suggest that systems biology pathway models are sufficiently mature to be used alongside traditional screening-based approaches in most applied drug discovery initiatives.

5.1. *Acknowledgments*

This work is funded by Ingenuity Systems and the Defense Threat Reduction Agency under award W81XWH-08-2-0002, for which the U.S. Army Medical Research Acquisition, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office. We thank COL Joseph Palma, COL George Christopher, COL George Korch, COL John Skvorak, Dr. Eric Van Gieson, Dr. Connie Smalljohn, and the Ingenuity and USAMRIID research communities for their support throughout this investigation. Portions of this work are available from Ingenuity at http://www.ingenuity.com/. Support materials for this work, including technical description of the causal algorithms are available at http://pages.ingenuity.com/PSB2013-Felciano-Supporting-Materials.html.

References

1. Berns, K. I. *et al.* Public health and biosecurity. Adaptations of avian flu virus are a cause for concern. *Science (New York, N.Y.)* **335**, 660–1 (2012).
2. Valdivia-Granda, W. A. Bioinformatics for biodefense: challenges and opportunities. *Biosecurity and bioterrorism : biodefense strategy, practice, and science* **8**, 69–77 (2010).

3. Fauci, A. S. Emerging and re-emerging infectious diseases: influenza as a prototype of the host-pathogen balancing act. *Cell* **124**, 665–70 (2006).
4. Cohen, O. J., Kinter, A. & Fauci, A. S. Host factors in the pathogenesis of HIV disease. *Immunological reviews* **159**, 31–48 (1997).
5. Shurtleff, A. C., Nguyen, T. L., Kingery, D. A. & Bavari, S. Therapeutics for filovirus infection: traditional approaches and progress towards in silico drug design. *Expert opinion on drug discovery* (2012).
6. Bowick, G. C. & Barrett, A. D. T. Comparative pathogenesis and systems biology for biodefense virus vaccine development. *Journal of biomedicine & biotechnology* **2010**, 236528 (2010).
7. Kash, J. C. *et al.* Genomic analysis of increased host immune and cell death responses induced by 1918 influenza virus. **443**, 578–581 (2006).
8. Felciano, R. M. Knowledge-based computational pathways analysis for integrative drug discovery. *American Institute of Chemical Engineers (AIChE) Annual Meeting* (2005).
9. Ficenec, D. *et al.* Computational knowledge integration in biopharmaceutical research. *Brief Bioinform* **4**, 260–278 (2003).
10. Calvano, S. E. *et al.* A network-based analysis of systemic inflammation in humans. *Nature* **437**, 1032–1037 (2005).
11. Ingenuity Systems Inc Ingenuity Pathway Analysis Software. *Ingenuity Technical Documentation* (2011).at <http://www.ingenuity.com/products/ipa/>
12. Callahan, A., Dumontier, M. & Shah, N. H. HyQue: evaluating hypotheses using Semantic Web technologies. *Journal of biomedical semantics* **2 Suppl 2**, S3 (2011).
13. Tran, N., Baral, C., Nagaraj, V. J. & Joshi, L. Knowledge-based framework for hypothesis formation in biochemical networks. *Bioinformatics (Oxford, England)* **21 Suppl 2**, ii213–9 (2005).
14. Ingenuity Systems Inc Ingenuity Knowledge Base. *Ingenuity Technical Documentation* (2011).at <http://www.ingenuity.com/science/knowledgebase/>
15. Ruttenberg, A., Rees, J. A., Samwald, M. & Marshall, M. S. Life sciences on the Semantic Web: the Neurocommons and beyond. *Briefings in bioinformatics* **10**, 193–204 (2009).
16. Lam, H. Y. K. *et al.* AlzPharm: integration of neurodegeneration data using RDF. *BMC bioinformatics* **8 Suppl 3**, S4 (2007).
17. Le Novère, N. *et al.* Meeting report from the first meetings of the Computational Modeling in Biology Network (COMBINE). *Standards in genomic sciences* **5**, 230–42 (2011).
18. Percha, B., Garten, Y. & Altman, R. B. Discovery and explanation of drug-drug interactions via text mining. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 410–21 (2012).
19. Felciano, R. M. *et al.* Systems Biology Approaches to Target and Mechanism Discovery in Infectious Diseases. *2010 CBD S&T Conference Conference Proceedings* (2010).
20. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* **34**, D668–72 (2006).
21. Zhu, F. *et al.* Update of TTD: Therapeutic Target Database. *Nucleic acids research* **38**, D787–91 (2010).
22. Huh, D. *et al.* Reconstituting organ-level lung functions on a chip. *Science (New York, N.Y.)* **328**, 1662–8 (2010).
23. König, R. *et al.* Human host factors required for influenza virus replication. *Nature* **463**, 813–7 (2010).
24. Cho, R. J., Chen, R. O., Felciano, R. M., Richards, D. R. & Norman, P. US Patent US6741986: Method and system for performing information extraction and quality control for a knowledgebase. (2001).
25. Novershtern, N., Regev, A. & Friedman, N. Physical Module Networks: an integrative approach for reconstructing transcription regulation. *Bioinformatics (Oxford, England)* **27**, i177–85 (2011).
26. Clark, T. & Kinoshita, J. Alzforum and SWAN: the present and future of scientific web communities. *Briefings in bioinformatics* **8**, 163–71 (2007).

A NOVEL MULTI-MODAL DRUG REPURPOSING APPROACH FOR IDENTIFICATION OF POTENT ACK1 INHIBITORS‡

SHARANGDHAR S. PHATAK

Integrated Molecular Discovery Laboratory (iMDL), The University Texas M.D. Anderson Cancer Center
School of Biomedical Informatics, The Univ. Texas Health Science Center
7000 Fannin St. Ste 600, Houston, Texas, 77030, USA
Email: sharangdhar@gmail.com

SHUXING ZHANG[1]

Integrated Molecular Discovery Laboratory (iMDL), The University Texas M.D. Anderson Cancer Center
1901 East Road, Unit 1950, Houston, TX, 77030, USA
Email: shuzhang@mdanderson.org

Exploiting drug polypharmacology to identify novel modes of actions for drug repurposing has gained significant attentions in the current era of weak drug pipelines. From a serendipitous to systematic or rational ways, a variety of unimodal computational approaches have been developed but the complexity of the problem clearly needs multi-modal approaches for better solutions. In this study, we propose an integrative computational framework based on classical structure-based drug design and chemical-genomic similarity methods, combined with molecular graph theories for this task. Briefly, a pharmacophore modeling method was employed to guide the selection of docked poses resulting from our high-throughput virtual screening. We then evaluated if complementary results (hits missed by docking) can be obtained by using a novel chemo-genomic similarity approach based on chemical/sequence information. Finally, we developed a bipartite-graph based on the extensive data curation of DrugBank, PDB, and UniProt. This drug-target bipartite graph was used to assess similarity of different inhibitors based on their connections to other compounds and targets. The approaches were applied to the repurposing of existing drugs against ACK1, a novel cancer target significantly overexpressed in breast and prostate cancers during their progression. Upon screening of ~1,447 marketed drugs, a final set of 10 hits were selected for experimental testing. Among them, four drugs were identified as potent ACK1 inhibitors. Especially the inhibition of ACK1 by Dasatinib was as strong as IC_{50}=1nM. We anticipate that our novel, integrative strategy can be easily extended to other biological targets with a more comprehensive coverage of known bio-chemical space for repurposing studies.

1. Introduction

The continual decline of the number of new small molecular entities from the pharmaceutical industry pipelines has been well documented[1]. The stop-gap measures such as mergers and outsourcing associated with the modern drug discovery process are unlikely to improve the drug discovery success rates in the long run[2]. Of several approaches under consideration to improve the

[*] To whom correspondence should be addressed.

‡ This work was partially supported by DOD BC085871 and Perot Foundation.

pipeline output, drug repositioning is the one that aims to increase the applicability of already discovered therapeutics to hitherto unknown clinical conditions. This approach may save time and costs associated with the discovery phase[2]. Drug repurposing certainly comes with some distinct advantages and the efforts have been driven by several important factors including: the access to increasing amounts of experimental data (e.g. kinase profiling[3]), better understanding of compound polypharmacology[4], biological data mining (BioCreative III)[5], and regulatory impetus from FDA and NIH[2]. Current successful examples are mostly from serendipitous discoveries such as the repurposing of buproprion from depression to smoking cessation as Zyban[6] and Duloxetine[7] from depression to stress urinary incontinence. Without doubt, there is an unmet need to develop novel, comprehensive methods for systematic drug repositioning to improve the efficiency.

In silico methods, either receptor-based or ligand-based, have been applied to drug repurposing projects. Keiser et al. predicted and validated 23 novel drug-target associations using two-dimensional chemical similarity approach (SEA)[8]. Recently the approach was employed for a large-scale prediction and testing of drug activity on side-effect targets[9]. Ligand-based quantitative structure-activity relationship (QSAR) models have been used by Yang et al. to predict indications for 145 diseases using the side effects as features[10]. With structure-based techniques, inverse docking was also used for drug repositioning[11, 12]. Likewise by mining drug phenotypic side effect similarities, Campillos et al. identified novel drug-target interactions[13]; Oprea et al. incorporated semantic method-based text mining for predicting novel drug actions[2]. With bipartite graph-based methods, Yildirim et al. linked FDA approved drugs to targets using binary associations[14], and Yamanishi predicted drug-target interactions using a combination of graph and chem-genomic approaches[15]. Our group recently conducted a comprehensive review of using molecular networks for drug discovery and development[16]. By developing models with other publicly available data, Dudley et al. repositioned Topiramate, an anti-convulsant drug to potential usage as an inflammatory bowel disease drug[17]. However, these unimodal approaches are likely to be limited by their respective shortcomings, e.g. inverse docking by scoring limitations[18]. Thus we propose that multimodal approaches may offer better solutions by offsetting the weakness of individual methods. In this study, we describe an integrative computational framework based on structure-based drug design and chemical-genomic similarity methods, combined with molecular network theories for drug repurposing. The approaches were applied to identification of existing drugs to target ACK1 for cancer treatment.

ACK1 (activated CDC42 kinase 1) is a ubiquitously expressed atypical non-receptor tyrosine kinase that integrates and delivers signals from multiple ligand-activated receptor tyrosine kinases such as EGFR, HER2 and PDGFR[19]. It also regulates several downstream proteins (e.g. AR, AKT and Wwox) implicated in cell survival roles[19, 20]. The activated ACK1 phosphorylates androgen receptor at Tyr-267 that leads to increased transcription of androgen receptors involved in the development of advanced metastatic prostate cancer or androgen independent prostate cancer[21, 22]. The knockdown of ACK1 increases cell apoptosis in prostate cancer cell lines, suggesting its importance as an anti-oncogenic drug target[22, 23]. Unlike the limited efficacies of conventional targeted therapeutics against RTKs, it has been hinted that ACK1 inhibitors may have higher efficacy for cancer treatment as it integrates signals from multiple RTKs and thus restraining the compensatory mechanisms of RTK signaling[20]. Although inhibitors targeting ACK1 have been developed, publicly available data on them are still limited and few late stage clinical trials are being conducted to date. Therefore, it is an attractive cancer target for drug repurposing.

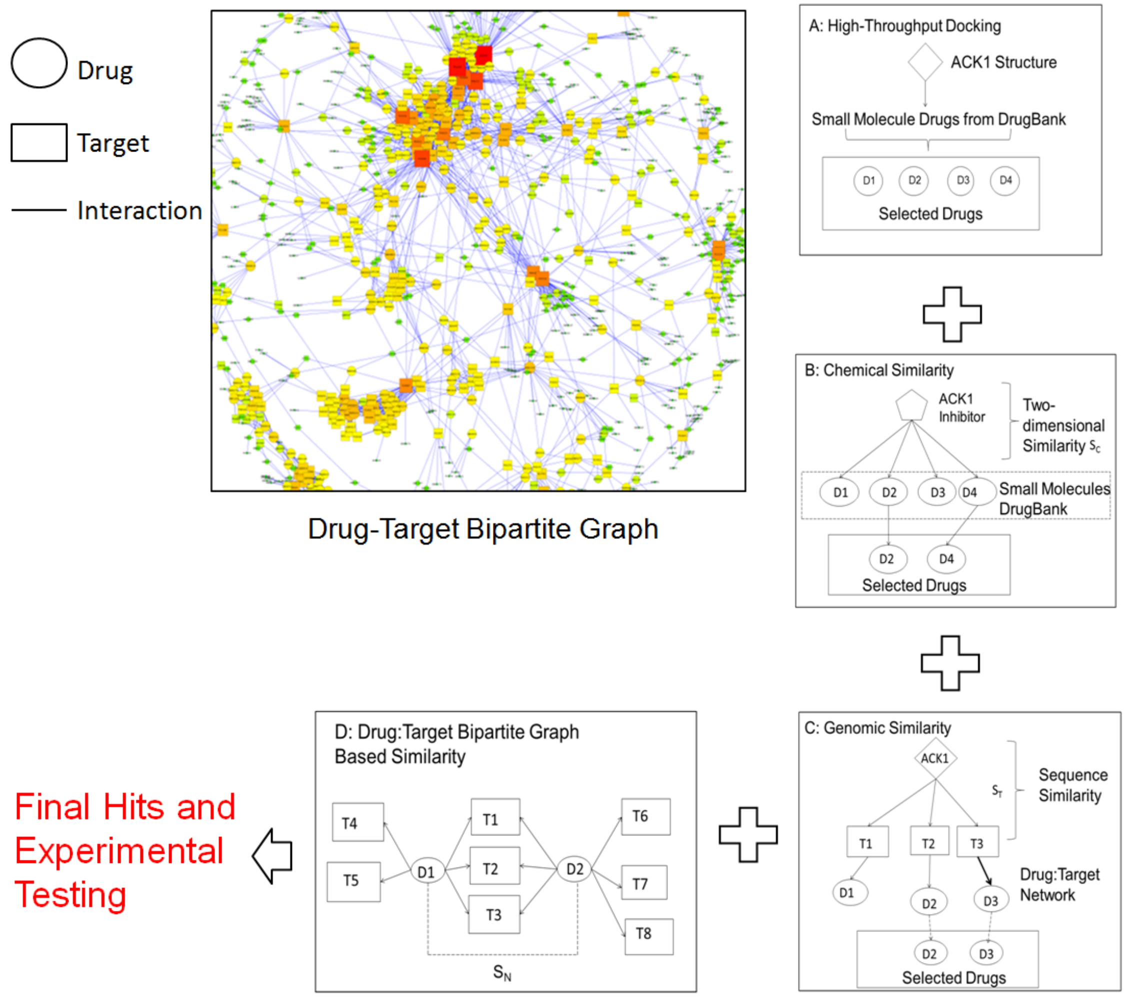


Fig. 1. The schematic Diagram of our modeling workflow. The first step is the construction of the drug-target bipartite graph. Drugs and targets are represented as circles and rectangles, respectively. Node sizes and color are proportional to the degree of each node. The larger shapes and the red color represent nodes with higher degrees. After three steps: **A.** high-throughput docking; **B.** Chemical similarity search using AIM-100, a known inhibitor of ACK1; **C.** Genomic similarity search of ACK1 against proteins in the drug-target graph to identify similar proteins and only the corresponding interacting drugs are selected; **D.** Using only the drug-target graph to identify drugs similar to those identified from steps **A-C**.

With our integrative approach consisting of classical structure-based drug design and chem-genomic similarity analysis approaches in tandem with the bipartite drug-target graph method, we identified 10 drugs for experimental testing. Four of them (Dasatinib, Sunitinib, Flavopiridol and Gefitinib) were confirmed active with IC_{50}<20uM. In particular, the IC_{50} of Dasatinib is as low as 1nM. Our results showed that integrative analysis of chemical-genomic features and molecular networks of drug-targeted interactions, combined with structure-based high-throughput docking could be successfully applied to drug repurposing for potent inhibitor discovery.

2. Methods and Materials

2.1. *Overall Approach*

Our drug repositioning workflow is illustrated by **Fig. 1** using an integrated three-level approach consisting of virtual screening, chemical genomic similarity, and bipartite-graph methods. The bipartite-graphs were developed based on the extensive data curation of DrugBank[23], Protein Data Bank (PDB)[24], and Protein Knowledge Base (UniProt)[25] using in-house developed python scripts. In brief, we employed high-throughput virtual screening followed by a pharmacophore-guided method to select a set of drugs as potential ACK1 inhibitors. Next, we evaluated if complementary results (hits missed by docking) can be obtained by using a novel chemo-genomic similarity approach based on chemical/sequence information. Finally, employing only the drug-target bipartite graph-based similarity, we identified a third set of drugs as potential ACK1 inhibitors. These three sets were further evaluated and merged into our final set consisting of 10 drugs which were evaluated using a qPCR-based kinase assays[26]. Four hits showed strong inhibition of ACK1 (1nM~20μM) and they can be potentially used for prostate cancer treatment.

2.2. *Virtual Screening*

Several structures of the ACK1 kinase domain are available in PDB. For virtual screening we chose two of them (3EQR and 1U4D) which are co-crystallized with very different ligands (T74 and DBQ, respectively). This strategy would implicitly accommodate for receptor flexibility and also possibly help us identify diverse chemotypes. Analysis of these two crystal structures revealed the importance of residues Ala208, Thr205, Glu206, Ala208 and Asp270 because they form hydrogen-bonding interactions with ligands. Particularly in 3EQR, the amine moiety on the 2,6-dimethylphenyl group of T74 interacts with the hydroxyl group on the conserved Thr205 residue. This hydrogen bond was found to significantly enhance the ACK1 inhibition in both biochemical and auto-phosphorylation assays as compared to its parent compound (N-aryl pyrimidine-5-carboxamide series)[27]. It suggests the importance of using this interaction as a pharmacophoric feature for subsequent hit selection. The high-throughput docking was conducted with the Glide software (www.schrodinger.com). Default parameters were used unless otherwise stated. The grid box with size 10Å X 10Å X 10Å was centered on the centroid of ligands (T74 or DBQ), and the active site flexibility was addressed with the induced-fit protocol. Only the approved/experimental drugs from DrugBank were selected for screening, and they were prepared with Epik, including their protonation and tautomer states at pH 7.0. The standard-precision (SP) mode was used for docking and scoring. To validate our protocol, both T74 and DBQ were re-docked into their respective co-crystallized crystal structure. In both cases, the ligands were docked within 1Å of their crystal structure binding poses. The Glide docking scores for T74 and DBQ were -10.4 and -9.26, respectively. Therefore, screened compounds with Glide scores above –9.26 were retained during hit selection via pharmacophore-based visual inspection. The pharmacophores were derived using MOE based on the analysis of the crystal structures and known ACK1 inhibitors (e.g., AIM-100)[28]. To be selected, the hits have to mimic at least three pharmacophoric features: **1).** a hydrophobic moiety in the nucleotide binding pocket surrounded by residues Ile190, Met203, and Leu207; **2).** hydrogen bonds with either Ala208, Thr205, Glu206 or Asp270; and **3).** a polar solvent exposed group in the phosphate binding region of ACK1 surrounded by Asp215 and Arg216. With this strategy, the aim was to reduce the false positives by eliminating the dependence on docking scores as the only parameter because frequently many high ranked compounds could have completely wrong poses due to inaccuracies in scoring functions.

2.3. *Chem-genomic Similarity*

To compensate for the limitations of docking methods (e.g. inaccurate scoring functions), we implemented a novel approach by combining chemical and genomic similarity metrics. This was to identify those missing ACK1 inhibitors from virtual screening. The underlying assumption of our chemical similarity metric is that similar chemistry may result in similar biological activity. To this end, the MACCS fingerprints were employed as they represent chemical substructures within compounds as a *bitstring* using pre-defined substructures and are suitable for such applications. The similarity was expressed with Tanimoto coefficient defined as

$$Tc\left(d_{ij}\right) = |Ai \cap Bj|/|Ai \cup Bj| \qquad \textbf{Eq. 1.}$$

Where: Tc(d_{ij}) = Tanimoto coefficient between drugs i and j. A_i = number of *on* bits (1 is for *on* and 0 is for *off*) in drug i, B_j = number of *on* bits in drug j. This cheminformatics approach was implemented using the Openbabel toolkit (www.openbabel.org). Briefly, a known ACK1 inhibitor AIM-100[29] was used as the query compound and compared with all of the small molecule drugs in our curation. In order to determine the cutoff Tanimoto coefficient, AIM-100 was compared with Dasatinib (Tc = 0.61) as it was shown to be active against ACK1 in our virtual screening study. Therefore, only those drugs that were similar to AIM-100 with ± 5% of Tc = 0.61 were selected, and their affiliated targets in our curated data were obtained.

Genomic-based approaches in such studies were reported to be complementary to their cheminformatics counterparts[30]. Hence to enable rational selection of hits for experimental testing, all protein sequences from PDB were compared with the ACK1 kinase domain. For those sequences/targets with a meaningful genomic similarity with ACK1 (defined as sequence identity>40%), their corresponding drugs, if available in our data curation, were selected for experimental testing. For this step, the Needleman-Wunsch algorithm was employed to identify proteins from PDB similar to ACK1 and the proteins must be represented in our bipartite drug-target graph (described below). We considered the drugs connected to these proteins in the bipartite-graph as likely inhibitor candidates against ACK1.

2.4. *The Unweighted Drug-Target Bipartite Graph*

To use drug-target networks[14] in this study, we extensively curated data (e.g., structures, annotations, etc.) from multiple databases including DrugBank, PDB and UniProt, and developed an unweighted drug-target bipartite graph[16, 23]. Once the proteins were identified (e.g. based on genomic similarity), the respective PDB codes would be obtained from PDB and their corresponding co-crystallized drugs would also be derived. However we only selected those drugs that were present in the drug-target bipartite graph but not identified either from virtual screening or from chem-genomic similarity search. To this end, the DrugBank database was downloaded from the website (www.drugbank.ca). The initial database containing 6,711 drug entries included

6,580 small molecule drugs. For this study, entries containing biotech/nutraceuticals, withdrawn, illicit and other non-small molecule like (as defined by the chemical filter developed for this study) were excluded. This eventually resulted in 1,447 approved drugs in our curation. At the time of this work, the drugcard information did not contain The PDB codes were mapped to their respective UniProt codes using a a Biopython (www.biopython.org) based protocol to rationally reduce the complexity of the drug-target bipartite graph by eliminating redundant degrees as one UniProt code can effectively represent multiple pdb codes. Denoting the drug set as D = $\{d_1, d_2, \dots, d_n\}$ and the target UniProt set as U =$\{u_1, u_2, \dots, u_n\}$, the drug-target bipartite graph was developed as G(D,U,E) where E= $\{e_{ij}: d_i \in D, u_j \in U\}$. A link ($e_{ij}$ in E) is established between d_i and u_j only when there is an explicit association in the respective drug record.

2.5. *Graph-based Similarity*

The unweighted and undirected bipartite graph of drugs from DrugBank is shown in **Fig. 1**. Here, drugs are represented as vertices and their corresponding proteins as edges. Since this graph follows the power-law probability distribution[31], it is feasible to calculate the similarity between two vertices (drugs) based on the shared edges (proteins). Once the similarity of two drugs is established, their affiliated edges (proteins), even unshared ones, may be established as a likely target for the drugs respectively. In our study, we attempted to identify those drugs that shared graph-based similarity with any hit identified from docking and chem-genomic approaches. For the similarity metric we utilized the *Salton's* cosine measure as it normalizes the similarity measures and does not penalize/favor vertices that may have larger number of edges. This graph could easily be represented as an $n \times m$ adjacent matrix $\{a_{ij}\}$ where $a_{ij} = 1$ if d_i and u_j (drug and UniProt, respectively) were connected, or 0 if not. In an undirected network as in our case, the number n_{ij} of common neighbors of vertices i and j is given by:

$$n_{ij} = \sum_k A_{ik} A_{jk} \qquad \textbf{Eq. 2.}$$

Where A is the matrix. Thus, as proposed by Salton, the cosine similarity can be represented as:

$$\sigma_{ij} = \cos\theta = (\sum_k A_{ik} A_{jk}) / (\sqrt{\sum_k A_{ik}^2} \sqrt{\sum_k A_{jk}^2}) \qquad \textbf{Eq. 3.}$$

As our drug-protein network is an unweighted graph, the elements of the adjacency matrix take only the values of 0 and 1, so that $A_{ij}^2 = A_{ij}$ for all i, j. Then $\sum_k A_{ik}^2 = \sum_k A_{ik} = k_i$ where k_i is the degree (number of connections) of vertex i. Thus:

$$\sigma_{ij} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{k_i k_j}} = \frac{n_{ij}}{\sqrt{k_i k_j}} \qquad \textbf{Eq. 4.}$$

In simple terms, the cosine similarity of i and j is therefore the number of common neighbors (in our case, proteins represented by UniProt IDs) between two vertices (represented as drugs) divided by the geometric mean of their degrees. Therefore in this approach, only graph-based geometric similarity is considered without including any chemical/biological information.

2.6. *Experimental Testing*

To validate our predictions, the selected drugs were experimentally tested using the proprietary screening platform with a quantitative qPCR-based assay[26]. This approach measures the amount of DNA-tagged kinase that is unable to bind to an immobilized ligand attached to a fixed support. The kinase assays were developed as kinase-tagged T7 phage strains that are grown in parallel in 24-well blocks in an *E. coli* host derived from the BL21 strain and tagged with DNA for qPCR detection. Streptavidin-coated magnetic beads treated with biotinylated small molecule ligands for 30 minutes at room temperature were used to measure binding affinities for kinase assays. All hits were prepared as 40x stocks in 100% DMSO and directly diluted in the assays. All reactions were performed in polypropylene 384-well plates in final volume of 0.04 ml. The assay plates were incubated at room temperature with shaking for 1 hour, and the affinity beads was washed with buffer (1 X PBS, 0.05% Tween 20). The beads were re-suspended in elution buffer (1 X PBS, 0.05% Tween 20 0.5μM non-biotinylated affinity ligand). The kinase concentration in the eluates was measured by qPCR. The compounds were screened at 0.1μM and 10μM. In addition to ACK1, five other kinases of our interest and implicated in important cancer signaling pathways were used to evaluate selectivity of these inhibitors. The results for primary screen binding interactions were reported as %Ctrl where lower numbers indicate stronger hits:

$$\%\text{Ctrl calculation} = \frac{\text{test compound signal} - \text{positive control signal}}{\text{negative control signal} - \text{positive control signal}} * 100 \qquad \textbf{Eq. 5}.$$

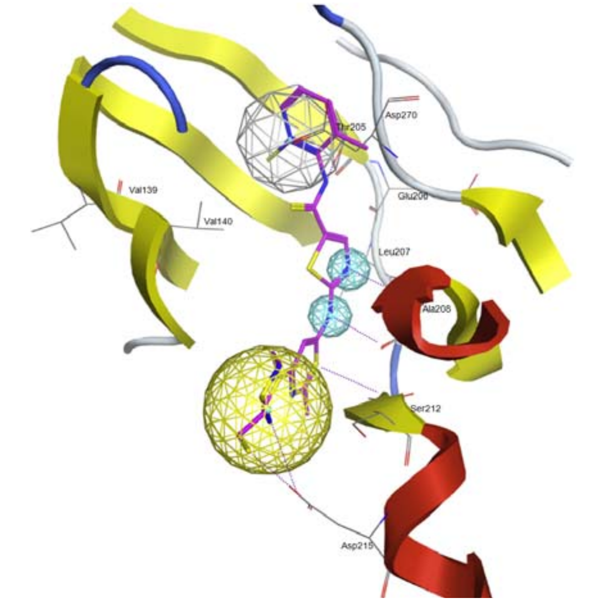


Fig. 2. Dasatinib (magenta sticks) docked into ACK1 (ribbon display). It was ranked top and has reasonable interactions with ACK1. The gray lines are critical residues in the active site. Hydrogen bonds are in magenta dashed lines. The spheres are pharmacophores: gray for hydrophobic, cyan for hydrogen bonds, and yellow for solvent exposed groups.

3. Results

3.1. *High-throughput virtual screening*

As described in the **Methods** section, small molecule drugs were docked and scored against two ACK1 crystal structures. Drugs scored above -9.26 were selected, also based on specific pharmacophoric features characterizing the binding poses. We particularly were interested in those hits with a hydrophobic moiety in the nucleotide binding pocket and forming hydrogen bonds with the Thr205 pocket. For example, Indinavir, a HIV protease inhibitor, was discarded despite being the best ranked hit (data not shown). On the other hand, although Dasatinib only ranked the 8th, it was selected because the drug demonstrated consistent binding pose with ACK1 (**Fig. 2**). Similarly, Amodiaquine, Flavoxate, Imatinib and Lapatinib were also selected based on our docking studies with 3EQR and Mebendazole with 1U4D crystal structures. These hits also exhibited similar shape properties to the co-crystallized ligands of the respective crystal structure. We found that hits from 3EQR had high average molecular weight of 461Da (T74 MWT is 514Da). Screening with 1U4D which has the smaller co-crysallized ligand (DBQ, MWT =254Da) resulted in smaller hit (e.g., Mebendazole MWT=295Da). This was in-line with our hypothesis that diverse chemotypes might be obtained when different crystal structures are used.

3.2. *Chem-genomics based inhibitor identification*

The fundamental principle behind this approach is: **a).** compounds with similar chemistry are likely to possess similar biological profiles, and **b).** if there is meaningful genomic similarity (e.g., high sequence identity) between two proteins (thus also similar tertiary profile), compounds binding to one protein may interact with the other protein as well. We employed AIM-100 inhibitor for chemical similarity search. To determine the Tanimoto coefficient (Tc) threshold, AIM-100 was compared with Dasatinib (a promising binder based on docking) and we obtained Tc=0.61. Hence, all similar drugs within ±5% of Tc were kept. The small range of Tc is to ensure that the hits would maintain a certain degree of both chemical similarity and diversity. Based on drug-target bipartite graph, the corresponding targets of these selected drugs were also identified.

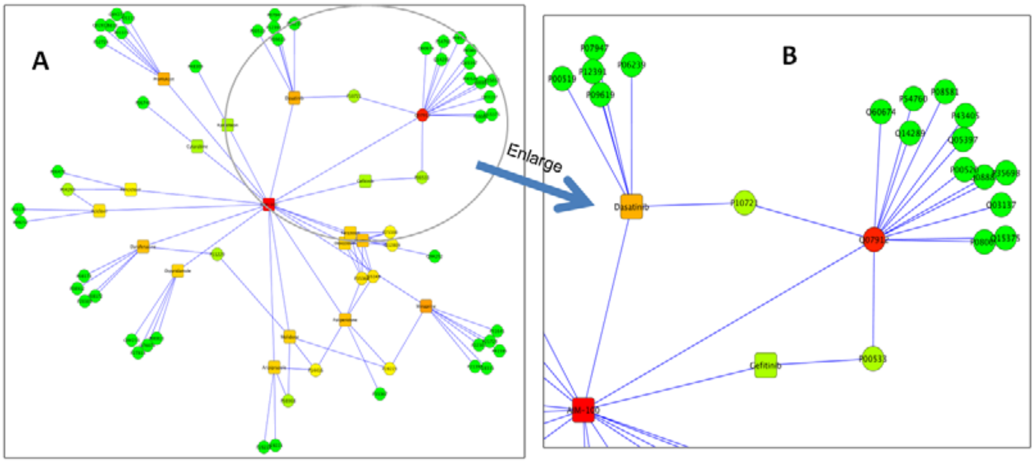


Fig. 3. A. The graph was derived from the drug chemical similarity and target genomic similarity. It represents the inhibitor AIM-100 (red square) and ACK1 (red circle) and those drugs obtained from the chemical similarity search (non-red squares) and proteins similar to ACK1 (green circles). **B.** The enlarged portion of graph **A** shows Gefitinib is similar to AIM-100 and its target (P00533) has significant genomic similarity to ACK1.

On the other hand, all proteins in our dataset were identified based on their genomic similarity to ACK1 (sequence identity>40%), and then their corresponding bound drugs were also obtained. These two sets of selected drug-target pairs were merged if two pairs shared the same target or the same drug. This resulted in a graph as demonstrated in **Fig 3**. Based on this combined chem-genomic similarity approach, Gefitinib, Sorafenib and Sunitinib were identified after excluding those (e.g., Imatinib) already identified by molecular docking. These observations were consistent with our postulation that combining *in silico* approaches, e.g. classical structure-based methods with molecular networks, might help identify unique and complementary sets of inhibitors.

3.3. *Graph-based similarity*

In this step, we attempted to identify potential ACK1 inhibitors based on their similarity to those already identified in the previous steps. However, the strategy was not based on chemical structure or genomic sequence similarities. Instead, the similarity was defined purely with our drug-target graph-based geometry (e.g., vertices and edges) without considering other chemical/biological information. We tried to investigate if this could provide us any extra hits. Using *Salton's* cosine index we calculated a similarity matrix based on the bipartite graph with the shared edges (proteins). A snapshot of the entire matrix is shown in **Fig. 4**. The hypothesis was that any small molecule drugs that showed some similarity to the previously identified inhibitors from the docking and chem-genomic similarity steps might be an inhibitor as well. As expected, we were able to identify the majority of the common hits such as Dasatinib and Imatinib (identified by both docking and similarity search methods). But we also identified new hits such as Flavopiridol as one of the ACK1 inhibitors, based on its graph similarity to Lapatinib. Though several other drugs were also identified, only Flavopiridol, along with another 9 drugs, was purchased for experimental testing due to the constraints of their commercial availability and our budgets.

3.4. *Experimental Results*

The *Kinomescan's* proprietary platform based on several thousands of profiled kinase inhibitors allowed the estimation of binding affinities of any compound based on their primary screening. The specific assay details of this approach are described elsewhere[26]. In addition to ACK1, we screened our selected compounds against several other kinases including EGFR, MEK1, PDPK1, PIK3CA and ABL2, because these targets are suggested to play important and diverse roles in various cancer pathways. EGFR, PDPK1 and PIK3CA are located in the signal transduction pathways that aid tumor growth and reduce apoptosis. MEK1 is located in the MAPK cell signaling that might affect the prognosis of the androgen-independent prostate cancer. We also tested ABL2 as it is the reported target of several drugs (e.g., Imatinib and Dasatinib).

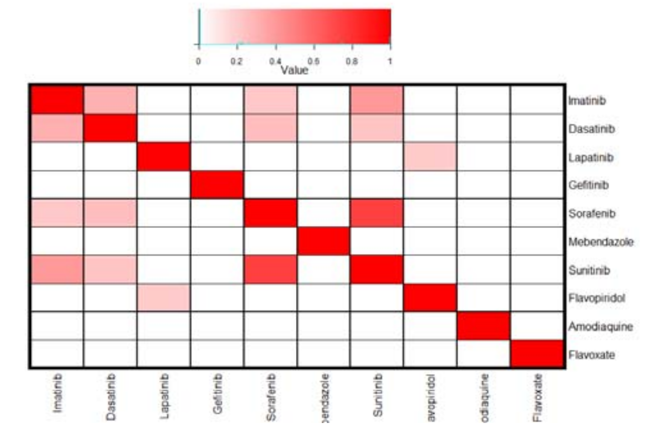


Fig. 4. A representative heat-map of purely graph-based cosine similarities of Flavopiridol against drugs identified from docking and chem-genomic similarity. The higher values (darker red) means higher graph-based similarity.

At the end, 10 hits were purchased and tested. Among them, four drugs including Dasatinib, Sunitinib, Flavopiridol and Gefitinib, showed significant inhibition of ACK1 with estimated $IC_{50}<25\mu M$. The activities of these compounds are illustrated in **Table 1**. These true ACK1 inhibitors were originally designed for different kinases, demonstrating the well-known polypharmacological properties of kinase inhibitors. In particular, Dasatinib was originally designed as a multi-BCR/ABL and Src family tyrosine kinase inhibitor approved for chronic myelogenous leukemia (CML). Here we demonstrated that it also strongly inhibited ACK1 (further experiments showed IC_{50}=1nM) which is implicated in advanced prostate cancer patients. This provided a strong mechanistic support of using Dasatinib to treat prostate cancer. Interestingly, just after our experimental testing of these ACK1 inhibitors, Dr. Whang's group from UNC Chapel Hill published their evaluation of Dasatinib on inhibiting ACK1-related prostate cancer progression *in vitro* and *in vivo*[32]. Their discovery highly conformed to our *in silico* predictions. Currently we are teaming up to further explore repurposing of our identified drugs to treat advanced prostate cancer by targeting ACK1.

Table1. Experimental screening results of in silico drug hits against six kinases.

Target	ACK1		PIK3CA		PDPK1		ABL2		EGFR		MEK	
Compounds	0.1μM	10μM	0.1μM	10μM	0.1μM	10μM	0.1μM	10μM	0.1μM	10μM	0.1μM	10μM
Amodiaquine	89	100	97	93	100	93	97	100	100	91	95	96
Gefitinib	100	77	100	100	100	1000	100	83	2.2	0	100	83
Lapatinib	100	93	98	92	100	100	99	100	0.25	0.05	95	87
Imatinib	100	82	100	92	100	100	43	3.2	100	71	95	94
Dasatinib	4	0	95	98	100	100	0.15	0	59	1.2	89	3.2
Sorafenib	100	89	100	99	100	85	97	33	100	96	100	99
Mebendazole	100	98	100	100	100	100	100	33	100	83	100	38
Flavoxate	100	94	100	100	100	100	100	94	100	88	94	95
Sunitinib	93	33	100	83	100	39	92	51	92	72	51	0.1
Flavopiridol	100	74	100	97	100	100	100	80	100	53	92	92

Sunitinib, Flavopiridol and Gefitinib were originally developed as PDGFR-Beta, CDK-2, and EGFR inhibitors, respectively, but also inhibited ACK1 based on our results. Imatinib and Sorafenib only showed moderate inhibition of ACK1. Flavoxate and Mebendazole were initially considered interesting as they are not kinase inhibitors but were predicted to inhibit ACK1. Unfortunately experimental results indicated that they were either false positives or weak ACK1 inhibitors. Therefore no further work is being performed on them but our efforts of identifying new chemotypes (non-kinase inhibitors) as ACK1 inhibitors are still undergoing.

3.5. *Comparison of Different Methods*

Our multi-modal approach clearly differs from other unimodal methods developed for drug repurposing such as SEA[30] and AERS-based method[33]. Cheng et al. recently evaluated multiple schemes and they found that their network-based interference (NBI) approach obtained better results in their cases[34]. In this study, we focused on a combined strategy but also investigated in details how each method is different from the others in their ability to identify ACK1 inhibitors. **Table 2** demonstrates docking-based virtual screening could reveal more diverse chemotypes including both kinase and non-kinase inhibitors. As expected, drugs uncovered with chemical structure and protein sequence based similarity analysis are all kinase inhibitors. Gefitinib and Sunitinib were shown to have low micromolar affinity to ACK1. Lastly, the graph-based similarity method, which does not include any chemical, biological, or sequence/structure information, identified a different chemotype drug -- Flavopiridol. It exhibits ~25uM inhibition of ACK1. The common hits (blue in **Table 2**) by these methods are Imatinib and Dasatinib, and in particular, the later demonstrates a nanomolar IC_{50}. Clearly this multi-modal approach shows improved performance over each individual methods in the present study.

Table 2. Drugs identified by different methods.

High-throughput Docking	Chem-genomic Similarity Analysis	Graph-based Similarity Analysis
Imatinib	Imatinib	Imatinib
Dasatinib	Sunitinib	Dasatinib
Lapatinib	Gefitinib	Flavopiridol
Mebendazole	Sorafenib	
Amodiaquine		
Flavoxate		

However, certain limitations still exist. For structure-based docking methods, the target 3D structures are usually required. To reduce false predictions, we incorporated as much known expert knowledge as possible such as using multiple ACK1-inhibitor complex structures to partially compensate for target flexibility[18]. We also filtered the top-ranked hits with protein pharmacophores[27]. The chemical similarity based methods are generally reliable, but combination with shape-based techniques may give better results[35]. For the graph-based analysis we were limited to the publicly available drug-protein interaction information. As the data increases, we expect our predictions will be continuously improved.

4. Conclusions

Understanding the drug polypharmacology may hold a great promise in our next generation of drug discovery and development. Along the line, drug repurposing applications are getting more and more attention as it may provide an efficient and effective way to fuel the current drug discovery engines. Both FDA and NIH have recently put a significant amount of funding and effort to promote drug repurposing. From *in silico* point of view, more multi-modal approaches and data integration are needed to increase our opportunity of success. To this end, our present study is to integrate the classical structure-based methods with chem-genomic similarity approaches, along with molecular graph theories to develop new strategies for drug repurposing. Our approach was applied to identification of existing drugs as ACK1 inhibitors for prostate cancer treatment, and multiple potent inhibitors have been discovered.

Our three-pronged approach consisted of curating currently available drug-target information into high-quality bio-chemical databases. Next, by combining the high-throughput molecular docking, chem-genomic similarity search and our in-house drug-target bipartite graphs, we identified 10 promising hits. Subsequent experimental profiling of these selected drugs against six kinases indicated that four of them, including Dasatinib, Sunitinib, Flavopiridol, and Gefitinib, could significantly inhibit ACK1. In particular, the IC_{50} of Dasatinib was as low as 1nM. Therefore we have demonstrated that, extensive analysis of chemical-genomic features, characterization of drug-target relations with graph-based approaches, and classical high-throughput docking are complementary to each other. The combination use of these methods can efficiently and accurately reveal strong inhibitors, corroborating our hypothesis of the need for an integrative approach for drug repurposing. In principle, this approach can be easily extended to other biological targets and chemical databases as a general tools for drug repurposing.

5. Acknowledgements

We specially thank Dr. Young Whang from UNC Chapel Hill School of Medicine for discussing ACK1 inhibitors for treating prostate cancer. We also thank Lu Chen for helping generate pharmacophores. This work was also supported by the MDACC IRG program and the Prostate SPORE Career Award (P50CA140388 SPORE). SSP was supported by UT-Health Innovation for Cancer Prevention Research Pre-Doctoral Fellowship from CPRIT Grant # RP101503.

References

1. B. H. Munos and W. W. Chin, *Sci Transl Med* **3**, 89cm16 (2011)
2. T. I. Oprea, S. K. Nielsen, O. Ursu et al., *Mol Inform* **30**, 100-111 (2012)
3. J. T. Metz, E. F. Johnson, N. B. Soni et al., *Nat Chem Biol* **7**, 200-2 (2011)
4. G. V. Paolini, R. H. Shapland, W. P. van Hoorn et al., *Nat Biotechnol* **24**, 805-15 (2006)
5. C. N. Arighi, P. M. Roberts, S. Agarwal et al., *BMC Bioinformatics* **12 Suppl 8**, S4 (2011)
6. M. S. Boguski, K. D. Mandl and V. P. Sukhatme, *Science* **324**, 1394-5 (2009)
7. T. T. Ashburn and K. B. Thor, *Nat Rev Drug Discov* **3**, 673-83 (2004)
8. M. J. Keiser, V. Setola, J. J. Irwin et al., *Nature* **462**, 175-81 (2009)
9. E. Lounkine, M. J. Keiser, S. Whitebread et al., *Nature* **486**, 361-7 (2012)
10. L. Yang and P. Agarwal, *PLoS One* **6**, e28025 (2011)
11. Y. Z. Chen and D. G. Zhi, *Proteins* **43**, 217-26 (2001)
12. Y. Y. Li, J. An and S. J. Jones, *Genome Inform* **17**, 239-47 (2006)
13. M. Campillos, M. Kuhn, A. C. Gavin et al., *Science* **321**, 263-6 (2008)
14. M. A. Yildirim, K. I. Goh, M. E. Cusick et al., *Nat Biotechnol* **25**, 1119-26 (2007)
15. Y. Yamanishi, M. Araki, A. Gutteridge et al., *Bioinformatics* **24**, i232-40 (2008)
16. J. K. Morrow, L. Tian and S. Zhang, *Crit Rev Biomed Eng* **38**, 143-56 (2010)
17. J. T. Dudley, M. Sirota, M. Shenoy et al., *Sci Transl Med* **3**, 96ra76 (2011)
18. L. Chen, J. K. Morrow, H. T. Tran et al., *Curr Pharm Des* **18**, 1217-39 (2012)
19. K. Mahajan and N. P. Mahajan, *J Cell Physiol* **224**, 327-33 (2010)
20. N. P. Mahajan, Y. E. Whang, J. L. Mohler et al., *Cancer Res* **65**, 10514-23 (2005)
21. M. E. Grossmann, H. Huang and D. J. Tindall, *J Natl Cancer Inst* **93**, 1687-97 (2001)
22. C. D. Chen, D. S. Welsbie, C. Tran et al., *Nat Med* **10**, 33-9 (2004)
23. L. Tian and S. Zhang, *Conf Proc IEEE Eng Med Biol Soc* **2009**, 2336-9 (2009)
24. H. M. Berman, K. Henrick, H. Nakamura et al., *Nat Biotechnol* **25**, 845-6 (2007)
25. R. Apweiler, A. Bairoch, C. H. Wu et al., *Nucleic Acids Res* **32**, D115-9 (2004)
26. M. A. Fabian, W. H. Biggs, 3rd, D. K. Treiber et al., *Nat Biotechnol* **23**, 329-36 (2005)
27. D. J. Kopecky, X. Hao, Y. Chen et al., *Bioorg Med Chem Lett* **18**, 6352-6 (2008)
28. L. Du-Cuny, L. Chen and S. Zhang, *J Chem Inf Model* **51**, 2948-60 (2011)
29. K. Mahajan, D. Coppola, Y. A. Chen et al., *Am J Pathol* **180**, 1386-93 (2012)
30. M. J. Keiser, B. L. Roth, B. N. Armbruster et al., *Nat Biotechnol* **25**, 197-206 (2007)
31. P. Sheridan, T. Kamimura and H. Shimodaira, *PLoS One* **5**, e13580 (2010)
32. Y. Liu, M. Karaca, Z. Zhang et al., *Oncogene* **29**, 3208-16 (2010)
33. M. Takarabe, M. Kotera, Y. Nishimura et al., *Bioinformatics* **28**, i611-i618 (2012)
34. F. Cheng, C. Liu, J. Jiang et al., *PLoS Comput Biol* **8**, e1002503 (2012)
35. S. R. Vasudevan, J. B. Moore, Y. Schymura et al., *J Med Chem* **55**, 7054-60 (2012)

PROTEIN-CHEMICAL INTERACTION PREDICTION VIA KERNELIZED SPARSE LEARNING SVM

YI SHI*[1], XINHUA ZHANG[1], XIAOPING LIAO[2], GUOHUI LIN[1], DALE SCHUURMANS[1]

[1] *Department of Computing Science, University of Alberta,*
Edmonton, Alberta T6G 2E8, Canada
E-mail: {ys3,xinhua2,guohui,daes}@ualberta.ca
[2] *Department of Agricultural, Food and Nutritional Science, University of Alberta,*
Edmonton, Alberta T6G 2P5, Canada
E-mail: xliao2@ualberta.ca

Given the difficulty of experimental determination of drug-protein interactions, there is a significant motivation to develop effective *in silico* prediction methods that can provide both new predictions for experimental verification and supporting evidence for experimental results. Most recently, classification methods such as support vector machines (SVMs) have been applied to drug-target prediction. Unfortunately, these methods generally rely on measures of the maximum "local similarity" between two protein sequences, which could mask important drug-protein interaction information since drugs are much smaller molecules than proteins and drug-target binding regions must comprise only small local regions of the proteins. We therefore develop a novel sparse learning method that considers sets of short peptides. Our method integrates feature selection, multi-instance learning, and Gaussian kernelization into an L_1 norm support vector machine classifier. Experimental results show that it not only outperformed the previous methods but also pointed to an optimal subset of potential binding regions. Supplementary materials are available at "`www.cs.ualberta.ca/~ys3/drug_target`".

Keywords: Drug-target interaction; SVM; Sparse learning; Kernelization.

1. Introduction

Proteins operate in highly interconnected networks ("interactome networks") that play a central role in governing cell functions. If a protein's conformation is changed, its function can be altered, thus affecting cell function. Drugs are small molecules that bind to target proteins to intensionally change the protein conformation, ultimately achieving treatment effects. The function of many classes of pharmaceutically useful protein targets, such as enzymes, ion channels, G protein coupled receptors (GPCRs), and nuclear receptors, can be modulated by ligand interaction. Identifying interaction between ligands and proteins is therefore a key to genomic drug discovery.

Various high-throughput technologies for analyzing the genome, the transcriptome, and the proteome have enhanced our understanding of the space populated by protein classes. Meanwhile, the development of high-throughput screening technology has enabled broader exploration of the space of chemical compounds.[1–3] The goal of the chemical genomics research is to identify potentially useful compounds, such as imaging probes and drug leads, by relating the chemical space to the genomic space. Unfortunately, our understanding of the relationship between the chemical and the genomic spaces remains insufficient. For example, the PubChem database at NCBI[4] contains information of millions of chemical compounds, but the number of compounds with known target proteins is limited. The lack of documented protein-chemical interactions suggests that many remain to be discovered, which motivates the need for improved methods for inferring potential drug-target interactions automatically and efficiently. To facilitate the study of protein-chemical interactions, Kuhn et al. created a protein-chemical interaction database called STITCH,[5] which, up to now, contains interactions for between 300,000 small molecules and 2.6 million proteins from 1,133 organisms.

By elucidating the interaction between proteins and drug molecules, 3D-structure based "docking analysis" has been the principle method for drug discovery.[6–8] In docking analysis, drug-protein binding affinities are modeled by non-covalent intermolecular interactions, such as hydrogen bonding, electrostatic interactions, hydrophobic and Van der Waals forces. Through establishing equations that model the physical interaction between a receptor and potential ligand, the potential energy of binding can be calculated. There are many docking software tools available, including DOCK,[8] GOLD,[6] and AutoDock.[7] All these methods require complete 3D structural information for the target, which might not be available in practice. Such a major disadvantage makes docking analyses infeasible for genome wide application.

Given the difficulty of experimental determination of compound-protein interactions,[9,10] there is a significant motivation to develop effective *in silico* prediction methods that can provide both new predictions for experimental verification and supporting evidence for experimental results. To predict compound-protein interactions various computational approaches have been developed. Keiser et al.[11] propose using the known structure of a set of ligands to predict target protein families. This method does not take advantage of available protein sequence information, and is thus limited to those between known ligands and protein families. Campillos et al.[12] propose predicting drug-target interaction based on similarities between side-effects of known drugs. Some results of this approach have been verified by *in vitro* binding assays, but the approach remains limited to predictions involving drugs with known side-effects. Yamanishi et al.[13] have investigated the relationship between drug chemical structure, target protein sequence, and drug-target network topology, and developed a regression-based learning method for predicting unknown drug-target interactions. In particular, they integrated the chemical and the genomic spaces into a unified space, referred to as the "pharmacological space", wherein chemical-chemical, protein-protein, and chemical-protein similarities can be modeled. Perlman et al.[14] used a combination of Smith-Waterman score, protein-protein interaction, and Gene Ontology information to measure the gene-gene similarity (similarity between targets), but these ancillary information is not always available making the prediction hard to extend to general case, and the way of combining different information sources is somehow tricky.

Most recently, classification methods have been adopted in drug-target prediction.[15–17] These methods firstly calculate the similarities between targets and/or drugs, then use these similarities to construct kernel matrices for the classifiers, such as the support vector machines (SVMs) for predicting novel drug-target interactions. The prediction can be cast into two ways, one for drug side or drug-to-target and the other for target side. For drug-to-target prediction, drug-drug similarities are first obtained, based on structural or pharmacological information; then a bipartite known drug-target interaction graph is constructed; for a new drug with known structural or pharmacological information, its similarities to known drugs are calculated to predict its interactions with known targets using the bipartite interaction graph.

Similarly for target-to-drug prediction, target-target similarities are first obtained using the primary amino acid sequences;[13,17,18] then for a new target with known primary sequence, its similarities to known targets are calculated to predict its interactions with known drugs again using the bipartite interaction graph.

It should be pointed out that in the state-of-the-art works of target-to-drug prediction, the target-target similarity is defined out of the normalized Smith-Waterman score.[17] This S-W score measures the maximum "local similarity" between two protein sequences,[19] thus reasonable, but the local similarity still uses the whole sequences and consequently might involve *long* substrings, which is unreasonable. In fact, long substrings could mask important interaction information, since drugs are usually much smaller molecules than proteins and the drug-target binding sites mostly comprise of only small local regions of the target.

In this work, we focus on the latter target-to-drug prediction to address the issues in the existing works. We first attempt to identify key local binding regions from the *common short* substrings shared by proteins that interact with the same drug. These key short substrings are then used to construct a vector representation for a protein sequence, to be used in the training and testing phases of a classifier. The use of key short substrings (i.e. potential binding regions) as features for the targets is a more direct and meaningful representation for drug interaction prediction. Additionally, the explicit vector representation of targets, as opposed to assessing similarity based on the S-W score, maps the targets into higher dimensional spaces, thus increasing the effectiveness of kernel-based classifiers. We remark that our use of common short substrings differ from the substring composition representation for proteins,[15] which uses all substrings while disregarding whether interactions exist.

The rest of the paper is organized as follows. In Section 2, we introduce the details of our prediction method, in which we focus on the SVM classifiers. We demonstrate in Section 3 the performance of our method compared against the existing ones. Lastly, in Section 4, we discuss the advantages and disadvantages of our method and propose future work.

2. Methods

The drug-target interaction prediction framework is the same as in Bleakley et al.,[17] in which we assume a dataset containing m drugs $d_1, d_2, \ldots, d_m$ and n targets $t_1, t_2, \ldots, t_n$, and the binary indicator on whether or not drug d_i interacts target t_j. The goal is to predict which of the drugs a new target t_c will interact.

2.1. *Target Vectorization*

In the *bipartite local model* (BLM) by Bleakley et al.,[17] to which our method will compare against, the similarity between two targets t and t' is defined as the normalized Smith-Waterman score:[17]

$$s(t, t') = \frac{SW(t, t')}{\sqrt{SW(t, t)}\sqrt{SW(t', t')}}, \tag{1}$$

where $SW(\cdot, \cdot)$ denotes the original Smith-Waterman score.[19] As we mentioned in the introduction, such a similarity measure might overlook the key short sequence regions to which a drug binds.

To address this issue, we want to identify the common short substrings of the targets that interact the same drug. We consider one drug, say d_i, at a time. From the dataset, we first retrieve the set of targets $T_i = \{t_{i1}, t_{i2}, \ldots, t_{in_i}\}$ interacting with d_i. By including the new target t_c, we obtain another set $T_i \cup \{t_c\}$. Using a substring length lower bound, we compute for each of the two sets T_i and $T_i \cup \{t_c\}$ the multi-set of pairwise maximal common substrings, denoted as *withoutSS* $= \{s_{i1}, s_{i2}, \ldots, s_{iq'}\}$ and *withSS* $= \{s_{i1}, s_{i2}, \ldots, s_{ip'}\}$, respectively. In each of the two multi-sets, if two substrings differ at at most one position, they are merged into one and their frequencies are summed together. This way, we obtain two *reduced* sets *withoutSS* $= \{s_{i1}, s_{i2}, \ldots, s_{iq}\}$ and *withSS* $= \{s_{i1}, s_{i2}, \ldots, s_{ip}\}$, containing q and p unique substrings respectively, and each substring is associated with its number of occurences.

Using the substrings in set *withSS* and their occurrences, we can map the n training targets and the new target t_c into the p dimensional Euclidean space, where each substring represents a dimension and the coordinate of target t in dimension s is calculated as the normalized match score between t and s in set *withSS*:

$$M(t, s) = \frac{L(t, s) \cdot c_s}{\sum_{i=1}^{p} c_{s_i}} \;, \tag{2}$$

where $L(\cdot, \cdot)$ is length of the longest common substring between the two sequences and c_s is the number of occurrence of substring s. Intuitively, if target t_c contains a long substring that is also frequent in the binding targets, then its match score for this feature substring will be high indicating a high likelihood of binding. We use $(M(t, s_1), M(t, s_2), \ldots, M(t, s_p))$ as the vector representation for target t.

This way we obtain an $n \times p$ training matrix X, where each row represents a training target, and a $p \times 1$ testing vector $\mathbf{x}_c$ representing the new target t_c, along with the $n \times 1$ binary training label vector $\mathbf{y}$ (with 1 indicating the target interacts with drug d_i and -1 otherwise). The task is to construct a classifier to return 1 if the new target t_c interacts with drug d_i, or -1 otherwise.

The classification problem can be analogously formulated using set *withoutSS* substring set. Next we show how to construct a classifier from the training data.

2.2. *Classification with Feature Selection*

In any classification problem, the quality of features used determines the accuracy of predictions. Here, features correspond to substrings of target proteins, which comprise potential binding regions between the proteins and drugs. Thus, selecting good features not only improves classification accuracy, but also provides candidate drug-target binding sites for further investigation. We investigated an approach that integrates feature selection in L_1-norm based support vector machine (SVM) classification method.

The primal form of L_1-norm SVM is:

$$\begin{array}{rc} & \min\limits_{\mathbf{w},\mathbf{b},\boldsymbol{\xi}} \beta\|\mathbf{w}\|_1 + \mathbf{1}^T\boldsymbol{\xi} \\ \textbf{s.t.}: & \boldsymbol{\xi} \geq \mathbf{1} - \triangle(\mathbf{y})(X\mathbf{w} - b\mathbf{1}), \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{3}$$

where $\triangle(\mathbf{y})$ denotes putting the vector $\mathbf{y}$ on the main diagonal of a square matrix. Here

$X \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \{+1, -1\}^n$, n is the number of data points (targets), and p is the number of features. Since by Micchelli et al.[20]

$$\|\mathbf{w}\|_1 = \min_{\boldsymbol{\gamma} \geq 0} \frac{1}{2} \sum_j (\frac{w_j^2}{\gamma_j} + \gamma_j) = \min_{\boldsymbol{\gamma} \geq 0} \frac{1}{2} (\mathbf{w}^T \triangle(\boldsymbol{\gamma})^{-1} \mathbf{w} + \boldsymbol{\gamma}^T \mathbf{1}),$$

so (3) becomes

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{b}, \boldsymbol{\xi}, \boldsymbol{\gamma}} \ & \frac{\beta}{2} (\mathbf{w}^T \triangle(\boldsymbol{\gamma})^{-1} \mathbf{w} + \boldsymbol{\gamma}^T \mathbf{1}) + \mathbf{1}^T \boldsymbol{\xi} \\ \textbf{s.t.}: \quad & \boldsymbol{\xi} \geq \mathbf{1} - \triangle(\mathbf{y})(X\mathbf{w} - b\mathbf{1}), \\ & \boldsymbol{\xi} \geq \mathbf{0}, \boldsymbol{\gamma} \geq \mathbf{0}. \end{aligned} \tag{4}$$

By introducing Lagrangian multipliers $\boldsymbol{\lambda} \geq \mathbf{0}$ and $\boldsymbol{\mu} \geq \mathbf{0}$, (4) becomes

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{b}, \boldsymbol{\xi}, \boldsymbol{\gamma}} \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \ & \frac{\beta}{2} (\mathbf{w}^T \triangle(\boldsymbol{\gamma})^{-1} \mathbf{w} + \boldsymbol{\gamma}^T \mathbf{1}) + \mathbf{1}^T \boldsymbol{\xi} + \boldsymbol{\lambda}^T (\mathbf{1} - \triangle(\mathbf{y})(X\mathbf{w} - b\mathbf{1}) - \boldsymbol{\xi}) - \boldsymbol{\mu}^T \boldsymbol{\xi} \\ \textbf{s.t.}: \quad & \boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\mu} \geq \mathbf{0}, \boldsymbol{\gamma} \geq \mathbf{0}. \end{aligned} \tag{5}$$

Let the objective function of (5) be L_1, and let $\frac{\partial L_1}{\partial \boldsymbol{\xi}} = 0$, we get $\boldsymbol{\lambda} = \mathbf{1} - \boldsymbol{\mu}$. Therefore, since $\boldsymbol{\mu} \geq \mathbf{0}$, we conclude that $\boldsymbol{\lambda} \leq \mathbf{1}$, hence $\mathbf{0} \leq \boldsymbol{\lambda} \leq \mathbf{1}$. By substitution, (5) becomes

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{b}, \boldsymbol{\gamma}} \max_{\boldsymbol{\lambda}} \ & \frac{\beta}{2} (\mathbf{w}^T \triangle(\boldsymbol{\gamma})^{-1} \mathbf{w} + \boldsymbol{\gamma}^T \mathbf{1}) + \boldsymbol{\lambda}^T \mathbf{1} - \boldsymbol{\lambda}^T \triangle(\mathbf{y}) X \mathbf{w} + b \boldsymbol{\lambda}^T \triangle(\mathbf{y}) \mathbf{1} \\ \textbf{s.t.}: \quad & \mathbf{0} \leq \boldsymbol{\lambda} \leq \mathbf{1}, \\ & \boldsymbol{\gamma} \geq \mathbf{0}. \end{aligned} \tag{6}$$

Let the objective function of (6) be L_2, and let $\frac{\partial L_2}{\partial b} = 0$. We get $\boldsymbol{\lambda}^T \mathbf{y} = 0$, so (6) becomes

$$\begin{aligned} \min_{\mathbf{w}, \boldsymbol{\gamma}} \max_{\boldsymbol{\lambda}} \ & \frac{\beta}{2} (\mathbf{w}^T \triangle(\boldsymbol{\gamma})^{-1} \mathbf{w} + \boldsymbol{\gamma}^T \mathbf{1}) + \boldsymbol{\lambda}^T \mathbf{1} - \boldsymbol{\lambda}^T \triangle(\mathbf{y}) X \mathbf{w} \\ \textbf{s.t.}: \quad & \mathbf{0} \leq \boldsymbol{\lambda} \leq \mathbf{1}, \\ & \boldsymbol{\lambda}^T \mathbf{y} = 0, \\ & \boldsymbol{\gamma} \geq \mathbf{0}. \end{aligned} \tag{7}$$

Let the objective function of (7) be L_3, and let $\frac{\partial L_3}{\partial \mathbf{w}} = 0$, we get $\beta \triangle(\boldsymbol{\gamma})^{-1} \mathbf{w} - X^T \triangle(\mathbf{y}) \lambda = \mathbf{0}$, so that $\mathbf{w} = \frac{1}{\beta} \triangle(\boldsymbol{\gamma}) X^T \triangle(\mathbf{y}) \boldsymbol{\lambda}$. By substitution, (7) becomes

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \max_{\boldsymbol{\lambda}} \ & \boldsymbol{\lambda}^T \mathbf{1} - \frac{1}{2\beta} \boldsymbol{\lambda}^T \triangle(\mathbf{y}) X \triangle(\boldsymbol{\gamma}) X^T \triangle(\mathbf{y}) \boldsymbol{\lambda} + \frac{\beta}{2} \boldsymbol{\gamma}^T \mathbf{1} \\ \textbf{s.t.}: \quad & \mathbf{0} \leq \boldsymbol{\lambda} \leq \mathbf{1}, \\ & \boldsymbol{\lambda}^T \mathbf{y} = 0, \\ & \boldsymbol{\gamma} \geq \mathbf{0}. \end{aligned} \tag{8}$$

Note that $\boldsymbol{\gamma}$ is the feature selection vector. Crucially, this problem is convex in $\boldsymbol{\gamma}$ and has no local minima,[21] hence it provides an optimal form of feature selection that can be efficiently obtained in conjunction with SVM training. Because a drug may bind to different regions of different proteins, i.e., different regions on different targets can bind to the same drug, each positive data point may correspond to a different set of important features (substrings). Therefore, the nature of this drug-target classification problem is essentially a multi-instance classification problem. To address this, we consider two ideas:

Idea (a) Use a radial basis function (RBF) kernel (Gaussian kernel), rather than a linear kernel since this addresses the multi-instance classification problem more effectively after implicitly mapping data points to an infinite dimensional space. After Gaussian kernelization, the original linear kernel matrix $K = X\triangle(\boldsymbol{\gamma})X^T$ becomes $K'_{ij} = e^{\frac{-1}{2\sigma^2}(\mathbf{x}_i - \mathbf{x}_j)^T\triangle(\boldsymbol{\gamma})(\mathbf{x}_i - \mathbf{x}_j)}$.

Idea (b) Because each positive data point may correspond to a unique set of important features, in principle each positive example $\mathbf{x}_i$ should employ its own feature selection vector $\boldsymbol{\gamma}_i^+$ while all negative examples should share a same vector $\boldsymbol{\gamma}^-$. So we get $K''_{ij} = e^{\frac{-1}{2\sigma^2}\|\boldsymbol{\gamma}_i \odot \mathbf{x}_i - \boldsymbol{\gamma}_j \odot \mathbf{x}_j\|^2}$ for all i and j, where $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_i^+$ if $y_i = +1$, and $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}^-$ if $y_i = -1$. Here $\odot$ stands for element-wise multiplication.

Idea (a) can be easily applied to (8) at the sacrifice of convexity, while applying Idea (b) to (8) will introduce too many extra coefficients which makes the model computationally expensive. To circumvent these issues, we introduce an efficient approach to re-weight the features. Intuitively, we wish to down-weight the features that are false positive indicator of binding, i.e. features that have a high score/value at some negative training examples (not bind). This motivation is similar to the case in multi-instance learning, where false positive indicators call for more strict control than true positive indicators. Towards this end, we introduce a p-dimensional weight vector $\mathbf{c}$ corresponding to the p features, and re-scale the feature matrix X by $\tilde{X} = X\triangle(\mathbf{c})$. A simple formula of $\mathbf{c}$ that concretizes our intuition is $c_j = \frac{1}{n}\sum_i a_{ij}$, where $a_{ij} = 1$ if $x_{ij} \leq 1 - \epsilon$ and $y_i = 1$, and $a_{ij} = 0$ otherwise. Here ϵ is a small positive number, and all elements in X are assumed to have been normalized to $[0, 1]$. Therefore by replacing X with $\tilde{X}$ in (8), we encourage using features that indicate less false positive, and formally we obtain

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \max_{\boldsymbol{\lambda}} \;\; & \boldsymbol{\lambda}^T\mathbf{1} - \frac{1}{2\beta}\boldsymbol{\lambda}^T\triangle(\mathbf{y})K'\triangle(\mathbf{y})\boldsymbol{\lambda} + \frac{\beta}{2}\boldsymbol{\gamma}^T\mathbf{1} \\ \textbf{s.t.}: \quad & \mathbf{0} \leq \boldsymbol{\lambda} \leq \mathbf{1}, \\ & \boldsymbol{\lambda}^T\mathbf{y} = 0, \\ & \boldsymbol{\gamma} \geq \mathbf{0}, \end{aligned} \tag{9}$$

where $K'_{ij} = \exp\left(\frac{-1}{2\sigma^2}(\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)^T\triangle(\boldsymbol{\gamma})(\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)\right)$.

We solve (9) by using a combination of L-BFGS-B (Limited-memory Broyden-Fletcher-Goldfarb-Shanno Bounded Optimization) and gradient decent method over $\boldsymbol{\gamma}$. After optimization, we get solutions for $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$. $\boldsymbol{\gamma}$ serves as a useful feature selector, with $\gamma_j > \epsilon$ indicating the j's features should be selected and otherwise not. $\boldsymbol{\lambda}$ can be used to construct the hyperplane in the SVM and to predict new data points. Given a test data point (target) $\mathbf{x}_c$, we can predict its label (binding to the drug or not) based on the sign of the classifier's output:

$$y_c = \sum_{i=1}^{n} \lambda_i y_i \exp\left(\frac{-1}{2\sigma^2}(\tilde{\mathbf{x}}_c - \tilde{\mathbf{x}}_i)^T\triangle(\boldsymbol{\gamma})(\tilde{\mathbf{x}}_c - \tilde{\mathbf{x}}_i)\right) - b. \tag{10}$$

As a key step for solving (9), we need the partial derivative of the objective function in (9) (denoted as L_4) with respect to the k's feature selector γ_k:

$$\frac{\partial L_4}{\partial \gamma_k} = \frac{1}{2\beta} \sum_{ij} \lambda_i \lambda_j y_i y_j \frac{\partial K'_{ij}}{\partial \gamma_k} + \frac{\beta}{2},$$

where

$$\frac{\partial K'_{ij}}{\partial \gamma_k} = K'_{ij} \Big[\frac{-1}{2\sigma^2} (\tilde{x}_{ik} - \tilde{x}_{jk})^2 \Big] = \ \exp\left(\frac{-1}{2\sigma^2} (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)^T \triangle(\boldsymbol{\gamma}) (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)^T \right) \Big[\frac{-1}{2\sigma^2} (\tilde{x}_{ik} - \tilde{x}_{jk})^2 \Big].$$

3. Experimental Results and Discussion

3.1. *Methods under Comparison*

We compared our method with the state-of-the-art method proposed by Bleakley et al.[17] In particular, we focused on target-side prediction of their method to make the two approaches comparable. Bleakley et al.[17] used the normalized Smith-Waterman score in (1) to evaluate the similarity between two target sequences. In the context of SVM classification, they used this target-target similarity matrix as the kernel matrix, i.e. the kernel matrix was fixed in their method. Based on this similarity measure, nearest neighbor (NN) classifiers can also be constructed as a baseline. We refer to Bleakley et al.'s approach as BLM_SVM and BLM_NN respectively. On the other hand, our methods include:

- SS_L1-SVM: L1-SVM with *withSS* feature (the main model of this paper),
- SS_L1-SVM: the classic L2 norm SVM with *withSS* feature,
- SS_NN_FS: nearest neighbor classifier based on the features selected by SS_L1-SVM,
- SS_NN_noFS: nearest neighbor classifier based on all *withSS* features.

3.2. *Experiment Settings*

The framework of our experiment is similar to Bleakley et al.[17] Specifically, we enumerated all pairs of drug d_i and protein t in the whole data set. For each (d_i, t) pair, we treated t as the single test example while the remaining proteins were used as training examples. To learn an L1 and L2 SVM, we chose the hyper-parameters (e.g. β and ϵ) by using three-fold cross validation on the training set, making sure that all the three folders contain at least one target that binds to the drug (i.e., at least one positive example). After the classification model was learned, we applied it to protein t in a way like (10), and obtained a score y_{it} that could be subsequently used to compute useful performance measures (see Section 3.4). All y_{it} calculated by ranging over all drugs d_i and target t in the data set constituted a drug-by-target score table.

We set the minimum length of a feature sub-sequence to 5 after trying all lengths from 4 to 12, noting that a too small cutoff generated excessively many features while a too big cutoff generated an insufficient number of features.

3.3. *Datasets*

We used drug-target interaction information from Bleakley et al.,[17] which was collected from the KEGG BRITE,[2] BRENDA,[22] SuperTarget[23] and DrugBank[24] databases. In particular, we

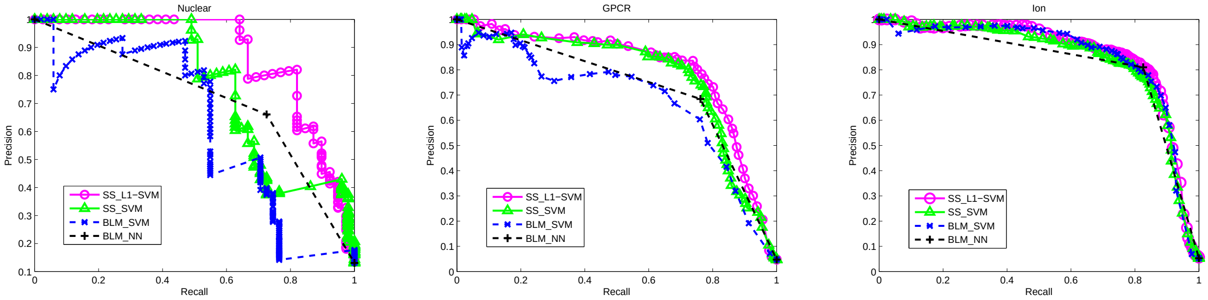


Fig. 1. The precision-recall curves of the four methods SS_L1-SVM, SS_SVM, BLM_SVM, BLM_NN, SS_NN_FS, SS_NN_noFS on three data set. The results are based on training data with drug interacting with at least 2 targets.

used three data sets—nuclear receptors, GPCRS, and ion channel—which have 54, 223, 210 drugs, 26, 95, 204 targets, and 90, 635, 1476 interactions, respectively. The three data sets used in this article are identical to those used in the state-of-the-art study,[17] which facilitates a fair comparison between the two methods. Since we only focused on target-side prediction, we did not require any drug structural or pharmacological information to obtain drug-drug similarity information. The amino acid sequences of the target proteins were obtained from the KEGG GENES database.[2]

3.4. *Classification results*

We used five measurements to evaluate the quality of drug-target prediction: Area under the Precision-Recall Curve (AUPR), Area under the ROC Curve (AUC), F-Measure, Precision (or Specification), and Recall (or Sensitivity). With the prediction score table y_{it} available from Section 3.2, these performance measures were all computed in a micro-average fashion. That is, given a cutoff point, all y_{it} could be converted into a binary label via thresholding (i.e., binding or not). By comparing these labels with the ground truth over the whole drug-by-target score table, we derived the number of false positive and false negative, which led to Precision, Recall, and F-Measure. The AUPR and AUC were derived by increasing the cutoffs with a fine step size, which led to thousands of points in the precision-recall curve. Of the five measurements, AUPR, AUC, and F-Measure are more robust than the others.

We only demonstrate the results based on *withSS* feature because the *withoutSS* feature set did not result in as good performance. Tables 1, 2, and 3 demonstrate the effectiveness of the different drug-target prediction methods over the five evaluation quantities. The F-Measure, Precision, and Recall scores reported in these tables were obtained at the cutoff point where F-Measure was maximized for respective methods. Figure 1 demonstrates the precision-recall curves of SS_L1-SVM and SS_SVM compared to BLM_SVM, BLM_NN, SS_NN_FS, and SS_NN_noFS on three data sets, namely Nuclear, GPCR, and Ion Channel from left to right.

Based on these evaluation, the SVM approaches that use *withSS* feature set (i.e., SS_L1-SVM and SS_SVM) outperform the current state-of-the-art methods BLM_SVM and BLM_NN. Moreover, the L1 norm feature selection method SS_L1-SVM is more effective than the traditional SVM method; it uses only 72.85%, 85.02%, and 62.86% of the original features

Table 1. Evaluations of classification quality on Nuclear data set. The F-Measure, Precision, and Recall scores were obtained at the cutoff point where F-Measure was maximized for respective prediction methods.

Performance comparison:	AUPR	AUC	F-Measure	Precision	Recall
SS_L1-SVM	**0.8756**	**0.9512**	**0.8205**	**0.8205**	0.8205
SS_SVM	0.7635	0.9277	0.7111	0.8205	0.6275
BLM_SVM	0.6163	0.8034	0.6353	0.7941	0.5294
BLM_NN	0.7111	0.8347	0.6916	0.6607	0.7255
SS_NN_FS	0.6985	0.8680	0.6415	0.5075	**0.8718**
SS_NN_noFS	0.6743	0.8459	0.6308	0.5190	0.8039

Table 2. Evaluations of classification quality on GPCR data set. The F-Measure, Precision, and Recall scores were obtained at the cutoff point where F-Measure was maximized for respective prediction methods.

Performance comparison:	AUPR	AUC	F-Measure	Precision	Recall
SS_L1-SVM	**0.8039**	**0.9603**	**0.7840**	**0.8360**	0.7381
SS_SVM	0.7720	0.9600	0.7607	0.8013	0.7240
BLM_SVM	0.6800	0.9435	0.6812	0.7152	0.6503
BLM_NN	0.7287	0.8721	0.7209	0.6842	0.7618
SS_NN_FS	0.7155	0.8878	0.6997	0.6219	**0.7996**
SS_NN_noFS	0.7219	0.8875	0.7081	0.6365	0.7977

Table 3. Evaluations of classification quality on Ion data set. The F-Measure, Precision, and Recall scores were obtained at the cutoff point where F-Measure was maximized for respective prediction methods.

Performance comparison:	AUPR	AUC	F-Measure	Precision	Recall
SS_L1-SVM	**0.8632**	0.9666	**0.8205**	**0.8260**	0.8151
SS_SVM	0.8450	**0.9690**	0.8045	0.8173	0.7921
BLM_SVM	0.8561	0.9568	0.8088	0.7785	**0.8416**
BLM_NN	0.8226	0.9075	0.8179	0.8101	0.8258
SS_NN_FS	0.7041	0.8542	0.6954	0.6647	0.7290
SS_NN_noFS	0.6702	0.8640	0.6497	0.5671	0.7606

in the Nuclear, GPCR, and Ion Channels datasets, respectively. The significant reduction in feature set size can not only make the classification more efficient and effective, it can also help biological practitioners to identify important features more accurately.

We further investigated the prediction result generated by the SS_L1-SVM method and the BLM_SVM method. At the prediction cutoff where both methods attained their own maximum F-Measure score, there were 8, 127, and 78 true positive interactions that SS_L1-SVM managed to identify but were missed by BLM_SVM. This was in comparison to 7, 16, 52 true positives that were identified by BLM_SVM but not by SS_L1-SVM. False positive is another important measurement of a method. On the three datasets Nuclear, GPCR, and Ion Channels, SS_L1-SVM generated 0, 73, and 139 false positive interactions, compared to 2, 85, 117 false positive interactions generated by BLM_SVM.

Some interesting case studies are in order. On the Nuclear dataset, the two nearest neighbors of the target protein RORB (KEGG Homo sapiens protein ID "hsa6096") under the normalized Smith-Waterman score are RORA ("hsa6095") and RORC ("hsa6097"), with scores 0.578 and 0.458 respectively. RORB and RORC share a common interacting drug *Tretinoin* (KEGG drug ID "D00094") while RORB and RORA do not. According to the BLM method, RORB will be predicted to have no interaction with *Tretinoin* because its nearest neighbor RORA does not interact with *Tretinoin.* On the contrary, our method can correctly identify the interaction between RORB and *Tretinoin* because the *withSS* feature set based method can discover two important substrings "EVVLVRMCRA-N" and "N-TV-FEGKYGGM" that exist in both RORB and RORC. Therefore, although the overall match score between RORB and RORC is not the highest, their feature vectors (with respect to the two feature substrings) are the most similar.

On the GPCR dataset, the five nearest neighbors of the target protein CHRM1 (KEGG Homo sapiens protein ID "hsa1128") under the normalized Smith-Waterman scores are CHRM5 ("hsa1133"), CHRM3 ("hsa1131"), CHRM4 ("hsa1132"), CHRM2 ("hsa1129"), and HRH3 ("hsa11255"), with scores 0.4707, 0.4536, 0.4237, 0.4228, and 0.2446 respectively. Although CHRM1 is supposed to bind to drug *Metoclopramide* (KEGG drug ID "D00726"), none of its five nearest neighbors bind to this drug. In fact binding occurs only with the 6-th nearest neighbor, HRH2 ("hsa3274"), whose SW_{norm} score with respect to CHRM1 is 0.2137. Therefore, the BLM methods can hardly predict that CHRM1 binds to *Metoclopramide.* In contrast, our method can correctly predict this interaction because the important substrings such as "KRTPRRAA", "Y-AKRTP-RAA-MI-L-W", and "NYFL-SLA-AD" are present in both CHRM1 and several proteins that bind to *Metoclopramide*, e.g., HTR1A ("hsa3350"), HTR1B ("hsa3351"), HTR1D ("hsa3352"), HTR1E ("hsa3354"), HTR1F ("hsa3355"), HTR2A ("hsa3356"), HTR2B ("hsa3357"), HTR2C("hsa3358"), HTR4("hsa3360"), HTR5A("hsa3361"), and HTR6("hsa3362"), which are all considered as faraway neighbors according to the SW_{norm} scores.

The binding regions discovered by our computational model can also be found on the Ion dataset. To provide potential drug-target binding regions for further investigation, we produced all the important common substrings selected by the SS_L1-SVM method, which are made available online at "`http://www.cs.ualberta.ca/~ys3/drug_target`".

4. Conclusions

In this article, we proposed a novel drug-target interaction prediction method based on potential drug-target binding regions. According to the evaluation metrics, the proposed method significantly outperformed the current state-of-the-art methods. More importantly, it identified a number of drug-target interactions that were missed by previous methods. We believe that the poor recall of previous methods is due to the use of a target kernel matrix based on Smith-Waterman score: a low overall similarity between two protein sequences does not mean they do not share common drug binding regions. This drawback was overcome in our approach by collecting a large number of candidate binding regions (i.e., common substrings) that subsequently played the primary role in interaction prediction. In addition, the use of an explicit vector representation, as opposed to implicit similarity measure, enabled the easy use of non-linear kernel expansions that were not possible for fixed kernel methods like BLM.

Besides the kernels based on substring feature vector, we believe the techniques of string kernel proposed in[25] could be useful in this problem. One straightforward benefit of using the string kernel is that it will automatically consider all substrings of a given sequence pair. It can also provide more intuitive understanding of substring-based sequence similarities than using Gaussian kernel. However, to employ the string kernel, one needs to customize the feature selection function into the model and to extend the non-gapped matching in string kernels.

We presented a feature selection method based on L_1-norm SVM that could not only predict the binding relations more accurately, but also find important candidate binding regions (features). It integrated feature selection directly into L_1-norm SVM and kernelized the optimization model. A drawback was that the sparse regularization term tended to select only a single feature from the candidate set. This is a well known problem with L_1 based regularization.[26] To avoid this limitation, we will investigate a combination of L_1 and L_2 norm regularizers, known as the elastic net,[26] which is generally more effective at group feature selection. Another possible extension is to adopt the OSCAR model,[27] which appears even more effective. We also discovered that the inference problem of drug-target interaction—in some cases—can be considered as a multi-instance learning problem. So we proposed using multiple feature selection vectors for each positive training example in theory and applied the feature cost vector to address the multi-instance problem in practice. We hypothesize that more advanced machine learning methods specifically tailored for multi-instance classification can further improve the accuracy of drug-target interaction prediction. Moreover, considering that protein 3D structures carry the essential binding information and an increasing amount of protein 3D structure is being made available (e.g., PSI Nature Structural Biology Knowledgebase[28]), incorporating protein 3D information in the prediction model in addition to sequence information would lead to promising improvement.

References

1. C. M. Dobson, Chemical space and biology, *Nature* **432**, 824 (2004).
2. M. Kanehisa, S. Goto, M. Hattori and et al., From genomics to chemical genomics: New developments in kegg, *Nucleic Acids Res.* **34**, D354 (2006).
3. B. R. Stockwell, Chemical genetics: Ligand-based discovery of gene function, *Nat. Rev. Genet.* **1**, 116 (2000).
4. D. L. Wheeler, T. Barrett, D. A. Benson and et al., Database resources of the national center for biotechnology information, *Nucleic Acids Res.* **34**, D173 (2006).
5. M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, L. J. Jensen and P. Bork, Stitch 3: zooming in on protein-chemical interactions, *Nucleic Acids Res.* **40**, 876 (2012).
6. G. Jones, P. Willett, R. C. Glen and et al., Development and validation for a genetic algorithm for flexible docking, *J. Mol. Biol.* **267**, 727 (1997).
7. G. M. Morris, D. S. Goodsell, R. S. Halliday and et al., Automated docking using a lamarckian

genetic algorithm and empirical binding free energy function, *J. Comput. Chem.* **19**, 1639 (1998).

8. B. K. Shoichet, D. L. Bodian and I. D. Kuntz, Molecular docking using shape descriptors, *J. Comput. Chem.* **13**, 380 (1992).
9. S. J. Haggarty, K. M. Koeller, J. C. Wong and et al., Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays, *Chem. Biol.* **10**, 383 (2003).
10. F. G. Kuruvilla, A. F. Shamji, S. M. Sternson and et al., Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays, *Nature* **416**, 653 (2002).
11. M. J. Keiser, B. L. Roth, B. N. Armbruster and et al., Relating protein pharmacology by ligand chemistry, *Nat. Biotech.* **25**, 197 (2007).
12. M. Campillos, M. Kuhn, A. C. Gavin and et al., Drug target identification using side-effect similarity, *Science* **321**, 263 (2008).
13. Y. Yamanishi, M. Araki, A. Gutteridge and et al., Prediction of drug-target interaction networks from the integration of chemical and genomic spaces, *Bioinformatics* **24**, i232 (2008).
14. L. Perlman, A. Gottlieb, N. Atias, E. Ruppin and R. Sharan, Combining drug and gene similarity measures for drug-target elucidation, *J. Comput. Biol.* **18(2)**, 133 (2011).
15. N. Nagamine and Y. Sakakibara, Statistical prediction of protein-chemical interactions based on chemical structure and mass spectrometry data, *Bioinformatics* **23**, 2004 (2007).
16. L. Jacob and J. P. Vert, Protein-ligand interaction prediction: An improved chemogenomics approach, *Bioinformatics* **24**, 2149 (2008).
17. K. Bleakley and Y. Yamanishi, Supervised prediction of drug-target interactions using bipartite local models, *Bioinformatics* **25**, 2397 (2009).
18. Y. Yamanishi, M. Kotera, M. Kanehisa and et al., Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework, *Bioinformatics* **26**, i246 (2010).
19. T. F. Smith and M. Waterman, Identification of common molecular subsequences, *J. Mol. Biol* **147**, 195 (1981).
20. C. Micchelli and M. Pontil, Learning the kernel function via regularization, *J. Mach. Learn. Res.* **6**, 1099 (2006).
21. G. Lanckriet, M. Cristianini, P. Bartlett and et al., Learning the kernel matrix with semi-definite programming, *J. Mach. Learn. Res.* **5**, 27 (2004).
22. I. Schomburg, A. Chang, C. Ebeling and et al., Brenda, the enzyme database: Updates and major new developments, *Nucleic Acids Res.* **32**, D431 (2004).
23. S. Gunther, M. Kuhn, M. Dunkel and et al., Supertarget and matador: Resources for exploring drug-target relationships, *Nucleic Acids Res.* **36**, D919 (2008).
24. D. S. Wishart, C. Knox, A. C. Guo and et al., Drugbank: A knowledgebase for drugs, drug actions and drug targets, *Nucleic Acids Res.* **36**, D901 (2007).
25. S. V. N. Vishwanathan and A. J. Smola, Fast kernels for string and tree matching, *NIPS* **15**, 569 (2002).
26. H. Zou and T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society B* **67**, 301 (2005).
27. W. Zhong and J. Kwok, Efficient sparse modeling with automatic feature grouping, *ICML* **28**, 9 (2011).
28. H. M. Berman, Psi nature structural biology knowledgebase (`www.sbkb.org/kb/index.html`) (2012).

DRUG TARGET PREDICTIONS BASED ON HETEROGENEOUS GRAPH INFERENCE

Wenhui Wang† , Sen Yang†, JING Li*
Department of Electrical Engineering and Computer Science, Case Western Reserve University
Cleveland, Ohio, 44106, USA
Emails:{wxw134@case.edu,sxy221@case.edu jingli@case.edu}

A key issue in drug development is to understand the hidden relationships among drugs and targets. Computational methods for novel drug target predictions can greatly reduce time and costs compared with experimental methods. In this paper, we propose a network based computational approach for novel drug and target association predictions. More specifically, a heterogeneous drug-target graph, which incorporates known drug-target interactions as well as drug-drug and target-target similarities, is first constructed. Based on this graph, a novel graph-based inference method is introduced. Compared with two state-of-the-art methods, large-scale cross-validation results indicate that the proposed method can greatly improve novel target predictions.

1. Introduction

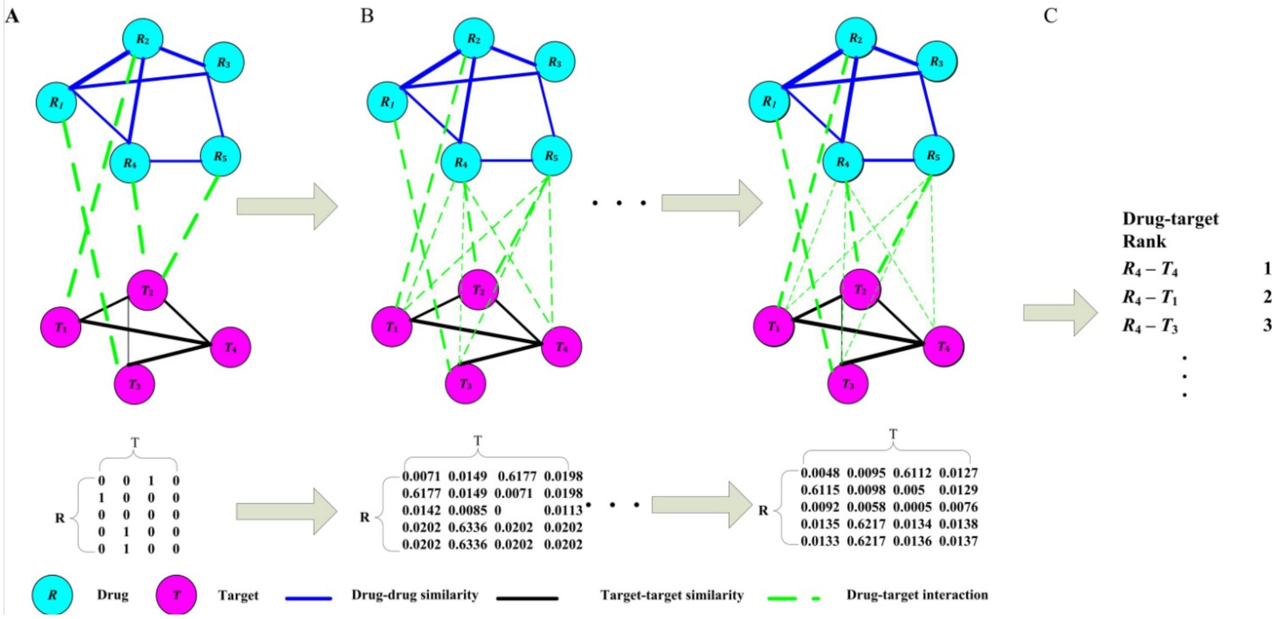
Drug targets are a class of molecular structures which could interact with drugs[1]. Establishing new connections between existing drugs and targets or finding novel targets for a given drug plays an important role in drug development. Experimental prediction of drug-target associations is a laborious and costly task[2]. So far there are only about a few hundreds of known targets[1]. In contrast, there are many more computationally predicted targets, e.g., the so called druggable genome[3,4]. The druggable genome denotes a set of human genes that encode proteins which might be able to bind drug-like molecules[3]. Though different sets of druggable genes have been predicted, the consensus on the number of druggable genes is around 3000[4]. Due to the large number of potential targets, examining each one of them with a specific drug becomes a tedious or even impossible task. From this point of view, an accurate druggable genome filtering or ranking approach becomes in urgent need.

In recent years, a large number of approaches have been proposed to address this problem. Zhu, et al. (2005)[5] attempted to mine implicit chemical compound and gene relations from their co-occurrences in the literature. However, their results were constrained to current knowledge. Furthermore, there are many inconsistencies in target names and drug names, which may adversely affect the accuracy of the results. By using some basic biophysical principles, the structure based maximal affinity model[6] could generate accurate predictions of druggability based solely on the crystal structure of a target's binding site. This method, however, is applicable only when the 3D structures of targets are known, which may not be available in general. More recently, several methods have combined drug-drug or target-target similarities into novel target predictions[7-13]. Phenotypic side-effect similarities were used to build a drug-drug relation network, based

† These authors contributed equally to this work. *To whom corresponds should be addressed.

on which novel drug-target associations were inferred[7]. Yamanishi and coauthors[8,9] formalized the drug-target interaction inference as a supervised learning problem on a bipartite graph. The learning process was based on a unified 'pharmacological space', which was constructed by combining chemical and genomic properties. It has also been shown that chemical similarities between drugs and ligands, small molecules that bind to molecular targets, can be used to predict unanticipated associations[10]. Bipartite local models (BLM) used supervised methods to predict target proteins of a given drug, then to predict drugs targeting a given protein, and finally these two steps were combined to give a final prediction for each drug-target interaction[11]. In another work, Perlman *et al.* (2011)[12] proposed a framework that combines multiple drug-drug and gene-gene similarity measures using a logistic regression model. The final classification score was used to indicate interactions between drugs and targets. Very recently, a network based inference (NBI) method was proposed to infer novel drug-target interactions[13], which ranks drugs for a specific target based on a two-step diffusion model on the bipartite drug-target graph.

The guilt-by-association principle has been widely used in many different domains and applications (e.g., Jeh and Widom (2002)[14]). Chiang and Butte (2009)[15] proposed a novel drug repositioning method based on the guilt-by-association principle. They claimed that suggestions for novel drug uses can be generated from the uses of drugs that cure the same diseases. This assumption was further extended by concluding that similar diseases tend to be connected with similar drugs and similar drugs tend to be connected with similar target[16]. Based on this assumption, the intra-similarity information can be incorporated into novel association predictions by constructing a *heterogeneous drug-target graph/network*, which includes both intra-similarity information (connections between the same kind of nodes, such as drug-drug connections and target-target connections) and interaction information (connections between different kinds of nodes, such as drug-target connections). In this paper, we propose a method, termed HGBI, for **H**eterogeneous **G**raph **B**ased **I**nference, for novel drug target predictions based on the guilt-by-association principle and an intuitive interpretation of information flow on the drug-target heterogeneous graph. The algorithm iteratively updates the measure of strength between unlinked drug-target pairs based on all the paths in the network connecting them. We show that when properly normalized, the proposed procedure will eventually converge and stable relationships between drugs and targets can be achieved. Fig. 1 shows the framework of HGBI. Based on large scale leave-one-out cross-validation experiments, we show that HGBI exhibited superior performance and achieved much higher AUC (area under the receiver operating characteristic, i.e., ROC curve) than two existing state-of-the-art novel drug target prediction methods, BLM[11] and NBI[13]. In particularly, when focusing on the top 1% ranked targets, HGBI successfully retrieved 1339 out of 1915 drug-target interactions, whereas BLM and NBI only retrieved 56 and 10 such interactions. Furthermore, HGBI can establish a novel interaction between a drug and a target even none of the two have directly associated targets/drugs. Some of these novel predictions are confirmed based on a new database, which is not used in this study.



$$R\begin{cases}\begin{matrix}0 & 0 & 1 & 0\\ 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 0\end{matrix}\end{cases} \Rightarrow R\begin{cases}\begin{matrix}0.0071 & 0.0149 & 0.6177 & 0.0198\\ 0.6177 & 0.0149 & 0.0071 & 0.0198\\ 0.0142 & 0.0085 & 0 & 0.0113\\ 0.0202 & 0.6336 & 0.0202 & 0.0202\\ 0.0202 & 0.6336 & 0.0202 & 0.0202\end{matrix}\end{cases} \cdots \Rightarrow R\begin{cases}\begin{matrix}0.0048 & 0.0095 & 0.6112 & 0.0127\\ 0.6115 & 0.0098 & 0.005 & 0.0129\\ 0.0092 & 0.0058 & 0.0005 & 0.0076\\ 0.0135 & 0.6217 & 0.0134 & 0.0138\\ 0.0133 & 0.6217 & 0.0136 & 0.0137\end{matrix}\end{cases}$$

Fig. 1 The framework of HGBI. A: A heterogeneous graph is constructed based on drug-drug similarities, target-target similarities, and drug-target interactions. B: Edge weights between drugs and targets are updated iteratively by incorporating all possible paths between each drug-target pair. C: For each drug, all candidate targets are ranked according to the final weights.

2. Materials and methods

2.1 Datasets collection

There are two intra-similarity matrices which represent the drug-drug similarities and target-target similarities, respectively. In addition, there is an interaction matrix, i.e. the drug-target interaction matrix, which represents the connections between drugs and targets. The drug-drug similarity matrix includes all the FDA-approved drugs from the DrugBank database[17]. The similarities are calculated based on their chemical structures. First, chemical structures of all drug compounds in the Canonical SMILES format[18] are downloaded from DrugBank[17]. Then, the Chemical Development Kit[19] is used to calculate a binary fingerprint for each drug. Finally, a similarity score of two drugs is calculated using Tanimoto score[20] based on their fingerprints, which is in the range of [0, 1]. A druggable gene is defined as a human protein coding gene that contributes to a disease phenotype and can be modified by a small molecule drug. The term “druggable genome” has been used to denote a list of computationally predicted genes that their proteins can serve as suitable targets for developing therapeutic drugs[21]. We use the term “druggable gene” and “target” interchangeably in this study. The list of druggable genes is downloaded from the Sophic Integrated Druggable Genome Database project[21], which includes genes from the ENSEMBL database[22], the DrugBank database[17] and the InterPro-BLAST database[23]. The target-target similarities are calculated using the Smith-Waterman algorithm[24] based on the amino acid sequences of their corresponding proteins. The similarities are normalized using the same method proposed by Bleakley and Yamanishi (2009)[11]. Initial drug-target interactions are collected from the DrugBank database[17], but limited to drugs that have associated diseases in the Online Mendelian Inheritance in Man (OMIM) database[25], which are the same as the one used in Ref[16]. For each drug-target interaction, their corresponding value in the drug-target interaction matrix is 1. All other items in the drug-target matrix are set to 0**.**

2.2 Basic statistics

The total number of drugs is 1409. The total number of targets is 3997. The matrix is very sparse with many isolated nodes (having no connections). The total number of connections among drugs and targets is only 2098, with 554 drugs having at least one known target and 602 targets connecting with at least one drug. Among the connected nodes, many of them have more than one connection, which means known information about drugs and targets is biased towards a very small subset of all drugs/targets. The degree distribution of each entity in the matrix is given in Fig. 2.

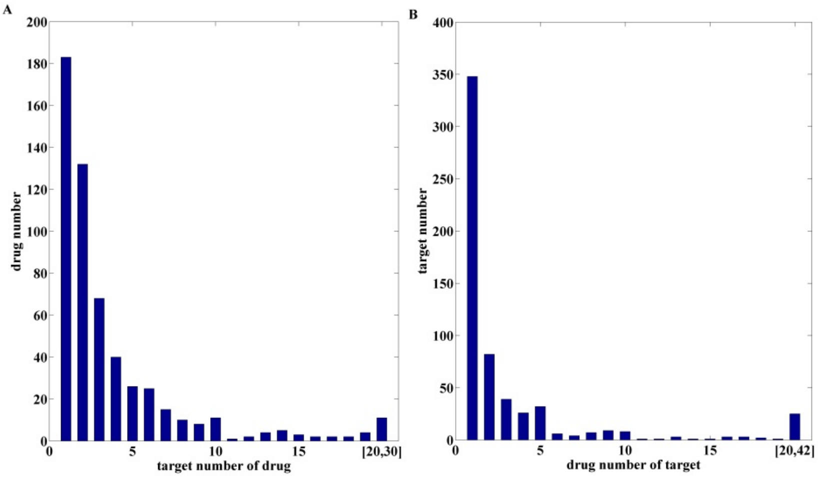


Fig. 2. Degree distributions of drugs (A) and targets (B) among the initial drug-target interactions.

2.3 Intra-similarity analysis

Before performing the proposed approach, we first study the statistical characteristics of these datasets. The distributions of the two intra-similarity matrices, i.e. the drug-drug and the target-target similarity matrices, are presented in Fig. 3A&C, which show the majority of the similarities are quite small. According to previous studies[26,27], low level similarity values provide little information for interaction inference. Furthermore, including the mess of low values could adversely affect prediction performance. Therefore, for the constructed heterogeneous graph, two nodes of the same type are connected only if their similarity score ≥ 0.3. It is also noted that although the entries between a node to itself have already been excluded, there are still some entries with value 1 (Fig. 3B&D), which is mainly due to the representation issue. To ensure that a node can only have a similarity score of 1 to itself, we replace those 1s with 0.99, which should not affect the final results much.

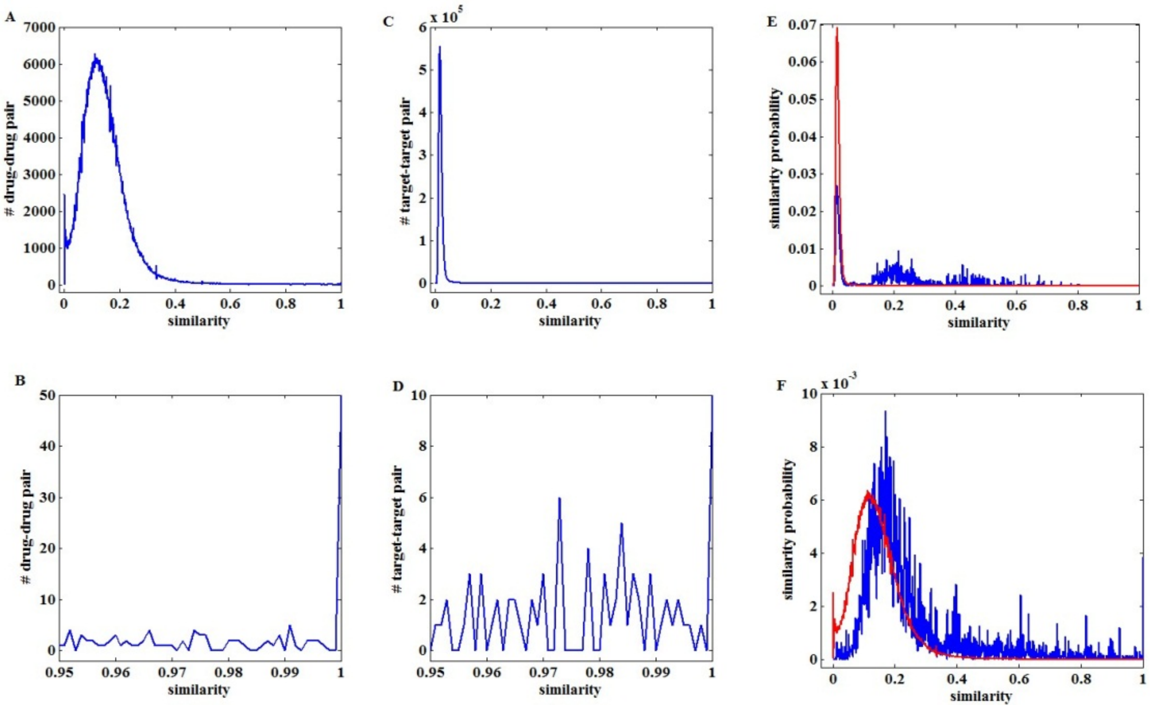


Fig. 3. Intra-similarity distributions. A: the distribution of drug-drug similarities; B: the tail distribution of drug-drug similarities in the range of (0.95, 1); C: the distribution of target-target similarities; D: the tail distribution of target-target similarities in the range of (0.95, 1); E: similarity distributions of targets sharing the same drugs (blue curve), and from different drugs (red curve); F: similarity distributions of drugs sharing the same targets (blue curve), and from different targets (red curve).

This study is based on the assumption that similar drugs tend to be associated with similar targets and dissimilar drugs are prone to be associated with dissimilar targets. To study the validation of the assumption on the collected real datasets, similarities of drugs for the same targets and similarities of drugs from different targets are compared. The average similarity of drugs for the same targets is calculated by averaging the similarities of all drug pairs that belong to the same target. To determine the average similarity of drugs from different targets, the similarity values of all drug pairs that are across different targets are averaged. Similarly, we examine the similarities among targets that share the same drugs and similarities among targets that do not share any drugs. The results are given in Fig. 3E&F. The average similarity of drug pairs from the same targets and the average similarity of target pairs from the same drugs are 0.2445 and 0.1836, respectively. In contrast, the average similarity of drug pairs from different targets and the average similarity of target pairs from different drugs are 0.1429 and 0.0231, respectively. We further test the differences of the corresponding distributions using the Wilcoxon rank sum test. Both tests reject the null hypothesis that the distributions are the same at the 5% significance level. Based on these results, it can be concluded that drugs (targets) associated with the same targets (drugs) possess higher similarity values than those associated with different targets (drugs). The guilt-by-association principle can be utilized in this study.

2.4 Drug-target association predictions

The drug-target heterogeneous graph has two kinds of nodes: drug nodes and target nodes. Let $R=\{R_1,R_2,\ldots,R_n\}$ denote the *n* drug nodes, and $T=\{T_1,T_2,\ldots,T_m\}$ denote the *m* target nodes. A drug is connected with another drug if and only if their similarity is greater than a pre-defined threshold (0.3 in this study), which is assigned as the weight of the edge. Edges and weights for target pairs are constructed similarly. Finally, a drug and a target are connected if they interact in the original drug-target interaction dataset. The weights of all drug-target edges are originally assigned 1. Let E_{rr}, E_{tt}, and E_{rt} represent drug-drug, target-target and drug-target edges, respectively, and W_{rr}, W_{tt}, and W_{rt} represent the weights on these three kinds of edges. The heterogeneous drug-target graph can be represented as $G_{RT}=\{\{R,T\},\{E_{rr},E_{tt},E_{rt},\},\{W_{rr},W_{tt},W_{rt},\}\}$. Based on this graph, the novel target prediction problem can be transformed into a novel drug-target edge prediction problem on the constructed drug-target graph. This means that the original heterogeneous graph G_{RT} can be considered as an incomplete graph with missing edges between *R* (drug) nodes and *T* (target) nodes. The objective is to capture hidden interactions between drugs and targets based on the drug-drug similarities, target-target similarities, and known drug-target interactions. The novel drug-target edge prediction problem can be formalized as follows:
Input: $G_{RT}=\{\{R,T\},\{E_{rr},E_{tt},E_{rt},\},\{W_{rr},W_{tt},W_{rt},\}\}$
Output: $G_{RT}^{new}=\{\{R,T\},\{E_{rr},E_{tt},E_{rt}^{new},\},\{W_{rr},W_{tt},W_{rt}^{new},\}\}$
where E_{rt}^{new} and W_{rt}^{new} represent the newly calculated edges and their weights respectively.

Based on the guilt-by-association assumption, the intra-similarities and drug-target associations can be combined together to predict novel interactions between drugs and targets. For example, given the graph G_{RT}, one way to calculate the association coefficient (i.e., weight) between each initially unconnected drug-target pair is based on the following equation,

$$w(r,t)=\sum_{r_i\in R}\sum_{t_j\in T} w(r,r_i)\times w(r_i,t_j)\times w(t,t_j) \qquad (1)$$

Here *r* is a drug and *t* is a target and they are not connected in the original graph. r_i's and t_j's are the neighbors of *r* and *t* that are connected with each other in G_{RT}. $w(r, r_i)$ is the weight between *r* and r_i, and $w(t, t_j)$ is the weight between *t* and t_j. Equation 1 basically means that one can establish a new weight between a drug and a target by summarizing all paths of length three, consisting one edge in each of E_{rr},E_{rt},E_{tt}. This is essentially the same idea adopted by NBI [13].

Naturally, once new relationships/weights between drugs and targets being established based on equation 1, they themselves can be utilized again to generate more relationships. An iterative procedure can be constructed, which can be represented as matrix multiplications: $W_{rt}^{i+1}=W_{rr}\times W_{rt}^{i}\times W_{tt}$. In general, there are two related issues that need to be resolved in order to make the proposed iterative approach to work. First, one may want to treat the initial links between drugs and targets differently from the inferred links because the initial links deserve more credibility. Second, it is desirable if the matrix W_{rt} will converge, which means that the information propagation is stabilized at the end. In this study, we propose an iterative approach based on equation 2, which naturally solves the first problem based on its formulation. Furthermore, we show that with proper normalization, it also solves the second problem.

$$W_{rt}^{i+1} = \alpha W_{rr} \times W_{rt}^{i} \times W_{tt} + (1-\alpha) W_{rt}^{0} \tag{2}$$

In this formula, $W_{rt}^{0} = W_{rt}$, represents the initial interactions between drugs and targets, α is a decay factor with its value between 0 and 1. In each iteration, the original drug-target interactions will contribute to the newly constructed interactions, and the contribution is controlled by the scale factor 1 - α. Theoretically, one can optimize α based on results from cross validations. In this study, we fix α = 0.4. By iteratively using this formula, the strength between a drug and a target will eventually include all the possible paths connecting them in the heterogeneous graph. We prove that when W_{rr} and W_{tt} are properly normalized, it is guaranteed that equation 2 will converge. The result is summarized as Theorem 1 and the details of the proof can be found in the appendix. To obtain the final solution based on equation 2, we use an iterative propagation-based algorithm[28]. Once the final result is given, for each drug, all the targets will be ranked according to the strength of their links to the drug.

THEOREM 1. When W_{rr} and W_{tt} are properly normalized utilizing equation 3, it is guaranteed that formula (2) will converge.

$$w(r_i, r_j) = \frac{w(r_i, r_j)}{\sqrt{\sum_{k=1}^{n} w(r_i, r_k) \sum_{k=1}^{n} w(r_k, r_j)}}, w(t_i, t_j) = \frac{w(t_i, t_j)}{\sqrt{\sum_{k=1}^{m} w(t_i, t_k) \sum_{k=1}^{m} w(t_k, t_j)}} \tag{3}$$

3 Experiments

3.1 Evaluation metrics

In order to systematically evaluate the proposed approach on the collected datasets, we adopt a leave-one-out cross-validation (LOOCV) strategy in our experiments. For each drug, one of its connections to a target is treated as the test data, and it is ranked with all other targets in descending order according to the calculated drug-target association coefficients using the remaining connections as training data. For each specific ranking threshold, if the rank of the testing connection is above the threshold, it is regarded as a true positive. On the other hand, if the rank of an unknown connection is above the threshold, it is regarded as a false positive. True positive rate (TPR) and false positive rate (FPR) are calculated by varying thresholds to construct the ROC curve[29]. The area under the curve (AUC) value represents the overall performance of the algorithm. In addition to LOOCV, we also perform 10-fold cross-validation, where all the drug-target connections are randomly partitioned into 10 subsets and each subset is treated as the test set in each iteration. Furthermore, in practice, it is natural that most researchers only focus on top ranked targets. Therefore, we also examine the performance of the algorithm on the top ranked results, i.e., the number of correctly retrieved connections based on various top percentiles (the most left side of the ROC curve). In addition, to test the capacity of the algorithm in detecting novel interactions for drugs with no known targets, we collect all drugs that only have a single known target and perform the experiment by removing the only interaction. Finally, using all the data as training data, we test our algorithm again and compare the top ranked targets with those in another database [30] that is not used in training.

3.2 Comparison with existing methods

To evaluate the proposed approach, we choose to compare its performance with BLM[11] and NBI[13]. BLM is considered one of the state-of-the-art approaches in drug-target interaction predictions. In this study, BLM is implemented the same way as the one in the original paper[11]. The predicted scores generated from SVM are used as the ranking criterion, which means that larger predicted scores yield higher ranks. In order to choose a proper number of negative samples for SVM training of BLM, we perform cross-validation. Based on the results (Fig S1 in appendix), the number of negative training samples was set to be max{20, 2 x *num_positive_samples*}. The result of BLM is obtained by averaging five runs with the same configuration but different negative training samples. We choose to compare with NBI[13] because it can be viewed as a simplified version of the proposed approach, in the sense that only a two-step diffusion of the matrices (similar to equation 1) is used in NBI, while our approach uses the converged matrix. NBI[13] is implemented according to the original paper.

3.3 Experimental results

3.3.1 Predictions for drugs with known connected targets

The cross-validation experiments for target predictions were conducted using all drugs with at least two known targets. In total, 371 such drugs and 1915 initial drug-target edges were considered. The ROC curves and AUC values of NBI, BLM, HGBI (LOOCV and 10-fold) are given in Fig.4A. It shows that HGBI (AUC:0.93) significantly outperforms both BLM (AUC:0.89) and NBI (AUC:0.73) for LOOCV. Furthermore, HGBI almost has the same performance when using 10-fold cross-validation. The numbers of correctly retrieved drug-target interactions according to different ranking thresholds are also given in Fig.4B. Results show that when focusing on the top ranked results, the performance of HGBI is much better compared with NBI and BLM, especially for the top 1% ranked targets, in which case HGBI correctly retrieved 1339 drug-target interactions, whereas BLM and NBI only retrieved 56 and 10 such interactions.

3.3.2 Predictions for drugs without known connected targets

To demonstrate the effectiveness of the proposed approach in detecting novel targets for drugs without known targets, only drugs with exactly one connected target in the dataset were collected in this experiment. There are in total 183 such drugs. Because BLM cannot predict novel targets for drugs without known targets, we only compare HGBI with NBI here. The ROC curves of NBI and HGBI are given in Figure 5. Again, HGBI (AUC:0.93) achieves much better performance than NBI (AUC:0.72).

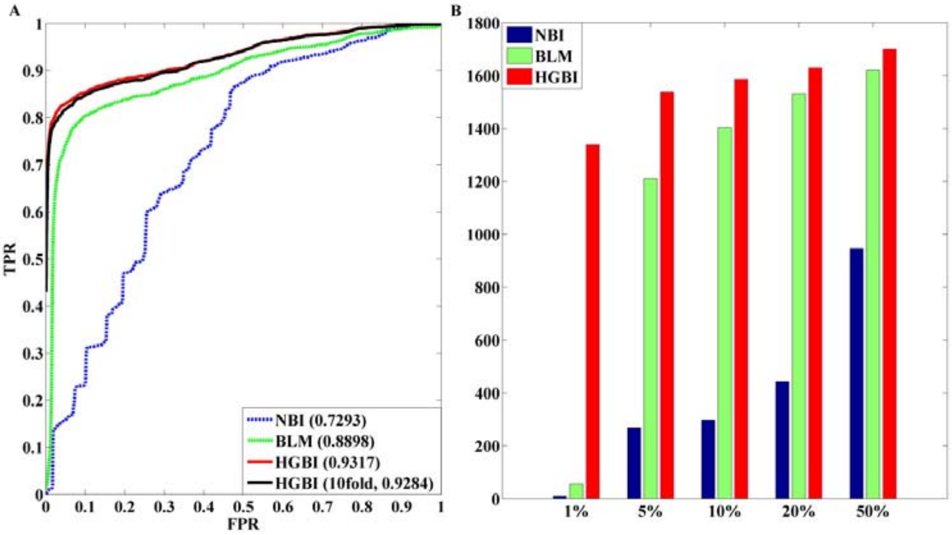


Fig.4 A: ROC curves of drug-target association predictions. AUC of each curve is indicated in the parentheses; B: The number of retrieved drug-target interactions using different thresholds. *x-axis* is the ranking thresholds in percentile.

3.3.3 Case studies on drug-target association predictions

Finally, using all the data as the training data, HBGI can make new predictions for all the drugs in the database. To further analyze its performance for practical usage, six drugs, i.e. Citalopram (Drugbank ID: DB00215), Escitalopram (DB01175), Terfenadine (DB00342), Diphenidol (DB01231), Fexofenadine (DB00950), and Naltrexone (DB00704), were randomly chosen for the case studies. For each drug, all their initial targets and the top 10 predicted targets were collected. A subset of these drugs, targets, and their connections are also illustrated in Fig 5B, which only shows upto 3 known targets and the top 3 predicted targets for clarity. Several observations can be made based on Fig 5B. First, similar drugs tend to share similar predicted targets, such as the drugs Diphenidol and Terfenadine. Second, predictions for drugs without known connected targets, such as the drug Fexofenadine, can be performed using HBGI. Because it is connected with the drug Diphenidol, one of Diphenidol's targets (target Entrez_ID: 1128 in Fig 5B) is predicted to be associated with it.

We further searched the Supertarget database[30], which is an extensive web resource for analyzing drug-target interactions. Some of the top ranked predictions by HBGI are supported by newly reported discoveries in the database. For example, in the Supertarget database, Citalopram (DB00215) has two new targets SLC6A3 (6531) and SLC6A2 (6530) that were not in the DrugBank database. They were ranked as the 2nd and the 17th among all target candidates by HBGI. Similarly, ADRA1D (146) & CHRM1 (1128) were not associated with Terfenadine (DB00342) in the DrugBank database. They were ranked as the 3rd & the 8th, respectively.

In a very recent study[31], the authors experimentally validated 123 unique drug-target relationships. In comparison of our prediction with these newly validated relationships, our median rank for this data set is 16 (out of 3997 targets) and 43 of them are ranked in top 10.

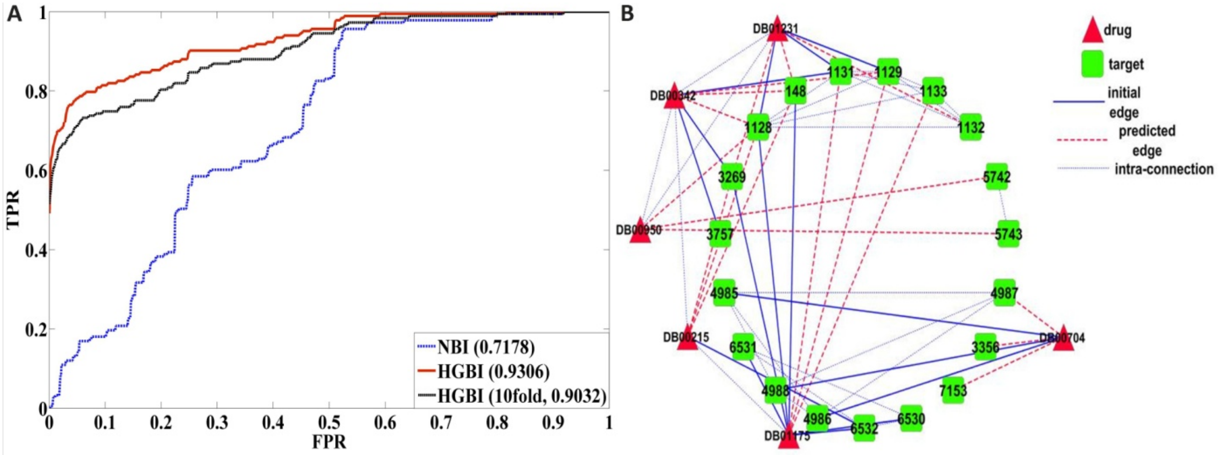


Fig.5 A. ROC curves of novel target predictions for drugs without known targets. B. The subsetwork for the case study.

4 Conclusion

In this paper, we have proposed a drug target prediction approach, HGBI, which integrates drug-drug similarities, target-target similarities and drug-target interactions into a heterogeneous graph and models the drug-target interactions as the stabilized information flow between them across the heterogeneous graph. Experiments have shown that HGBI significantly outperforms two existing methods in predicting novel targets for drugs with and without known targets. A case study has illustrated that HGBI can be used in practice to rank candidate targets for drugs and many top ranked ones can be utilized for further investigations.

Although equation 2 is similar to the framework of random walk with restart (RWWR)[32], it is different in the sense that 1) equation 2 is defined on a heterogeneous graph and only connections between nodes of different types need to be derived; 2) because of the heterogeneous graph, all the information on similar drugs, similar targets and drug-target interactions has been used in predicting new drug-target associations.

For future directions, first, instead of using top ranked targets, it is possible to adopt an automatic threshold to declare significant predictions using the same idea in Ref[33]. Furthermore, many approaches have been developed to identify disease genes. However, the relationships between disesae genes and drug targets are not totally characterised. It would be interesting to include disease gene information in drug target predictions.

5. Acknowledgements

This research is supported by an NIH/NLM grant (LM008991).

References

1. P. Imming, *et al.*, *Nat Rev Drug Discov.* **10**, 821 (2006).
2. S. Haggarty, *et al.*, *Chem Biol.* **5**, 383 (2003).
3. A. Hopkins and C. Groom, *Nat Rev Drug Discov.* **9**, 727 (2002).
4. A. Russ and S. Lampel, *Drug Discov Today.* **23-24**, 1607 (2005).
5. S. Zhu, *et al.*, *Bioinformatics*, **suppl 2**, ii245 (2005).
6. A. Cheng, *et al.*, *Nat Biotechnol*, **1**, 71 (2007).

7. M. Campillos, *et al.*, *Science*, **5886**, 263 (2008).
8. Y. Yamanishi and M. Araki, *et al.*, *Bioinformatics,* **13**, i232 (2008).
9. Y. Yamanishi, *Proceedings of NIPS,* **21**, 1433 (2008)..
10. M. Keiser, *et al.*, *Nature,* **7270**, 175 (2009).
11. K. Bleakley and Y. Yamanishi, *Bioinformatics,* **18**, 2397 (2009).
12. L. Perlman, *et al.*, *J Comput Bio,* **2**, 133 (2011).
13. F. Cheng, *et al.*, *PLoS Comput Biol,* **5**, e1002503 (2012).
14. G. Jeh and J. Widom, *KDD,* 538 (2002).
15. A. Chiang, and A. Butte, *Clin Pharmacol Ther,* **5**, 507 (2009).
16. A. Gottlieb, *et al.*, *Mol Syst Biol*, **496**, 496 (2011).
17. C. Knox, *et al.*, *Nucleic Acids Res*, **Database issue**, 1035 (2011).
18. D. Weininger, *Journal of Chemical Information and Modeling*, **1**, 31 (1988).
19. C. Steinbeck, *et al.*, *Curr Pharm Des,* **17**, 2111 (2006).
20. T. Tanimoto, *IBM Internal Report 17th Nov*, (1957).
21. Sophic,http://www.sophicalliance.com/documents/sophicdocs/White%20Paper%20Update%201-27-11/The%20Druggable%20Genome012511.pdf, (2012).
22. P. Flicek, *et al.*, *Nucleic Acids Res*, **Database issue**, 800 (2011).
23. S. Hunter, *et al.*, *Nucleic Acids Res*, **Database issue**, 211 (2009).
24. T. Smith and M. Waterman, *J Mol Biol*, **1**, 195 (1981).
25. A. Hamosh *et al., Nucleic Acids Res*, **30**, 52 (2002).
26. Y. Chen, *et al.*, *Bioinformatics*, **13**, i167 (2011).
27. M. van Driel, *et al.*, *Eur J Hum Genet*, **5**, 535 (2006).
28. V. Oron, *et al.*, *PLoS Comput Biol*, **1**, e1000641 (2010).
29. T. Sing, *et al.*, *Bioinformatics*, **20**, 3940 (2005).
30. N. Hecker, *et al.*, *Nucleic Acids Res,* **Database issue**, D1113 (2012).
31. E. Lounkine, *et al.*, *Nature,* **486**, 361 (2012).
32. Tong, *et al.*, *Proceedings of ICDM,* 613-622(2005).
33. Chen, *et al.*, *PLoS One*, **6**, e21137 (2011).

6. Appendix

PROOF of Theorem 1: To make the proof process clear, Let *A*, *B* and *X* denote W_{rr}, W_{tt}, and W_{rt} respectively. *A*, *B* and *X* are *n*×*n*, *m*×*m* and *n*×*m* matrices respectively. A_i and A^j denote the *i*-th row of *A* and *j*-th column of *A* respectively. a_{ij} is used to represent the value at the *i*-th row and *j*-th column of matrix *A*. These conventions are also used for matrix *B* and *X*.

Then according to equation (2), we have $x_{ij} = \alpha A_i X B^j + (1-\alpha) x_{ij}^0$. For X^1, we can also get

$$\begin{bmatrix} x_{1,1} \\ \vdots \\ x_{n,1} \end{bmatrix} = \alpha \begin{bmatrix} a_{1,1}b_{1,1}, \cdots, a_{1,n}b_{1,1}, \cdots, a_{1,1}b_{m,1}, \cdots, a_{1,n}b_{m,1} \\ \vdots \\ a_{n,1}b_{1,1}, \cdots, a_{n,n}b_{1,1}, \cdots, a_{n,1}b_{m,1}, \cdots, a_{n,n}b_{m,1} \end{bmatrix} \begin{bmatrix} x_{1,1} \\ \vdots \\ x_{n,1} \\ \vdots \\ x_{1,m} \\ \vdots \\ x_{n,m} \end{bmatrix} + (1-\alpha) \begin{bmatrix} x_{1,1}^0 \\ \vdots \\ x_{n,1}^0 \end{bmatrix}$$

If we use $A_i \times B^j$ to denote $\begin{bmatrix} a_{i,1}b_{1,j} & \cdots & a_{i,n}b_{1,j} & \cdots & a_{i,1}b_{m,j} & \cdots & a_{i,n}b_{m,j} \end{bmatrix}$, then equation (2) can be written as

$$\begin{bmatrix} X^1 \\ \vdots \\ X^m \end{bmatrix} = \alpha \begin{bmatrix} A_1 \times B^1 \\ \vdots \\ A_n \times B^1 \\ \vdots \\ A_1 \times B^m \\ \vdots \\ A_n \times B^m \end{bmatrix} \begin{bmatrix} X^1 \\ \vdots \\ X^m \end{bmatrix} + (1-\alpha) \begin{bmatrix} X^{10} \\ \vdots \\ X^{m0} \end{bmatrix} \tag{4}$$

Let *C* denote $\begin{bmatrix} A_1 \times B^1 & \cdots & A_n \times B^1 & \cdots & A_1 \times B^n & \cdots & A_n \times B^m \end{bmatrix}^T$ and

$i = sn + t$, $j = rn + \theta$, $s = sI\{t > 0\} + (s-1)I\{t = 0\}$, $t = tI\{t > 0\} + nI\{t = 0\}$, $r = rI\{\theta > 0\} + (r-1)I\{\theta = 0\}$, $\theta = \theta I\{\theta > 0\} + nI\{\theta = 0\}$, $0 \le t, \theta < n$.

Then we get: $c_{i,j} = a_{t,\theta} b_{r+1,s+1}$ and $c_{j,i} = a_{\theta,t} b_{s+1,r+1}$.

By comparing the above two equations, we can easily find that *C* is a symmetrical matrix with row and column number *n×m*. If we use *X** to represents $\begin{bmatrix} X^1 & \cdots & X^n \end{bmatrix}^T$, then equation (4) can also be written as:

$$X^* = \alpha C X^* + (1-\alpha) X^{*0} \tag{5}$$

According to (Vanunu, et al. 2010)[29], in order to get a converged solution for equation (5), *C* is normalized as: $C^{norm} = D^{-1/2} C D^{-1/2}$, where *D* is diagonal matrix with $d_{i,i}$ equals to the sum of the *i*-th row of *C*. Therefore, we also have $c_{i,j}^{norm} = \frac{c_{i,j}}{\sqrt{d_{i,i} d_{j,j}}}$ and $d_{i,i} = \sum_{u=1}^{nm} c_{i,u} = \sum_{u=1}^{nm} a_{t,\theta_u} b_{r_u+1,s+1} = \sum_{p=1}^{n} a_{t,p} \sum_{q=1}^{m} b_{q,s+1}$

where $u = r_u n + \theta_u$. By incorporating the above equation into $c_{i,j}^{norm}$, we can get

$$c_{i,j}^{norm} = \frac{a_{t,\theta}, b_{r+1,s+1}}{\sqrt{\sum_{p=1}^{n} a_{t,p} \sum_{q=1}^{m} b_{q,s+1}} \sqrt{\sum_{p=1}^{n} a_{\theta,p} \sum_{q=1}^{m} b_{q,r+1}}} = \frac{a_{t,\theta}}{\sqrt{\sum_{p=1}^{n} a_{t,p} \sum_{p=1}^{n} a_{\theta,p}}} \frac{b_{r+1,s+1}}{\sqrt{\sum_{q=1}^{m} b_{q,s+1} \sum_{q=1}^{m} b_{q,r+1}}}$$

Therefore, if we normalize *A* and *B* as $a_{i,j}^{norm} = \frac{a_{i,j}}{\sqrt{\sum_{p=1}^{n} a_{i,p} \sum_{p=1}^{n} a_{j,p}}}$ and $b_{i,j}^{norm} = \frac{b_{i,j}}{\sqrt{\sum_{q=1}^{n} b_{q,i} \sum_{q=1}^{n} b_{q,j}}}$

We can get $c_{i,j}^{norm} = a_{t,\theta}^{norm} b_{r+1,s+1}^{norm}$.

With this equation, we can rewrite equation (5) as $X^* = \alpha C^{norm} X^* + (1-\alpha) X^{*0}$

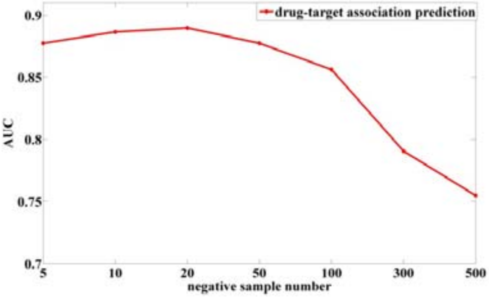


Figure S1 Cross-validation results of BLM using different numbers of negative samples.

EPIGENOMICS

A. J. HARTEMINK

Department of Computer Science, Box 90129
Duke University
Durham, NC 27708-0129, USA
Email: amink@cs.duke.edu

M. KELLIS

Computer Science and Artificial Intelligence Laboratory
Broad Institute of MIT and Harvard
Stata Center - 32D.524 Cambridge, MA 02142, USA
E-mail: manoli@mit.edu

W. S. NOBLE

Department of Genome Sciences, Box 355065
University of Washington
Seattle, WA 98109, USA
E-mail: william-noble@uw.edu

Z. WENG

Biochemistry & Molecular Pharmacology
364 Plantation Street, LRB
University of Massachusetts Medical School
Worcester, MA 01605, USA
E-mail: Zhiping.Weng@umassmed.edu

Epigenomics involves the global study of mechanisms, such as histone modifications or DNA methylation, that have an impact on development or phenotype, are heritable, but are not directly encoded in the DNA sequence. The recent availability of large epigenomic data sets, coupled with the increasing recognition of the importance of epigenetic phenomena, has spurred a growing interest in computational methods for interpreting the epigenome.

Keywords: Epigenomics, histone modifications, chromatin, DNA methylation

Scientists have known for a long time that the sequence of nucleotides that comprise the genome is not sufficient to explain the heritability of traits from one generation to the next, nor is that sequence sufficient to drive the myriad functions of a living cell. Recently, however, catalyzed by the rapid acquisition of a wide variety of genome-scale data sets from projects such as ENCODE,[1] modENCODE,[2] and Roadmap Epigenomics,[?] scientists have begun to characterize just how much information is encoded beyond the primary DNA sequence. Accordingly, many of the central questions facing biology today concern the interpretation and integration of epigenomic data with our existing knowledge of the molecular pathways within the cell, including DNA, RNA, proteins, and metabolites. This session includes three papers, each of which describes a novel computational method for the analysis and interpretation of one or more types of epigenomic data.

The first paper analyzes a single type of data, derived from a DNase 1 sensitivity assay. The endonuclease DNase 1 has long been known to preferentially cleave in short regions of open chromatin, known as DNase 1 hypersensitive sites.[3] Such regions are of great interest because they correspond to various types of regulatory elements, including promoters, enhancers, insulators and boundary elements. Recently, a series of DNase 1-based assays have been described for ascertaining the cleavage profile of DNase 1 across the entire genome. Originally based on quantitative PCR[4] and microarrays,[5,6] these assays were quickly adapted for next-generation sequencing platforms.[7,8] Importantly, in addition to recognizing classical hypersensitive sites, which have a typical size of 225–250 bp, subsequent work demonstrated that a detailed DNase 1 cleavage profile could localize protein-binding events at basepair resolution.[9,10]

Given the importance of transcription factor binding for gene regulation, and given the increasing availability of DNase 1 data for a wide variety of human cell types, a variety of computational methods have been developed to interpret DNase 1 sensitivity data. Luo and Hartemink contribute to this literature by introducing a method, called Millipede, that aims to identify transcription factor binding events on the basis of DNase 1 sensitivity data as well as analysis of the primary sequence. Millipede improves upon the previously described Centipede algorithm[11] by reducing the number of parameters and switching from unsupervised to supervised learning. Luo and Hartemink benchmark Millipede using data from human and yeast.

The second paper, by Sahu et al., proposes a machine learning approach to enhancer detection. An enhancer is a gene regulatory element that is responsible for upregulating one or more genes. Enhancers are notoriously difficult to detect because they often do not occur proximal to their target gene, relying instead upon DNA looping or other complex chromatin structures to carry out their regulatory effect. No single high-throughput assay can be used to identify the "enhancerome" because different types of enhancers presumably rely upon different regulatory mechanisms. The gold standard method for identifying an enhancer involves knocking it out and observing the resulting downregulation of the target gene. This approach, obviously, does not scale to whole-genome analysis. Currently, closest proxy we have for genome-wide enhancer detection is ChIP-seq for the DNA-binding protein p300. Although almost all p300 binding sites are enhancers, many known enhancers are not bound by p300.

This lack of a high-quality and high-throughput enhancer assay has led to the development of a series of computational methods that aim to identify putative enhancers.[12–15] Sahu et al. contribute to this ongoing project by introducing a support vector machine classifier that learns to identify enhancers on the basis of ChIP-seq histone modification and DNase 1 sensitivity data. They demonstrate that, not only does their classifier perform well in cross-validation, but it also can be used to identify putative enhancers associated with SNPs from genome-wide association studies of cardiac phenotypes.

Finally, the paper by Ahn and Wang describes a statistical testing methodology for identifying genomic regions in which patterns of variability in DNA methylation across individuals may be indicative of disease. DNA methylation involves the addition of a methyl group either to an adenine or (most commonly in animals) a cytosine. Methylation is used extensively by the cell to shut off expression of individual genes or large chromosomal regions, and plays a critical role in regulating cellular processes such as embryonic development, X chromosome inactivation, genomic imprinting and chromosome stability.[16] Methylated cytosines can be identified by first subjecting the DNA to bisulfite conversion, which changes cytosine residues to uracil unless the cytosines are methylated, and then sequencing the converted DNA. The result, by comparison to a reference genome, is a map of the frequency of methylation at each cytosine residue. Methylation is associated with a set of heritable syndromes—imprinting disorders—that result from asymmetric expression of the alleles of one or more genes, as well as with a variety of repeat-instability diseases.[16] More recently, aberrant methylation has been increasingly implicated in various types of cancer.[17]

The primary goals of Ahn and Wang's work is to improve our ability to detect patterns of aberrant methylation that are potentially associated with a given disease. Their proposed statistical framework draws upon the observation that such loci differ not only in the mean level of methylation but also its variance. Accordingly, Ahn and Wang propose a regression-based testing framework that captures more features of the methylation profile of a given locus and, in so doing, boosts statistical power relative to approaches based only on the mean.

The topics covered by these three papers are quite diverse, reflecting the wide range of challenging computational and statistical problems posed by epigenomic data.

References

1. ENCODE Project Consortium, *Nature* **489**, 57 (2012).
2. T. modENCODE Consortium, *Science* **330**, 1775 (2010).
3. C. Wu, *Nature* **286**, 854 (1980).
4. P. J. Sabo, R. Humbert, M. Hawrylycz, J. C. Wallace, M. O. Dorschner, M. McArthur and J. A. Stamatoyannopoulos, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4537 (2004).
5. P. J. Sabo, M. S. Kuehn, R. Thurman, C. Grant, B. Johnson, S. Johnson, H. Kao, M. Yu, J. Goldy, M. Weaver, M. A. Singer, T. Richmond, M. Dorschner, P. Navas, R. Green, W. S. Noble and J. A. Stamatoyannopoulos., *Nature Methods* **3**, 511 (2006).
6. G. E. Crawford, S. David, P. C. Scacheri, G. Renaud, M. J. Halawi, M. R. Erdos, R. Green, P. S. Meltzer, T. G. Wolfsberg and F. S. Collins, *Nature Methods* **3**, 503 (2006).
7. P. J. Sabo, M. Hawrylycz, J. C. Wallace, R. Humbert, M. Yu, A. Shafer, J. Kawamoto, R. Hall, J. Mack, M. O. Dorschner, M. McArthur, and J. A. Stamatoyannopoulos, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 16837 (2004).
8. A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey and G. E. Crawford, *Cell* **132**, 311 (Jan 2008).
9. J. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields and J. A. Stamatoyannopoulos, *Nature Methods* **6**, 283 (2009).
10. A. P. Boyle, L. Song, B. K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford and T. S. Furey, *Genome Research* **21**, 456 (2011).
11. R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad and J. K. Pritchard, *Genome Research* **21**, 447 (2011).
12. D. Lee, R. Karchin and M. A. Beer, *Genome Research* **21**, 2167 (2011).
13. M. Fernandez and D. Miranda-Saavedra, *Nucleic Acids Research* **40**, p. e77 (2012).
14. D. May, M. J. Blow, T. Kaplan, D. J. McCulley, B. C. Jense, J. A. Akiyama, A. Holt, I. Plajzer-

Frick, M. Shoukry, C. Wright, V. Afzal, P. C. Simpson, E. M. Rubin, B. L. Black, J. Bristow, L. E. Pennacchio and A. Visel, *Nature Genetics* **44**, 89 (2011).

15. L. Narlikar, N. J. Sakabe, A. A. Blanski, F. E. Arimura, J. M. Westlund, M. A. Nobrega and I. Ovcharenko, *Genome Research* **20**, 381 (2010).
16. K. D. Robertson, *Nature Reviews Genetics* **6**, 597 (2005).
17. M. A. Dawson and T. Kouzarides, *Cell* **150**, 12 (2012).

A POWERFUL STATISTICAL METHOD FOR IDENTIFYING DIFFERENTIALLY METHYLATED MARKERS IN COMPLEX DISEASES

SURIN AHN

Email: surin.ahn@gmail.com

TAO WANG[†]

Department of Epidemiology and Population Health, Albert Einstein College of Medicine of Yeshiva,
1300 Morris Park Ave, Bronx, NY 10461
Email: tao.wang@einstein.yu.edu

DNA methylation is an important epigenetic modification that regulates transcriptional expression and plays an important role in complex diseases, such as cancer. Genome-wide methylation patterns have unique features and hence require the development of new analytic approaches. One important feature is that methylation levels in disease tissues often differ from those in normal tissues with respect to both average and variability. In this paper, we propose a new score test to identify methylation markers of disease. This approach simultaneously utilizes information from the first and second moments of methylation distribution to improve statistical efficiency. Because the proposed score test is derived from a generalized regression model, it can be used for analyzing both categorical and continuous disease phenotypes, and for adjusting for covariates. We evaluate the performance of the proposed method and compare it to other tests including the most commonly-used t-test through simulations. The simulation results show that the validity of the proposed method is robust to departures from the normal assumption of methylation levels and can be substantially more powerful than the t-test in the presence of heterogeneity of methylation variability between disease and normal tissues. We demonstrate our approach by analyzing the methylation dataset of an ovarian cancer study and identify novel methylation loci not identified by the t-test.

[†] This work is supported in part by the CTSA Grant UL1 RR025750 and KL2 RR025749 and TL1 RR025748 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH) and NIH roadmap for Medical Research, R21HG006150 from National Human Genome Research Institute (NHGRI).

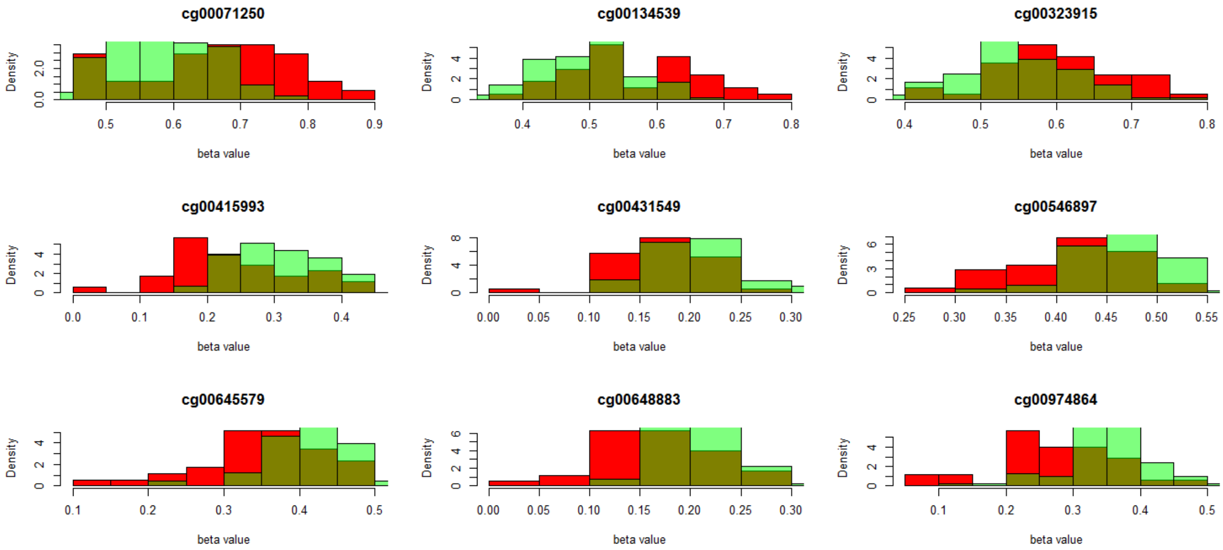
1. Introduction

DNA methylation is an important epigenetic modification that regulates transcriptional expression and plays an important role in complex diseases including cancer. [1] Recently, tremendous amounts of DNA methylation data have been generated from high-throughput DNA methylation platforms for many complex diseases. Compared to patterns of other molecular profiling, e.g. gene expression, DNA methylation has unique features. One example is that not only the mean but also the standard deviation of methylation levels can vary across age groups.[2,3] New statistical approaches designed to incorporate these features are desirable because they could be more robust and efficient than conventional methods. As such, Chen et al. proposed a test to evaluate the overall statistical significance of association by combining p-values from different age groups and showed it was more robust and usually more powerful than existing tests. [4]

Another phenomenon that has recently received attention is the increased methylation variability at relevant loci of cancer. [5-7] It has been found that differential variability between normal and cancer tissues can be very useful for identifying methylation markers of cancer [6-8] However, commonly-used statistical methods, such as the t-test and linear regression, which do not directly detect differences in variability, are statistically inefficient in the presence of heterogeneity of methylation variability. In the statistical literature, various tests, e.g. the Barlett's test[9] and the Levene's test[10], have been proposed for testing homogeneity of variance between two groups. In general, the Levene's test is less sensitive than the Bartlett's test to departures from normality. Figure 1 shows methylation distributions of several representative loci in cancer and normal tissues from an ovarian cancer study.[3] One important feature of these loci is both the mean and variability of methylation levels are different between cancer and normal tissues. For these loci, it may be useful to combine information from both the first and second moments of methylation distribution to improve power to identify methylation markers. One approach to combine the results for testing mean and variability is Fisher's method of combining p-values. However, it requires that the mean and variance are independent, which is often not true for methylation data. Another approach is to use tests, e.g. Kolmogorov-Smirnov test, to compare the empirical distribution of methylation data, which, however, is often not statistically efficient[11].

In this article, we propose a new statistical test that incorporates changes in both mean and variability to identify methylation markers of diseases, and demonstrate how jointly testing the mean and variability can identify methylation markers that are otherwise missed by testing the mean alone. More specifically, we first define two score tests for testing methylation differences in mean and variability, respectively, under a generalized regression model. Then, we develop a new joint test by combining these two statistics, while accounting for their correlation. As such, the new test may not require intensive sampling approaches to evaluate p-values. We evaluate the performance of the proposed approach and compare it to the conventional tests including the commonly-used two-sample t-test through simulations. We show that the validity of the proposed test is robust to departures of the normal distribution of methylation levels and can be substantially more powerful than the t-test in the presence of heterogeneity of variability between two groups. Finally, we apply our approach to the methylation data of an ovarian cancer study and identify cancer relevant loci that other tests could fail to identify.

Fig1: Histograms of DNA methylation values of pretreatment cancers and control groups at 9 selected methylation loci. Red bars represent cancers and green bars represent controls.



2. Methods

We consider detecting the association of individual methylation loci with disease based on a case-control study. For individual i $(i = 1, 2, ..., n)$, the trait value is denoted as Y_i, and the methylation value is denoted as X_i. To identify methylation loci that are relevant to disease, we consider the statistical hypothesis $H_0 : \mu_0 = \mu_1 \text{ and } \sigma_0^2 = \sigma_1^2$ versus $H_1 : \mu_0 \neq \mu_1 \text{ and } \sigma_0^2 \neq \sigma_1^2$, in which μ_0 and μ_1 are means of methylation levels for controls and cases, respectively, and σ_0^2 and σ_1^2 are the corresponding variances.

To compare the average methylation levels of disease and normal tissues, we consider a generalized linear model,

$$logit[P(Y_i = 1)] = \alpha + \beta X_i,$$

in which α and β are regression coefficients. Under this model, a score statistic to test the difference of the average methylation levels of two groups is given by $U_1 = \sum_i (Y_i - \bar{Y}) X_i$. By treating X_i as the variable, the variance of the score statistics can be estimated by $\hat{\sigma}_{U_1}^2 = \sum_i (Y_i - \bar{Y})^2 \hat{\sigma}_X^2$, where $\hat{\sigma}_X^2$ is the estimated variance of methylation levels. As such, the score test can be formed by

$$T_1 = \frac{U_1^2}{\hat{\sigma}_{U_1}^2}.$$

This test is closely related to the commonly-used t-test as they both test the difference of means between two groups and has a centered χ_1^2 under the null hypothesis for a large sample size.

To test the difference in methylation variability between disease and normal tissues, we first define a variability score for each sample by $Z_i = (X_i - \bar{X})^2$, in which $\bar{X}$ is the sample mean of methylation levels. With the variability score, a similar logistic regression can be constructed with the variability score as the independent variable. Then, the score statistic is given by $U_2 = \sum_i (Y_i - \bar{Y}) Z_i$. It can be easily seen that this score statistic is proportional to the difference of estimated variances between disease and normal tissues, i.e. $U_2 \propto \hat{\sigma}_1^2 - \hat{\sigma}_0^2$. The variance of U_2 can be estimated by $\hat{\sigma}_{U2}^2 = \sum_i (Y_i - \bar{Y})^2 \hat{\sigma}_Z^2$, in which $\hat{\sigma}_Z^2$ is the estimated variance of the variability score. As such, the score test based on the variability score is

$$T_2 = \frac{U_2^2}{\hat{\sigma}_{U_2}^2}.$$

Similarly, T_2 also has χ_1^2 under the null hypothesis for a large sample size.

A joint test statistic for both mean and variability of methylation levels may be simply formed as $T_1 + T_2$ that has a χ_2^2 under the null hypothesis, or by Fisher's method for combing p-values when T_1 and T_2 are independent. However, T_1 and T_2 are generally not independent. To take into account the correlation between T_1 and T_2, it is necessary to estimate the covariance of U_1 and U_2. To do this, we denote the joint score statistic as $U_{joint} = (U_1, U_2)$ and its variance-covariance matrix can conveniently be estimated by $\hat{\Sigma}_{Ujoint}^2 = \sum_i (Y_i - \bar{Y})^2 \hat{\Sigma}_S^2$, in which $\hat{\Sigma}_S^2$ is the estimated variance-covariance matrix of X and Z. Then, the joint test is defined by

$$T_{joint} = U_{joint} \Sigma_{U_{joint}}^{-1} U_{joint}^T.$$

For a large sample size, T_{joint} has a centered χ_2^2 under the null hypothesis. When sample size is small, we could use a fast permutation procedure by randomly shuffling the order of the trait values of Y_is. Of note, the inverse of $\hat{\Sigma}_S^2$ does not require to be calculated at each replicate.

3. *Results*

3.1 Simulation study

We evaluated the performance of the proposed joint test through simulations. To evaluate the type I error rate, we first considered a case-control study with various sample sizes (n=20, 30, 50 and 100) for each group and sampled methylation values of cases and controls from various distributions (the standard normal, t distribution with 10 degrees of freedom or χ^2 with 2 degrees of freedom). For each scenario, we used 10,000 replicates to evaluate type I error rate. It is also of interest to examine the false positive rate at a more stringent threshold as a large number of loci are now routinely examined in methylation studies. We simulated 10 million replicates to evaluate type I error rate for a sample size of 100 cases and 100 controls. With this large number of simulations, we estimate the false positive rate with reasonable accuracy for a threshold of 10^{-5}. Finally, we examined the type I error rate of the proposed test after adjusting for batch effects. We assumed different proportions of cases and controls were assayed in two batches (30% in batch 1 for cases and 70% in batch 1 for controls), yielding difference methylation variability between cases and controls due to batch effects.

Table 1: The empirical type I error rate at the statistical significance level of 0.05

Distribution	Sample size	t-test	Levene	KS	T_{joint}	$T_{permutation}$
N(0,1)	20	0.050	0.040	0.037	0.039	0.053
	30	0.047	0.043	0.033	0.039	0.049
	50	0.050	0.042	0.037	0.045	0.050
	100	0.050	0.050	0.036	0.048	0.050
t_{10}	20	0.048	0.049	0.041	0.034	0.051
	30	0.050	0.043	0.037	0.036	0.050
	50	0.049	0.048	0.042	0.042	0.050
	100	0.052	0.047	0.037	0.046	0.051
χ_2^2	20	0.050	0.039	0.033	0.034	0.049
	30	0.051	0.044	0.034	0.038	0.050
	50	0.050	0.046	0.041	0.036	0.049
	100	0.054	0.040	0.048	0.043	0.049

We further compared power of the proposed joint test with the t-test, Levene's test and Kolmogorov-Smirnov test (KS) at the statistical significance level of 0.05. First, we simulated the methylation values of controls from a standard normal distribution, and cases from a normal distribution with various means and standard deviations (sd). The sample size was set at 100 for each group. Second, we simulated situations when different levels of heterogeneity exist in cancer tissues by sampling cases from a mixed normal distribution,

$$\pi_0 N(0,1) + (1-\pi_0)N(\mu^2, \sigma^2).$$

In this simulation, we set π_0 at 0.5 and varied μs and σ^2s to simulate different changes in the mean and variances between cancer and normal tissues. The sample size was set at 200 for each group. For each scenario we used 1,000 replicates to evaluate power. P-values of the proposed method were assessed using both the asymptotic distribution and the empirical null distribution obtained by the permutation procedure. The number of permutation was set at 1,000.

Table 1 shows the empirical type I error rate at the statistical significance of 0.05 for the proposed joint test (T_{joint}), the joint test based on permutation ($T_{permuation}$), the Levene's test Kolmogorov-Smirnov test (KS), and the t-test. We can see all tests maintained a good control of type I error rate under simulated scenarios. However, T_{joint} tended to be slightly conservative when sample size is small (n<50) and the distribution is highly skewed (χ_2^2 distribution). For a more stringent threshold of 10^{-5}, we found a similar pattern of the type I error rate for T_{joint}, which tended to be slightly conservative with the type I error rate at around 0.4×10^{-5}.

Table 2 shows the type I error rate of different tests when there is a difference in methylation variability between cases and controls due to batch effects. We can see the proposed test maintained a good control of type I error rate by incorporating a batch variable for adjustment for batch effects, while the Levene's test tend to have an inflated type I error rate.

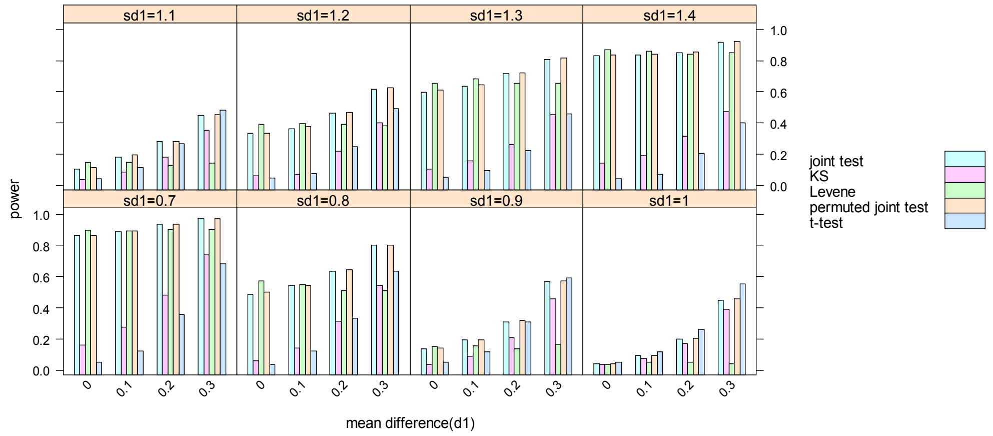
Table 2: The empirical type I error rate at the statistical significance level of 0.05 in the presence of heterogeneity in variability between cases and controls due to batch effects (n=100)

SD1*	SD2	t-test	KS	Levene	T_{joint}	$T_{permutation}$
1	1.1	0.053	0.036	0.061	0.059	0.058
1	1.2	0.046	0.038	0.066	0.042	0.044
1	1.3	0.038	0.033	0.068	0.046	0.047
1	1.4	0.046	0.039	0.091	0.042	0.042
1	1.5	0.045	0.031	0.090	0.042	0.042

*SD1 and SD2 are standard deviations of methylation values in batch 1 and 2, respectively. 70% cases are assumed to be assayed in batch 1 and 30% controls are assayed in batch 2.

Fig 2: The empirical power of the proposed test and the two-sample t-test at significance level of 0.05 to detect methylation loci associated with disease. (a) Controls are simulated from a standard normal distribution and cases are simulated with varied means and standard deviations (sds). The x-axes indicate varied means of cases and different panels represent varied sds. The sample size is 100 for each group. (b) Controls are simulated from a standard normal distribution and cases are simulated from a mixture normal distribution, i.e. $0.5N_0(0,1)+0.5N_1(d,sd)$. The x-axes indicate the means of $N_1(d,sd)$ and panels represent sds of $N_1(d,sd)$. The sample size is 200 for each group.

(a)



(b)

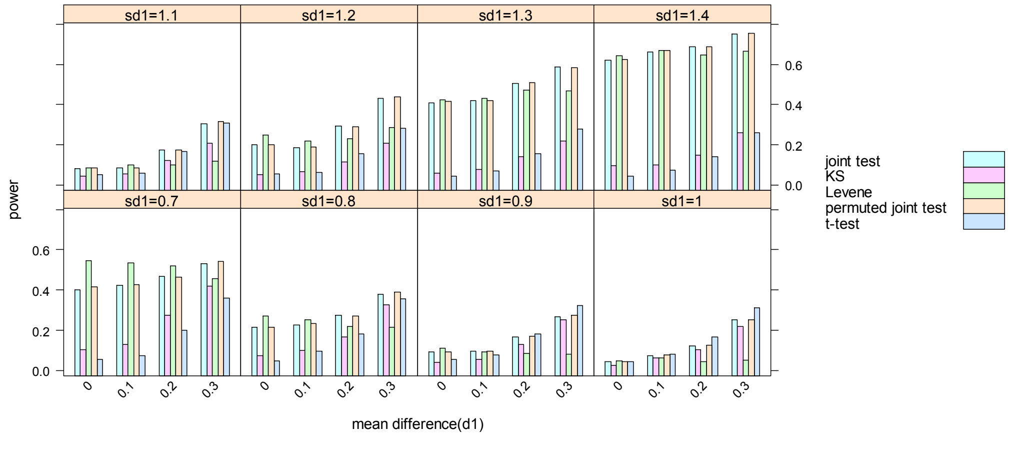


Figure 2 compares the empirical power of different tests at the significance level of 0.05 to detect methylation loci associated with disease under various situations. Based on our simulations, we have the following observations. First, T_{joint} was slightly less powerful than $T_{permutation}$. In situations when cases were sampled from an admixture distribution, the gain of power for $T_{permutation}$ appeared more obvious, which might reflect the conservative nature of the asymptotic test when the normal assumption does not hold. Second, the proposed tests were substantially more powerful than the t-test in the presence of heterogeneity of methylation variability between cases and controls. Third, T_{joint} and $T_{Permutation}$ were only slightly less powerful than the t-test when there was no heterogeneity of variability between cases and controls.

3.2 Application to an ovarian cancer study

To demonstrate the utility of the proposed test, we applied the proposed method to the data of United Kingdom Ovarian Cancer Population Study (UKOPS)[3]. This dataset is available at the NCBI Gene Expression Omnibus (http:///www.ncbi.nlm.nih.gov/geo) with accessing number GSE19711. The data includes 266 cases with 131 treatment and 135 post-treatment patients, and 274 age-matched healthy controls. To avoid the heterogeneity between age groups, we chose to analyze the 50-60 year group with 35 pretreatment patients and 82 controls. The data with 27,578 GpG loci were generated by Infinium assay with the HumanMethylation27 DNA Analysis beadchip. After background correction and normalization for the raw fluorescent intensities, a summarized value, i.e. β value, is calculated based on about 30 replicates in the same array by max(M,0)/[max(M,0)+max(U,0)+100], where M is the average signal from a methylated allele and U is from an unmethylated allele. Hence, the range of the β value is between 0 (unmethylated) and 1 (fully methylated). Because of the small sample size, we calculated both T_{joint} and $T_{permutation}$, and compared them to other tests. For computational reasons, the number of permutations for each locus was determined adaptively. Initially, 10^3 simulations were performed. If the resulting empirical p value was less than 0.01, 10^4 simulations were performed. If the p value from 10^4 simulations was less than 0.001, 10^5 simulations were performed.

Table 3: Number of loci with p-values smaller than the given cutoff from different tests

P-value	t-test	Levene	KS	T_{joint}	$T_{permutation}$	t-test and T_{joint} ($T_{permutation}$)
<0.01	750	157	1044	1047	1318	549(750)
<0.001	267	18	353	463	582	214(267)
<0.0001	62	4	85	169	250	51(62)

Fig 3: The correlation of –log10 p-values between the t-test, T_{joint} and $T_{permutation}$ for comparing pre-treatment cases and controls in the age group of 50-60 years. (a) the t-test and T_{joint} (b) the t-test and $T_{permutation}$ and (c) T_{joint} and $T_{permutation}$.

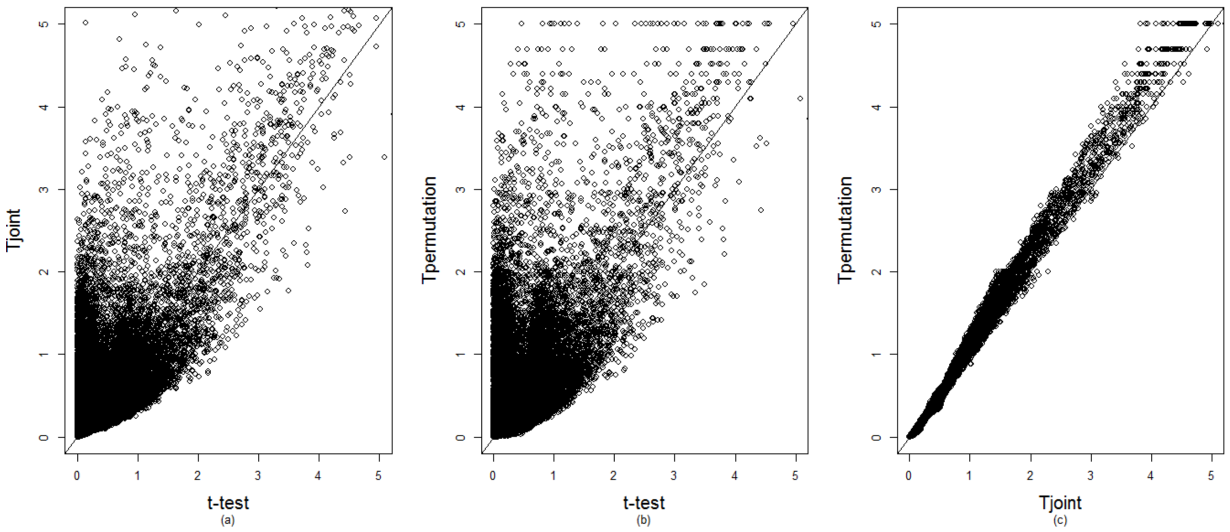


Table 3 shows the number of significant loci at different significance levels fordifferent tests. As expected, $T_{permutation}$ identified slightly more loci than T_{joint} because T_{joint} tends to be conservative for small sample sizes. However, T_{joint} and $T_{permutation}$ identified many more loci than the t-test. We further compared –log10 p-values of different methods for all loci (Figure 3). It can be seen that for many loci, T_{joint} and $T_{permutation}$ provided much lower p-values than the t-test, suggesting a large proportion of loci may have significant changes in the methylation variability between cases and controls. However, T_{joint} and $T_{permutation}$ had similar p-values, although $T_{permutation}$ tended to generate slightly smaller p-values. The analysis has also been performed on other age groups (60-70 years and >70 years) and yielded similar findings (data not shown).

4. Discussion

Although in recent cancer studies suggested the difference of methylation levels in both mean and variability was observed between cancer and normal tissue[5-7,12,13], so far most methods to identify differentially methylated loci examine the methylation mean and variability separately. To overcome this drawback, we propose a new statistical score test that achieves higher power than the t-test when there is heterogeneity in methylation variability between cases and controls. The traditional t-test gives less significant p-values in this case as it ignores information provided by the second moment of the methylation distribution. When there is no heterogeneity in methylation variability, the proposed method, although it is not optimal in terms of power, generally has robust power. Additionally, because the proposed test is very simple and hence can be calculated in a fast fashion, it is computationally feasible to be applied to very large methylation datasets, e.g. Illumina 450K. Our simulations and application to an ovarian cancer demonstrated the utility of our new method for discovering new methylation markers of complex diseases.

Essentially, the proposed method is an attempt to combine tests for mean and variance of methylation levels between two groups. With the normal assumption of methylation levels, one may perform the t-test for comparing means and F-test for comparing variances; and the joint test statistic can be obtained by Fisher's method for combining p-values.[14] However, the normal assumption is in general not true for methylation data. Moreover, a normal transformation is often not feasible for a large number of genome-wide methylation loci, since each can have a unique distribution. One of the consequences due to departures of normal distribution is that test statistics for the mean and variance are no longer independent, resulting in an inflated type I error rate when Fisher's method of combining p-values is used. To obtain valid p-values, computationally extensive sampling procedures, e.g. permutation, may be necessary. However, for highly significant p values, sampling is not a trivial task as such a procedure can be very inefficient. To address the issue of correlation, we propose a score test in which the correlation between test statistics for the mean and variance can be naturally adjusted. Another consequence of non-normality, in particular when the distribution is skewed, is that the t-test may lead to loss in power. The underlying assumption of our test statistic is that there is a linear relationship between independent variables and risk of disease. Because the linear relationship does not hold when the distribution is skewed, the power of our method may also be sensitive to skewness of the methylation distribution, although the validity of our method is quite robust. Of note, the permutation procedure itself would not improve power in this case. Further research is necessary to develop or identify statistical tests that can maintain good power when the distribution is highly skewed.

In the application to a real dataset from an ovarian cancer study, our method achieves higher statistical significance than the t-test at some loci. Indeed, a relatively large proportion of markers are only identified by the proposed test. The main reason for this might be that heterogeneity of methylation variability between cancer and normal tissue is a common phenomenon. In our study of both simulated and real datasets, the t-test performs better than our method when there is no difference in variance between cases and controls as an extra degree of freedom is used for testing variance in our method. However, Figure 2 (a) and (b) show that the relative power gain of the t-test is not very dramatic.

The proposed method can be generalized in different ways. In this paper we consider a case-control study. However, our score test is developed from a generalized linear regression model. As such, our method could be generalized for both continuous, e.g. age, and other categorical disease phenotypes. Another advantage of our method is that it can easily generalize to incorporate covariates. As such, our method can differentiate the true biological difference from the technical difference of variance between cases and controls, e.g. the batch effect, by incorporating an addition batch variable as a covariate. As shown in our simulation result, our method maintained a good control of the type I error rate after adjustment for batch effects. When there is no obvious variable, the technical difference can also be corrected by using a “genomic control”, in which the null distribution of the test statistic can be estimated from random methylation loci in the genome[15]. In addition, the application of our method can naturally extend beyond the analysis of a single methylation locus to the region-based (or gene-based) analysis under the framework of generalized linear regression. The advantage of the region-based analysis is it can make use of information of correlated loci in a spatial region. One challenge in applying the method for testing variances is the interpretation. Because the proposed test is an omnibus test that can simultaneously account for methylation mean and variability, it may be useful to further examine the independent effect of the change in methylation mean and variability when an association is identified. Various reasons could cause the change of methylation variability in disease tissues. One possibility is the heterogeneity of disease itself. However, it has also suggested that methylation variability may play an important biological role in the development of complex diseases[5]. Understanding the cause of heterogeneity of variance could have fundamental biological implications.

In summary, our results demonstrate that simultaneously testing differences in means and variances of methylation levels between cases and controls could identify disease related loci that are otherwise missed. Our method has the potential to be an efficient tool for screening potential methylation markers of diseases as our method does not require computationally intensive sampling to obtain valid p-values, and provides higher power than the t-test in the presence of differences in variability.

5. Acknowledgments

S.A is a high school student and worked as a summer intern in this project. T.W. is supported in part by the CTSA Grant UL1 RR025750 and KL2 RR025749 and TL1 RR025748 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH) and NIH roadmap for Medical Research, R21HG006150 from National Human Genome Research Institute (NHGRI).

References

1. Laird, P.W. & Jaenisch, R. DNA methylation and cancer. *Hum Mol Genet* **3 Spec No**, 1487-95 (1994).
2. Christensen, B.C. *et al.* Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet* **5**, e1000602 (2009).
3. Teschendorff, A.E. *et al.* Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res* **20**, 440-6 (2010).
4. Chen, Z., Liu, Q. & Nadarajah, S. A new statistical approach to detecting differentially methylated loci for case control Illumina array methylation data. *Bioinformatics* **28**, 1109-13 (2012).

5. Feinberg, A.P. & Irizarry, R.A. Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc Natl Acad Sci U S A* **107 Suppl 1**, 1757-64 (2010).
6. Hansen, K.D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**, 768-75 (2011).
7. Jaffe, A.E., Feinberg, A.P., Irizarry, R.A. & Leek, J.T. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics* **13**, 166-78 (2012).
8. Teschendorff, A.E. & Widschwendter, M. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics* **28**, 1487-94 (2012).
9. Snedecor, G.W.a.C., William G. *Statistical Methods*, (Iowa State University Press, 1989).
10. Levene, H. *In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, (Stanford University Press, 1960).
11. Chakravarti, L.a.R. Handbook of Methods of Applied Statistics. Vol. 1 392-394 (John Wiley and Sons, 1967).
12. Issa, J.P. Epigenetic variation and cellular Darwinism. *Nat Genet* **43**, 724-6 (2011).
13. Feinberg, A.P. *et al.* Personalized Epigenomic Signatures That Are Stable Over Time and Covary with Body Mass Index (vol 3, 65er1, 2011). *Sci Transl Med* **2**(2010).
14. Perng, S.K.a.L., R.C. A test of equality of two normal population means and variances. *Journal of the American Statistical Association* **71**, 968-970 (1976).
15. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).

USING DNASE DIGESTION DATA TO ACCURATELY IDENTIFY TRANSCRIPTION FACTOR BINDING SITES

KAIXUAN LUO[1] and ALEXANDER J. HARTEMINK[1,2]
[1] *Program in Computational Biology and Bioinformatics, and*
[2] *Department of Computer Science, Duke University, Durham, NC 27708, USA*
E-mail: kaixuan.luo@duke.edu, amink@cs.duke.edu

Identifying binding sites of transcription factors (TFs) is a key task in deciphering transcriptional regulation. ChIP-based methods are used to survey the genomic locations of a single TF in each experiment. But methods combining DNase digestion data with TF binding specificity information could potentially be used to survey the locations of many TFs in the same experiment, provided such methods permit reasonable levels of sensitivity and specificity. Here, we present a simple such method that outperforms a leading recent method, CENTIPEDE, marginally in human but dramatically in yeast (average auROC across 20 TFs increases from 74% to 94%). Our method is based on logistic regression and thus benefits from supervision, but we show that partially and completely unsupervised variants perform nearly as well. Because the number of parameters in our method is at least an order of magnitude smaller than CENTIPEDE, we dub it MILLIPEDE.

1. Motivation

Identifying binding sites of transcription factors (TFs) is a key task in deciphering transcriptional regulation. Methods based on chromatin immunoprecipitation (ChIP) permit the identification of TF binding sites at varying degrees of precision (ChIP-chip < ChIP-seq < ChIP-exo),[1] but they can only survey the genomic locations of a single TF per experiment.

To increase throughput, a complementary strategy based on the genomic digestion products of deoxyribonuclease I (DNase I, which we will simply call DNase) might be considered. DNase cleaves DNA in a manner that depends, *inter alia*, on the chromatin state of the genome, with nucleotides bound by proteins being cleaved less frequently than unbound nucleotides. Thus, the frequency with which a particular nucleotide is cleaved provides (noisy) information about the degree to which that nucleotide is bound by a protein. The primary motivation for using DNase digestion is that it applies non-specifically to all proteins binding the genome, regardless of their identity. This non-specific property is both a strength—in that it overcomes the one-TF-at-a-time limitation of ChIP—and a weakness, since simply knowing that a nucleotide is bound does not reveal the identity of the protein that binds it.

However, the binding specificities of many DNA-binding proteins are known (in this work, we assume specificities are modeled using a position weight matrix (PWM), but our method is general and can use any binding specificity model). This raises the prospect that a computational method combining DNase digestion data with prior knowledge of TF binding specificities might be able to identify binding sites in a TF-specific manner, at least for TFs with sufficiently distinct binding specificities. This prospect has spurred the development of a number of promising methods over the past few years. Though these methods all use DNase data in conjunction with binding specificity information, they adopt one of two distinct strategies:

(1) *TF-generic DNase signature.* Early methods started by scanning the mapped DNase data

for signatures of TF binding (roughly: the cleavage frequency is elevated, then drops for a short interval, and is then elevated again). Once sites with these signatures are detected, the TF(s) putatively bound to each site may be assigned by searching for matches to known PWMs. This strategy was first adopted by Hesselberth *et al.* in yeast, where initial site-detection was performed using a greedy approach.[2] As a technical refinement applied to the same data, Chen *et al.* developed a dynamic Bayesian network (DBN) approach for initial site-detection.[3] Boyle *et al.* developed a hidden Markov model (HMM) approach for initial site-detection, and applied it to DNase data from human cells.[4] More recently, Cuellar-Partida *et al.* formulated an informative positional prior from the human DNase data, and then looked for strong posterior evidence of binding, using PWM matches for the likelihood;[5] this method is essentially DNase-weighted motif scanning.

(2) *TF-specific DNase signature.* One disadvantage of the previous strategy is that the DNase signature of TF binding is necessarily the same for all TFs. A more effective strategy might be to start by scanning the genome for sites that match TF binding specificities; these will be called ‘candidate binding sites’. The DNase data in the genomic region surrounding each candidate binding site can then be used (along with other relevant information, such as the strength of the PWM match) to estimate whether the TF is indeed bound there. The first (and to our knowledge only) such method, given the moniker CENTIPEDE, was developed by Pique-Regi *et al.* and tested exclusively on human DNase data.[6]

Figure 1 shows two examples—Reb1 in yeast and REST (also known as NRSF) in human—of DNase data surrounding candidate binding sites that arise from PWM scanning of the genome. The figure illustrates that methods capable of using TF-specific DNase signatures are more likely to be effective at identifying TF binding sites. As such, in what follows, we focus exclusively on this second strategy.

2. Background

CENTIPEDE learns TF-specific DNase signatures in an unsupervised manner, using an EM algorithm to optimize a Bayesian mixture model.[6] The model discriminates bound from unbound sites using DNase data, plus other prior information (e.g., strength of match to PWM, degree of conservation, proximity to TSS). The likelihood of the DNase data is modeled in terms of both the total number of DNase cleavage events (‘cuts’) in the region around the candidate binding site (using a negative binomial), and the specific ‘shape’ of the cuts as they are arranged in the region (using a multinomial). CENTIPEDE’s discrimination power is largely driven by the PWM component of the prior and the multinomial component of the likelihood. However, the use of the highly flexible multinomial means that the model has the potential to over-fit irrelevant details in the shape of the DNase data—both in and around candidate binding sites—including random noise, systematic bias in DNase digestion, or artifacts arising from EM becoming trapped in a local mode. Indeed, the authors attempt to address the likely over-fitting by employing shrinkage estimators for their multinomial parameters, which improves certain evaluation metrics like area under the ROC curve (auROC), but at the expense of others, such as sensitivity at 1% false positive rate (FPR).

The authors of CENTIPEDE also explored the use of activating and repressing histone

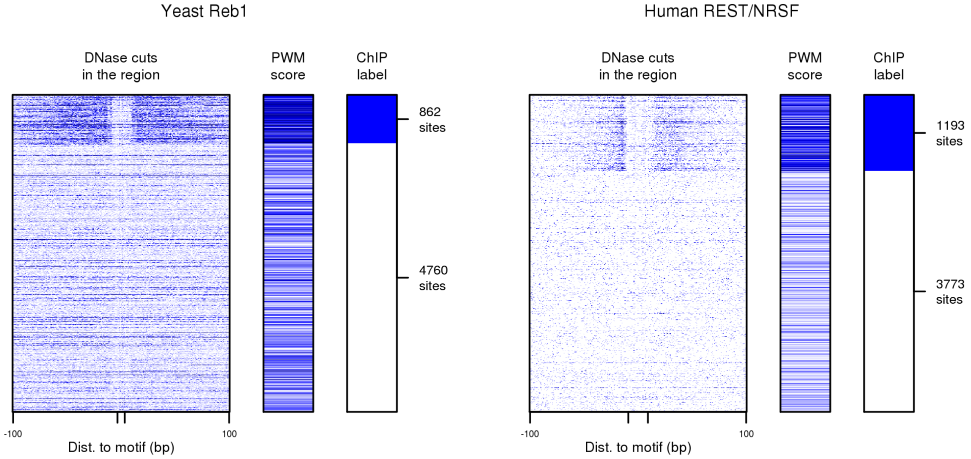


Fig. 1. DNase digestion data used in conjunction with TF binding specificity information can be used to identify TF binding sites. Left panel shows data for Reb1 candidate binding sites in yeast, and right panel shows data for REST (also known as NRSF) candidate binding sites in human. In each case, rows represent candidate sites based on PWM matches; rows are grouped by ChIP labels into positive and negative sets and then randomly ordered within each set. For each candidate binding site, columns depict DNase cuts in the region 100bp up- and downstream, PWM score, and ChIP label. Darker blue in data columns indicates higher number of DNase cuts at each position or higher PWM score. The figure makes clear that (a) both DNase data and TF binding specificity information provide noisy evidence of TF binding, and (b) since DNase cuts near Reb1 binding sites have a distinct pattern from DNase cuts near REST binding sites, methods using TF-specific DNase signatures are more likely to be effective at identifying TF binding sites.

marks in their likelihood, but reported limited benefit. Cuellar-Partida *et al.* later explored individual histone modifications, and proposed a TF-generic DNase signature method called H-p, intended to make better use of histone modification data (alongside DNase data and PWM scores).[5] With the same six human TFs of Pique-Regi *et al.*, they evaluated their proposed H-p method in comparison with (a) D-p, their method using DNase data and PWM scores, but omitting histone modifications, (b) D-s, a straw-man method using only the total number of nearby DNase cuts, but omitting both histone modifications and PWM scores, and (c) CENTIPEDE. Surprisingly, in terms of auROC, the proposed H-p model was the worst performer across the board for all six human TFs, suggesting that histone modifications are not likely to be helpful for this task. Even more interestingly, the straw-man method D-s outperformed H-p and D-p, and was competitive with CENTIPEDE, though it was not very sensitive at 1% FPR. This surprising set of results motivated us to ask the following questions:

(1) Based on the observation that D-s was performing competitively even though it lacked a TF-specific DNase signature and ignored the strength of the PWM match, could we develop an effective model that would address these two shortcomings of D-s, yet be much simpler than CENTIPEDE, to minimize the possibility of over-fitting?
(2) Would such a model perform well across organisms, not only in the human data of Boyle *et al.*[4] but also in the yeast data of Hesselberth *et al.*?[2] This is important for three reasons: (a) it would ensure that our conclusions about the relative merits of various approaches

are not specific to human, (b) it would allow us to evaluate using a larger set of TFs because many TFs have been profiled by ChIP in yeast, and (c) we had observed that CENTIPEDE performed quite poorly when applied to DNase data from yeast. In summary, could we develop a model that worked at least as well as CENTIPEDE in human, but improved dramatically upon CENTIPEDE in yeast?

In this paper, we describe a conceptually simple method for combining DNase data with information on TF binding specificity to identify TF binding sites. We show that our simple method outperforms CENTIPEDE marginally in human and dramatically in yeast. Its superiority is robust to the choice of evaluation metric, as well as the definition of positive and negative binding sites. Our method is based on logistic regression and thus benefits from supervision, but we show that partially and completely unsupervised variants perform nearly as well. Because the number of parameters in our method is at least an order of magnitude smaller than CENTIPEDE, we dub it MILLIPEDE.

3. The MILLIPEDE framework

MILLIPEDE improves upon CENTIPEDE in two primary ways. First, it reduces the number of parameters to be estimated by 1–2 orders of magnitude. This reduces the potential to over-fit irrelevant details in the DNase data, speeds up computation, and simplifies interpretation. The reduced number of parameters is a consequence of aggregating the DNase cut data within bins. One specific motivation for—and benefit of—such a strategy is that it reduces the prospect of fitting structure in the DNase signal that arises from digestion bias, which we show later may be important to address (though a more sophisticated approach would be to model the bias explicitly). Second, it is supervised, allowing the model to be trained to discriminate between bound and unbound sites rather than simply having to guess which sites are bound and unbound, as CENTIPEDE does. When labeled data are not available (for instance, if one is interested in identifying binding sites for a TF that has no ChIP data), we show later that partially and completely unsupervised variants of MILLIPEDE still perform admirably.

3.1. *Bins for aggregating DNase cuts*

Consider a candidate TF binding site, always oriented with respect to the strand on which the PWM is matched. As illustrated in Figure 2, within the binding site we construct two bins representing the left and right half of the site. Then, within the 100bp regions flanking the binding site both up- and downstream, we construct five equal-size bins (the choice of five is arbitrary: it allows the 100bp flanking regions to have some substructure, but not an excessive amount; we discuss this choice later). The result is 12 total bins across a genomic region of size $200 + w$, where w is the width of the TF PWM.

If we use all 12 bins in MILLIPEDE, we call the model M12. However, not all of these bins may be important, so we can construct various model simplifications by merging or dropping bins. For example, we can merge the two bins of the left and right halves of the binding site into a single bin, resulting in 11 total bins, so we call this model M11. Next, we can merge bins in the up- and downstream flanking regions: by merging the more proximal three bins

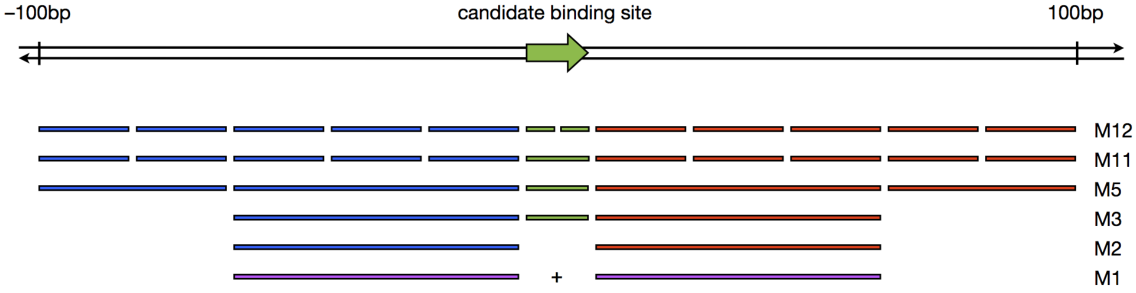


Fig. 2. Understanding the relationships between bins in various MILLIPEDE models. All bins are defined relative to the orientation of the candidate binding site: green bins are within the binding site, blue bins are upstream, and red bins are downstream. Models are arranged from most to least complex, so that each simpler model is derived from the one above it by merging or dropping bins. The simplest model M1 arises when the upstream and downstream bins of M2 are merged (thus shown in purple).

into a single bin, and merging the more distal two bins into a single bin, we are left with two bins upstream, two bins downstream, and one bin for the binding site, so we call this model M5. We can then drop the up- and downstream distal bins altogether, resulting in model M3. We can further drop the binding site bin, resulting in model M2. Finally, we can merge the two bins of M2, resulting in a model that has only one bin: M1. Specifically, M1 is a model that aggregates DNase cuts in the union of the two 60bp windows upstream and downstream of the binding site.

It is also possible to make the model more complex, for example by distinguishing between the forward and reverse strands when strand-specific DNase cleavage data is available. Strand-specific information was not available in the yeast DNase data from Hesselberth *et al.*,[2] but was available in the human DNase data from Boyle *et al.*[4] For example, if we start with model M12 but elect to distinguish between forward- and reverse-strand cuts, we have a model with 24 bins, which we call M24.

3.2. *Logistic regression*

The MILLIPEDE framework is based on standard logistic regression. Natural extensions with regularization (shrinkage or selection) are easily applied (though we do not explore them in this paper). Any relevant variables can be included, which makes the framework flexible and extensible. In what follows, the logistic regression covariates at each candidate binding site are simply (a) $\log_2$-transformed counts of aggregate DNase cuts within each bin, (b) the PWM score, and (c) optionally, a score measuring the degree of conservation. Formally, the full MILLIPEDE model for estimating the probability p_i that candidate binding site i is bound is:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \sum_{b=1}^{B} \beta_b \times \mathrm{D}_{b,i} + \beta_{\text{PWM}} \times \text{PWM}_i + \beta_{\text{CONS}} \times \text{CONS}_i$$

where $\mathrm{D}_{b,i}$ is the $\log_2$-transformed count of aggregate DNase cuts in bin b relative to site i, B is the total number of bins being considered in the model, and PWM_i and CONS_i are the PWM and conservation scores of site i, respectively. Note that we could also choose to include other variables if they were deemed relevant. Specifically, we could add a term $\beta_{\text{TSS}} \times \text{TSS}_i$ to include a score measuring proximity to the TSS for any TFs where that might be informative. We tested this but observed that TSS proximity scores were of negligible benefit; in what follows, we therefore omit them for simplicity.

When MILLIPEDE is run in a supervised mode, we learn its various coefficients from training data (and we can interpret the resulting model by examining the learned coefficients). We describe later how MILLIPEDE can also be run in completely or partially unsupervised modes.

4. Experimental methods

4.1. *Human data*

To facilitate comparison with the work of Pique-Regi *et al.*[6] and Cuellar-Partida *et al.*[5] in human, we used the exact same data wherever possible, kindly shared with us by Roger Pique-Regi. We used the same DNase digestion data in GM12878 cells, originally collected in the lab of Greg Crawford and reported in Boyle *et al.*[4] We used the same candidate binding sites as reported by Pique-Regi *et al.*; to avoid mappability bias, Pique-Regi *et al.* filtered out candidate sites whose surrounding region contained more than 20% unmappable nucleotides. We used the same PWM, conservation, and TSS scores as reported by Pique-Regi *et al.* (eventually deciding to omit the TSS scores, as discussed above). For training and evaluation, we studied the same six TFs, constructing positive and negative sets using the same ENCODE ChIP-seq data in GM12878 cells, as processed by Pique-Regi *et al.*

4.2. *Yeast data*

We used DNase digestion data in α-factor arrested yeast cells, collected in the lab of John Stamatoyannopoulos and reported in Hesselberth *et al.*[2] When scanning for candidate binding sites, we used PWM models of TF binding specificities from MacIsaac *et al.*,[7] and the sacCer2 (June 2008) version of the yeast genome. Following Pique-Regi *et al.*, to avoid mappability bias, we filtered out candidate sites whose surrounding region contained more than 20% unmappable nucleotides. For training and evaluation, we studied 20 TFs, constructing positive and negative sets using ChIP-exo data from Rhee *et al.*,[1] where available (Reb1, Rap1, and Phd1), as well as the ChIP-chip data of Harbison *et al.*,[8] as processed by MacIsaac *et al.*[7] MacIsaac *et al.* used conservation information to define positive binding sites, so to avoid potential bias, we omit all conservation data when evaluating performance in yeast. In practice, the usefulness of conservation information when applying MILLIPEDE in human suggests that it would likely also be useful in yeast.

4.3. *Gold standard evaluation sets regarding TF binding*

Cuellar-Partida *et al.*[5] describe a ‘peak-centric’ approach for constructing gold standard evaluation sets, in contrast to what they term the ‘site-centric’ approach of Pique-Regi *et al.* As it happens, the two approaches construct positive sets quite similarly—requiring positive TF binding sites to have both sufficiently strong ChIP signal and sufficiently strong PWM score—but construct negative sets quite differently (more on this below).

4.3.1. *Positive TF binding sites*

Since site-centric and peak-centric approaches construct positive sets in roughly the same fashion, we defined our positive TF binding sites in a manner analogous to Pique-Regi *et al.*: among all candidate binding sites determined by PWM scanning along the genome, positives are those that fall within a ChIP peak. In human, we used the exact same positive set as Pique-Regi *et al.*, while in yeast, we constructed our own positive set using ChIP-exo peaks (where available) and the TF binding sites of MacIsaac *et al.*, derived from ChIP-chip signals (requiring a ChIP-chip p-value < 0.005, and a 'moderate' level of conservation). One small caveat is that although the peaks from ChIP-exo (for Reb1, Rap1, and Phd1 in yeast) and ChIP-seq (in human) are likely of high enough quality to serve as a fairly accurate gold standard, peaks from ChIP-chip in yeast are perhaps better described as a bronze standard.

4.3.2. *Negative TF binding sites*

Since we are trying to predict whether or not candidate binding sites are bound, a natural choice for a negative set would be all candidate binding sites that are not in the positive set (do not fall within a ChIP peak); these are the negative sets we use in this paper, and we refer to these as 'MILLIPEDE gold standards'. Under such a construction, every candidate binding site is either positive or negative. The two previous approaches for constructing negative sets are notably different. The negative sets of Pique-Regi *et al.* are roughly subsets of ours because they require both of the following: (a) the candidate binding site does not fall within a ChIP peak, and (b) the ChIP treatment signal is less than the ChIP control signal at the site. This reduces the size of the negative set by including only those sites with the strongest negative signal, which makes the discrimination task easier and may thus over-estimate performance. In contrast, the negative sets of Cuellar-Partida *et al.* are roughly supersets of ours because negatives are defined as all genomic sites that do not fall within a ChIP peak (whether they are candidate binding sites or not). However, since we are only making predictions on candidate binding sites, our definition of negatives is equivalent to that of Cuellar-Partida *et al.* for this task. To ensure our results do not depend importantly on our choice of negative sets, we also evaluate each method's performance in human using the same negative sets that Pique-Regi *et al.* considered, referring to these as 'CENTIPEDE gold standards'.

4.4. *Evaluation metrics*

We evaluate the predictions of each model using four different metrics. To facilitate comparison with previous work, we report both area under the ROC curve (auROC) and sensitivity at 1% FPR. However, we also report area under the precision-recall curve (auPR) and precision at 1% FPR, which may be more realistic metrics of performance with imbalanced evaluation sets. When MILLIPEDE is supervised, reported results are averages based on 5-fold cross-validation.

4.5. *Availability*

Software, data, complete numerical results, and other Supplemental Material are all available from `http://www.cs.duke.edu/~amink`.

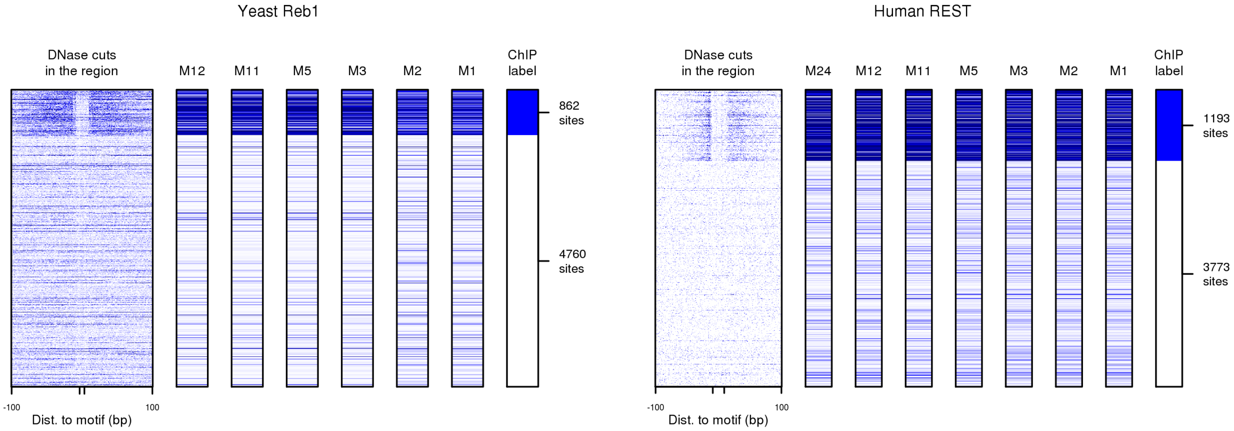


Fig. 3. MILLIPEDE models with various numbers of bins perform similarly. As in Figure 1, rows represent candidate binding sites based on PWM matches to Reb1 in yeast (left) or REST in human (right). For each candidate binding site, columns depict DNase cuts in the region 100bp up- and downstream, probability of being bound under various MILLIPEDE models, and ChIP label. Darker blue in data columns indicates higher number of DNase cuts or higher probability of being bound under the respective MILLIPEDE model.

5. Results

We compared the performance of our MILLIPEDE model with CENTIPEDE for 20 yeast TFs in G1-arrested cells and six human TFs in GM12878 cells. We used the MILLIPEDE gold standard for both yeast and human TFs (in Supplementary Material, we also show results using the CENTIPEDE gold standard for human TFs).

Since MILLIPEDE can use various numbers of bins as covariates in its logistic regression model, we first explored the effect of merging and dropping bins to produce simplifications of the full MILLIPEDE model. As illustrated in Figure 3, different simplifications of MILLIPEDE have surprisingly similar performance; although we use yeast Reb1 and human REST as running examples in the manuscript, this is true across all TFs and all four of our evaluation metrics (full results in Supplementary Material). Looking more closely, we observe that model M5 generally shows the best performance with DNase data for both yeast and human TFs, with a mean auROC using the MILLIPEDE gold standard of 94.2% across 20 yeast TFs, and 97.6% across six human TFs (the latter number becomes 98.6% when using the CENTIPEDE gold standard). As an aside, we note that M12 usually outperforms M24 in human, suggesting the strand-specific information may not be too informative, at least for these six TFs.

As demonstrated in the various panels of Figure 4 and the bar chart in Figure 5, our MILLIPEDE M5 model achieves significantly better ROC performance for yeast TFs compared to CENTIPEDE, and slight improvement for human TFs. In addition, M5 largely outperforms CENTIPEDE (nearly 10% higher on average) when considering other metrics like auPR, and sensitivity or precision at 1% FPR, for both yeast and human TFs using the MILLIPEDE gold standard (Supplementary Material). Finally, in terms of sensitivity at 1% FPR for human TFs using the CENTIPEDE gold standard, MILLIPEDE improves noticeably on the D-s straw-man method of Cuellar-Partida *et al.*,[5] achieving 82.2% with M5 and 84.7% with M12, each at least

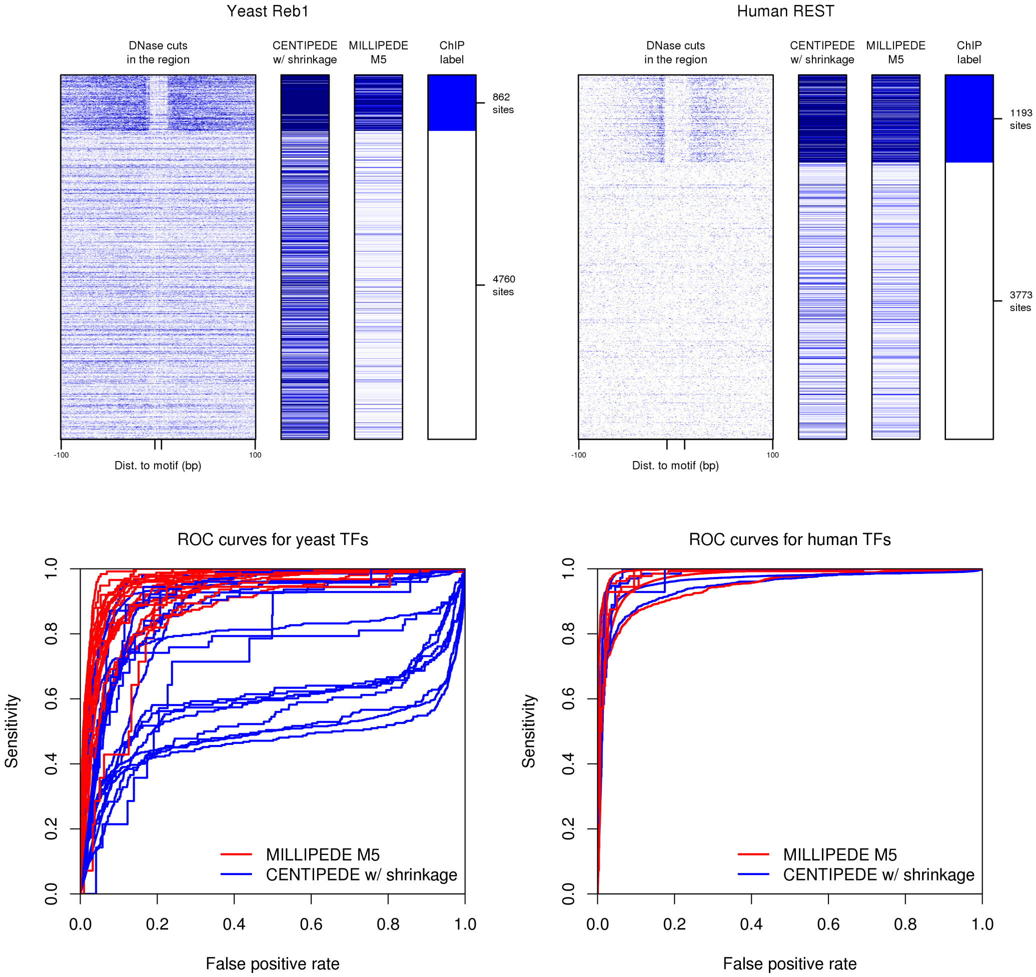


Fig. 4. Comparing MILLIPEDE model M5 to CENTIPEDE across all yeast and human TFs. Top panels are akin to those of Figure 3, but compare MILLIPEDE model M5 to CENTIPEDE with shrinkage estimates of multinomial parameters for yeast Reb1 and human REST. To reduce clutter, we only show CENTIPEDE results with shrinkage, since this performs noticeably better in an ROC setting than without (as shown in Figure 5). To confirm that results hold beyond the specific cases of Reb1 and REST, bottom panels show ROC curves for MILLIPEDE (red) and CENTIPEDE (blue) across all 20 yeast TFs (left) and all six human TFs (right). Two other yeast factors are shown later in Figure 6.

10% higher than D-s (Supplementary Material). Compared to D-s, MILLIPEDE's inclusion of PWM scores increases its ability to properly recognize the identity of the bound TF (versus other TFs that may be bound at those same candidate sites).

Normally, MILLIPEDE is run in a supervised mode to achieve high accuracy with the help of ChIP training data. However, when no ChIP data are available, we can run MILLIPEDE in a completely unsupervised mode: we choose a simple model and set the various coefficients to 1 (or -1 for bins that are either within a candidate binding site or distal). As shown in Figure 5, an unsupervised version of the MILLIPEDE M2 model still exhibits quite satisfactory auROC performance across both yeast and human TFs. As an intermediate scenario, if we

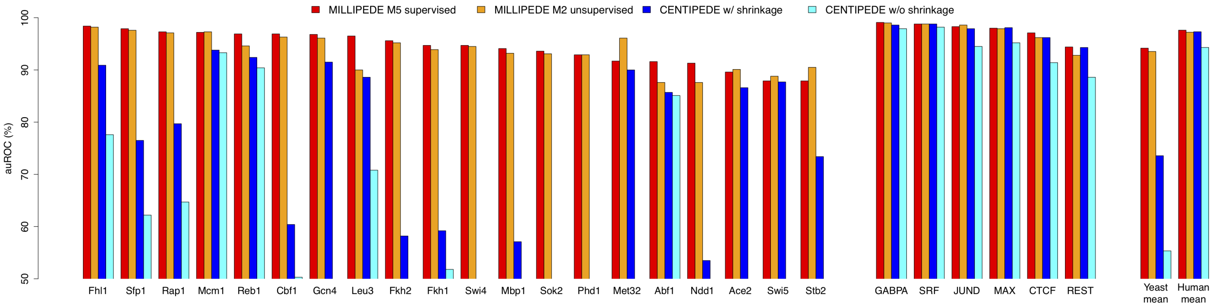


Fig. 5. Area under the ROC curve for 20 yeast and six human TFs. Red bars are MILLIPEDE model M5 run in a supervised mode, orange bars are MILLIPEDE model M2 run in a completely unsupervised mode, blue bars are CENTIPEDE with shrinkage, cyan bars are CENTIPEDE without shrinkage. Bars start at 50% since that represents random performance for an ROC curve; values below 50% are just not shown (e.g., for CENTIPEDE on Swi4, Sok2, and Phd1). The 20 yeast TFs are listed before the six human TFs; within each organism, the TFs are sorted such that the red bars decrease in height (the poor performance for Ace2 and Swi5 is perhaps unsurprising since the yeast DNase data are from cells arrested in G1). As summarized in the far right bars, mean performance is remarkably similar across the six human TFs, but MILLIPEDE improves dramatically upon CENTIPEDE in yeast, even when run completely unsupervised.

have some ChIP data available, but not for our TF of interest, we can run MILLIPEDE in a 'partially unsupervised' mode. To do so, we simply use coefficients trained on other TFs and apply those same coefficients to the new DNase data and PWM scores. This crude form of transfer learning results in very high prediction accuracy. Using the coefficients of MILLIPEDE M2 model trained on yeast Reb1 and applying it to the other 19 yeast TFs achieves a mean auROC of 93.9%, while using the coefficients of M2 trained on human REST and applying it to the other five human TFs leads to a mean auROC of 98.5% using the CENTIPEDE gold standard. These results suggest that even a single ChIP experiment can go a long way toward learning effective MILLIPEDE models.

While examining the DNase cleavage patterns for yeast TFs, we sometimes found strikingly similar DNase cleavage patterns for both bound and unbound sites, as with Abf1 and Mcm1, shown in Figure 6. Hesselberth *et al.* showed a significant match between Mcm1's DNase cleavage pattern within the binding site and the crystal structure of Mcm1-DNA contact,[2] but the similar patterns we see across all unbound sites (not just borderline cases) suggest the detailed cleavage patterns within the binding site are more likely a sequence-dependent artifact, perhaps arising from DNase digestion bias. To further test this claim, we also looked at the digestion patterns for Swi4, whose consensus binding sequence is CGCGAAA. Examining the more than 29,000 candidate binding sites that are unbound by Swi4, the number of cuts in the CG-rich left half of the candidate site is noticeably lower than the number in the AT-rich right half of the site (Supplementary Material).

Finally, we observed strong correspondence between DNase cleavage patterns in bound sites and the model coefficients learned in MILLIPEDE models. For most TFs, including our running examples of Reb1 and REST, we see positive coefficients for the bins proximal to the binding site, and negative coefficients for bins within the binding region or distal to it. These models therefore recapitulate the TF-generic DNase signatures of early papers[2–4] in this

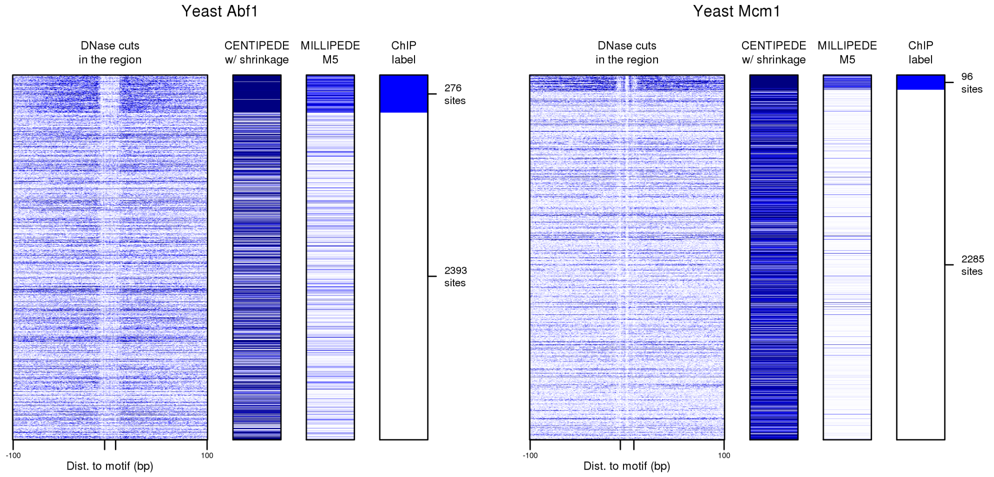


Fig. 6. DNase data can exhibit systematic artifacts such as sequence-dependent digestion bias. Left and right panels show yeast Abf1 and Mcm1 candidate binding sites, respectively. Notice that some fine details of the DNase cut data are preserved within and around candidate binding sites, whether the site is bound or unbound. CENTIPEDE is prone to over-fit these details both because of the large number of parameters in the multinomial component of its likelihood, and because it is unsupervised and uses EM to assign labels to candidate sites.

area: cleavage frequency rises to elevated levels near the binding site, then drops for a short interval, and is then elevated before gradually falling again. However, for individual factors, we saw subtly distinct patterns, lending credence to the importance of TF-specific DNase signatures. We even observed striking exceptions for a few TFs. For instance, for yeast Fkh1, MILLIPEDE models have significant positive coefficients for bins in the binding site and the bin immediately downstream, whereas for yeast Fkh2, MILLIPEDE models have significant positive coefficients for bins in the binding site and the bin immediately upstream. Correspondingly, we also see elevated DNase digestion in those regions without clear depletion in the binding site. As Fkh1 and Fkh2 are known to bind with other TFs like Mcm1 and Ndd1, this result may reflect consistent positioning of each TF relative to other co-factors along the genome.

6. Discussion

MILLIPEDE models achieve accurate and robust prediction performance under all four of our evaluation metrics across both yeast and human TFs. We have therefore demonstrated that a very simple model using only the most salient information from DNase data can perform as well as or better than more complex models like CENTIPEDE, with the further attendant advantages of fast computation, easy interpretation, and low potential of over-fitting.

Because our MILLIPEDE model is so simple, many variants can be imagined. For example, the number, widths, and locations of our bins have not been optimized in any way, though we briefly explored whether our results were sensitive to our admittedly arbitrary choices; we did not observe any notable change. Also, other covariates might be added to the model: MILLIPEDE's logistic regression framework permits great flexibility in including new covariates, should more information become available to further improve its performance. If the number of covariates becomes large, shrinkage or selection could be used to regularize the parameters.

Interestingly, we often observed detailed DNase cleavage patterns inside unbound candidate binding sites (especially in yeast), suggesting that some of the detail may be induced by sequence-dependent DNase digestion bias rather than actual protein-DNA protection at the single nucleotide level. This might also partially explain why CENTIPEDE does not work nearly as well for identifying TF binding sites in yeast. By declining to fit the detailed signal at every nucleotide, MILLIPEDE focuses its attention on the large-scale differences between bound and unbound sites, making it robust to biases that might arise at the single nucleotide level.

Since current technology for profiling TF occupancy requires a separate ChIP experiment for each TF being profiled, gaining a comprehensive understanding of the dynamic TF occupancy across the genome for all TFs across many tissues and conditions using only ChIP is utterly impractical. The prospect of using a complementary assay like DNase digestion has been tantalizing, but the sensitivity and specificity gap with ChIP has been too large to date. However, as more accurate methods like MILLIPEDE are developed to close the gap, efficient means for profiling TF occupancy across the genome for many TFs at once may become a reality. Intriguingly, since it can operate in a supervised mode, MILLIPEDE can leverage available ChIP data to train its models for identifying TF binding sites from DNase data alone.

7. Acknowledgments

We are grateful to Roger Pique-Regi for generous assistance in supplying code and data for his CENTIPEDE method, and to Jason Belsky and Matt Eaton for providing code for PWM scanning. We would also like to thank Jason Belsky, Gürkan Yardımcı, Greg Crawford, Dave MacAlpine, and Uwe Ohler for helpful discussions. This work was funded in part by grants from NIH (P50 GM081883-01) and DARPA (HR0011-09-1-0040) to A.J.H.

References

1. H. S. Rhee and B. F. Pugh, *Cell* **147**, 1408 (December 2011).
2. J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields and J. A. Stamatoyannopoulos, *Nat. Methods* **6**, 283 (April 2009).
3. X. Chen, M. M. Hoffman, J. A. Bilmes, J. R. Hesselberth and W. S. Noble, *Bioinformatics* **26**, i334 (June 2010).
4. A. P. Boyle, L. Song, B.-K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford and T. S. Furey, *Genome Res.* **21**, 456 (March 2011).
5. G. Cuellar-Partida, F. A. Buske, R. C. McLeay, T. Whitington, W. S. Noble and T. L. Bailey, *Bioinformatics* **28**, 56 (January 2012).
6. R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad and J. K. Pritchard, *Genome Res.* **21**, 447 (March 2011).
7. K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo and E. Fraenkel, *BMC Bioinformatics* **7**, p. 113 (2006).
8. C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. MacIsaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel and R. A. Young, *Nature* **431**, 99 (September 2004).

EPIGENOMIC MODEL OF CARDIAC ENHANCERS WITH APPLICATION TO GENOME WIDE ASSOCIATION STUDIES

Avinash Das Sahu[1,3], Radhouane Aniba[1,3], Yen-Pei Christy Chang[2] and Sridhar Hannenhalli[1,*]

[1] *Center for Bioinformatics and Computational Biology,*
University of Maryland, College park, MD 20742, USA
[2] *School of Medicine, University of Maryland, Baltimore, MD 20201, USA*
[3] *co-first authors*
[*] *E-mail: sridhar@umiacs.umd.edu*

Mammalian gene regulation is often mediated by distal enhancer elements, in particular, for tissue specific and developmental genes. Computational identification of enhancers is difficult because they do not exhibit clear location preference relative to their target gene and also because they lack clearly distinguishing genomic features. This represents a major challenge in deciphering transcriptional regulation. Recent ChIP-seq based genome-wide investigation of epigenomic modifications have revealed that enhancers are often enriched for certain epigenomic marks. Here we utilize the epigenomic data in human heart tissue along with validated human heart enhancers to develop a Support Vector Machine (SVM) model of cardiac enhancers. Cross-validation classification accuracy of our model was 84% and 92% on positive and negative sets respectively with ROC AUC = 0.92. More importantly, while P300 binding has been used as gold standard for enhancers, our model can distinguish P300-bound validated enhancers from other P300-bound regions that failed to exhibit enhancer activity in transgenic mouse. While GWAS studies reveal polymorphic regions associated with certain phenotypes, they do not immediately provide causality. Next, we hypothesized that genomic regions containing a GWAS SNP associated with a cardiac phenotype might contain another SNP in a cardiac enhancer, which presumably mediates the phenotype. Starting with a comprehensive set of SNPs associated with cardiac phenotypes in GWAS studies, we scored other SNPs in LD with the GWAS SNP according to its probability of being an enhancer and choose one with best score in the LD as enhancer. We found that our predicted enhancers are enriched for known cardiac transcriptional regulator motifs and are likely to regulate the nearby gene. Importantly, these tendencies are more favorable for the predicted enhancers compared with an approach that uses P300 binding as a marker of enhancer activity.

Keywords: Enhancer, Epigenomics, SVM, heart disease

1. Introduction

Eukaryotic transcription is intricately regulated at multiple levels including chromatin reorganization through epigenomic modifications and sequence specific binding of transcription factors (TF) to either proximal promoter or to distal enhancer/repressor regions of the gene.[1,2] Enhancers can regulate their target genes from long distances, up to a megabase away and are especially important in regulating developmental and tissue-specific genes.[3,4] Numerous genome wide association studies (GWAS) have revealed genomic loci associated with various human traits.[5] Going from association to causality is however a major challenge, because a vast majority of GWAS signals lie in non-coding regions, often far from any gene, and our understanding of functional consequences of non-coding mutations is incomplete. It is possible that many of these associations are mediated via regulatory regions.[6] By investigating putative polymorphic enhancers near GWAS signals, we might be able to identify the causal links between genetic variability and disease, at least in some cases. Thus, both for our fundamental understanding of transcriptional regulation as well as for interpretation of genotype-phenotype relationships, a comprehensive knowledge of context-specific enhancers is critical.

Large scale identification of enhancers is challenging because they do not have sufficiently discriminating sequence properties (except for their tendency to harbor homotypic binding motifs[7]) and their location is not restricted relative to the location of the target gene. Moreover, enhancers are often tissue and cell-type specific and are detectable only under the appropriate conditions. Recent revolution in sequencing technologies have triggered several large scale profiling of epigenomic marks and analysis of these marks have revealed strong associations between enhancers and specific epigenomic marks (either positive or negative[8–10]). Using genome-wide profiling of several epigenomic marks, Ernst et al. segmented the genome into 51 segment classes, where each segment class is defined by a specific combination of epigenomic marks.[8,11] They designated two of these segment classes as strong and weak enhancers. Apart from epigenomic marks, histone acetylase P300 is known to bind to tissue-specific enhancers, with high rate of experimental validation using mouse transgenic.[10,12] However, it is argued that while P300 may mark tissue-specific enhancers, those enhancers are not necessarily active in a specific context.[13] This assertion is consistent with less than perfect validation rate of P300 bound regions as enhancers. Despite this, previous approaches to predict enhancers have used P300 bound regions as the gold standard to assess the methods prediction accuracy.[14,15]

Here we report an SVM trained specifically on 83 validated cardiac enhancers using four epigenomic profiles marks (H3K4me1, H3K27me3, P300 and DNase hypersensitivity) in human heart tissue. Our model achieves a cross-validation classification accuracy of 84% and 92% on positive and negative sets respectively. It was encouraging that our model can distinguish validated enhancers from those that were bound by P300 but failed to exhibit enhancer activity in transgenic mouse. Next, starting with a comprehensive set of 229 SNPs associated with cardiac phenotypes in 36 GWAS studies, we identified putative enhancers harboring SNPs in linkage disequilibrium (LD) with the GWAS SNP. We found that our predicted enhancers are enriched for binding sites for all known core cardiac transcriptional regulators GATA, MEF2, STAT, NF-AT, Nkx, and FOX. Using a novel approach we show that the predicted enhancers are likely to regulate the nearby gene. Our predicted enhancers uniquely point to a few genes highly relevant to the heart disease. Moreover, these tendencies of having enriched cardiac transcriptional motifs and likelihood of regulating nearby genes are more favorable for the predicted enhancers compared with an approach that uses P300 binding as a marker of enhancer activity. Overall, we show that a SVM model trained exclusively on validated enhancers performs better than those that use P300 binding as gold standard and that GWAS studies can be better interpreted in light of predicted polymorphic enhancers.

2. Results

2.1. *SVM model for cardiac enhancers*

2.1.1. *Data*

Heart tissue was chosen for our analysis because of the availability of both relevant epigenetic data (H3K4me1, H3K27me3, P300 and DNase hypersensitivity) and validated human enhancers. We collected 83 experimentally heart enhancers validated in mouse transgenic from VISTA browse and split them into 1kb regions (step size 500 bps) to be used as positive training set. Negative set was constructed by mixing random samples of 1 Kb long regions from the genome and randomly selected promoters. H3K4me1, H3K4me3, H3K27me3, P300 and DNase-I epigenetic markers, which have previously been shown to be associated with tissue-specific enhancers, were collected for the heart tissue from the GEO database. For each epigenetic mark we calculated its average signal strength across every 1 Kb genomic region as feature vector of the region. In order to normalize the feature vectors of the positive and negative set to zero mean and unit variance, we randomly sampled 40,000 1 Kb regions across the genome to estimate mean and variance of feature vector.

2.1.2. *Training*

Epigenetic marks relevant to enhancers are relatively sparse in the genome. If the negative example in the training set only included random regions then SVM would choose subset of these inactive regions as its support vectors and would create a classifier hyperplane separating inactive regions from any epigenetically active region, resulting in high false positive rate. Therefore, in our negative set, in addition to random genomic regions, we added gene promoters as examples of epigenetically active non-enhancer regions. Figure 1 shows the effect of varying the proportion of promoters region in negative training set. In general, we found that a greater proportion of promoters in negative set improves positive set accuracy with relatively smaller decline in negative set accuracy, at least initially. This suggests that including a small fraction of promoters in the negative training set results in a better classification. Therefore, we constructed the negative training set by mixing 1000 random genomic regions and 250 randomly selected gene promoters.

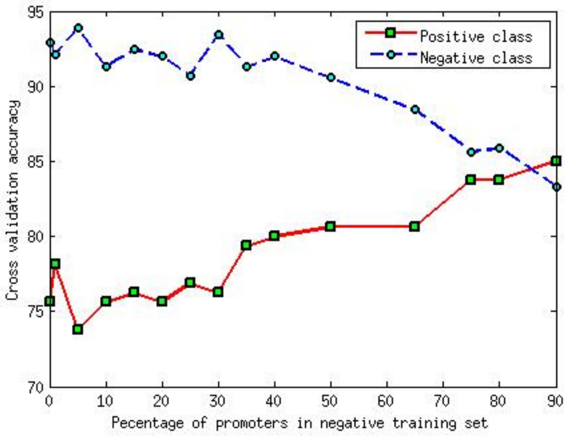


Fig. 1: Effect of variation of proportion of promoter region on accuracy of model. Two fold cross validation is used for positive set. Negative set accuracy is calculated by running the trained model on large number of random 1 kb genomic regions not including those used for training.

2.1.3. *Testing*

We used 5-fold cross validation for positive set accuracy estimate. For negative test set we randomly sampled 1000 1kb genomic regions. On performing grid search (see Methods) to train the SVM model the average testing classification accuracy on positive set was 84.1% and on negative set was 92%. The roc curve for the model prediction is shown in Figure 2. The AUC of the model was 0.9231.

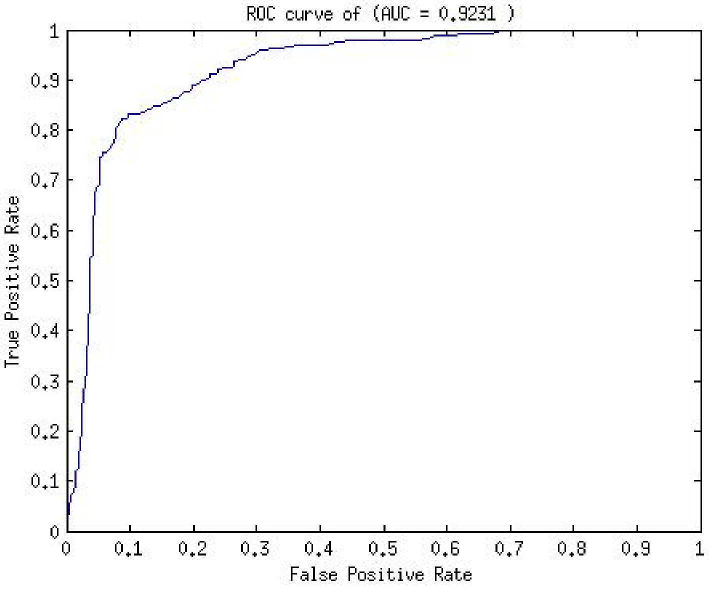


Fig. 2: ROC curve of SVM model

Despite some evidence to the contrary, a number of previous works have assumed P300-bound regions to be active enhancers and used them as gold standards to train and evaluate enhancer prediction tools. Next, we tested whether our model trained on validated enhancer and oblivious of P300 binding can nevertheless distinguish active and inactive P300-bound regions. We tested our model with 12 P300 peaks in human heart which were found not to have enhancer activity.[16] Interestingly, the model classified 10(83%) of these cases as non-enhancers. Although based on a small set of examples, this suggests that our model can distinguish inactive P300-bound regions from active enhancers.

Narlikar et al.[17] proposed a model based on specific motifs as features for cardiac enhancer identification. To compare performance of our model with their's, 83 validated enhancers were separated into 60 training and 23 testing instances. SVM was trained on the 60 instances. We extracted the 1Mb regions flanking each of the 23 test enhancers and predicted enhancer in those genomic regions using the trained SVM. We first checked how well P300 can retrieve the validated enhancers. We found that there are only 69 P300 peaks in adult human heart in the 23 genomic regions, out of which only one overlapped with a validated enhancer. In other words, P300 peaks are poor predictor of enhancer activity in this context.

Using our trained SVM model we scored each 1 Kb region in the test set. Cardiac enhancer predicted in Narlikar et al. [17] are typically much shorter. For fair comparison with Narlikar et al. [17] (1) we extended each of their enhancer to 1 Kb region flanking the reported location, and (2) used a threshold on the enhancer score such that the predictions made by our SVM and the Motif based model cover almost the same number of enhancers (same basepair coverage as well due to extension) in the genomic test set. Among the 8522 enhancer regions predicted by the SVM, 21 of the 23 validated enhancers were included, while among 8551 enhancer regions predicted by Narlikar et al.[17] only 13 were covered. we repeated the above comparison between our method, P300 peaks and Narlikar et. al. 10 times with different sets of 60 training and 23 testing instances out of total 83 enhancers. Figure 3 shows the number of enhancer predicted by each method across different iterations.

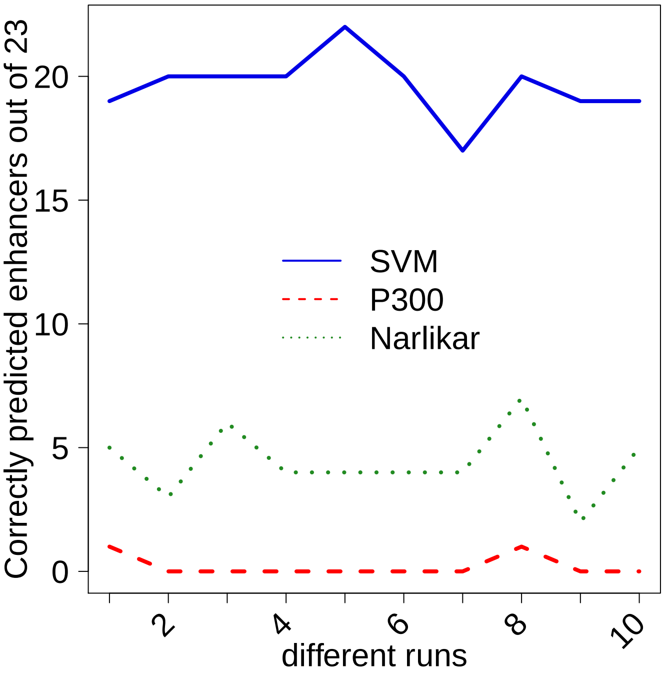


Fig. 3: Number of enhancers (out of 23) predicted by SVM, P300 peaks and Narlikar et. al.

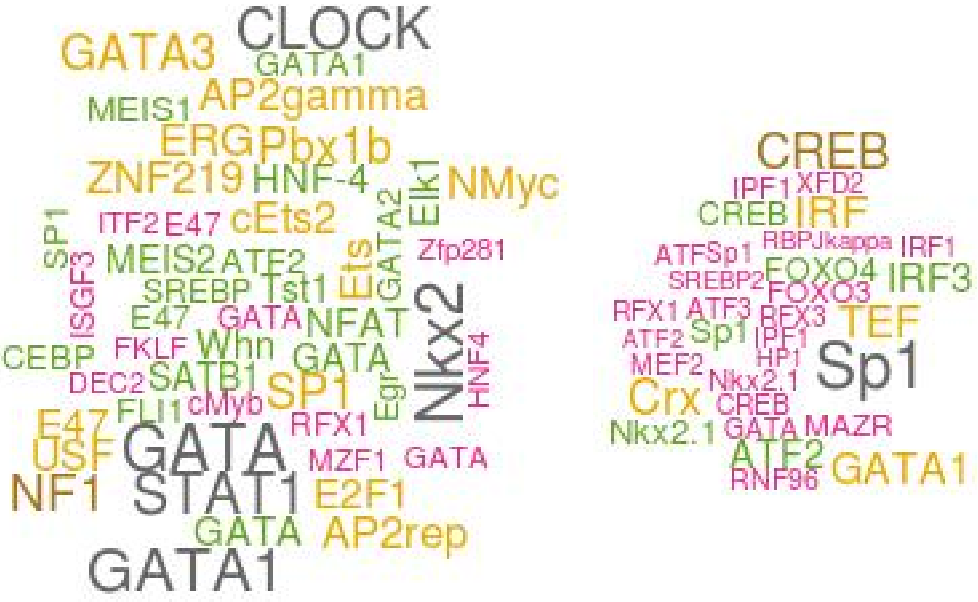
Taken together, these results suggest that the SVM model trained on epigenomic data is more suitable for identifying cardiac enhancers than are P300 binding or motif based models.

2.2. *Identification of cardiac enhancers near SNPs associated with cardiac phenotypes*

Next, we hypothesized that the causal variants underlying GWAS signals might lie within an enhancer element and affect gene regulation. We tested this hypothesis on SNPs associated with a variety of cardiomyopathies. Starting with NHGRI's GWAS catalog,[5] which includes 1332 studies revealing 6852 SNPs, we manually selected studies for cardiovascular disease traits. This yielded 229 SNPs from 36 studies. We then extended this seed SNPs set to include all other SNPs in Linkage Disequilibrium (LD) with a seed SNP using Broad Institutes SNAP server.[18] We included all SNPs within 500kb from a seed SNP with $r^2 \geq 0.3$. The extended SNP were merged from the 1000 Genome Project and multiple HapMap releases (Consortium 2003; Consortium 2010). For each of the resulting 14233 SNPs, we scored 1kb flanking region using our SVM model to prioritize them as potential cardiac enhancers. Of all SNPs, the SVM scored 1054 as having enhancer probability ≥ 0.8. We found that distance of these enhancers from the corresponding GWAS SNP was significantly shorter than expected (Wilcoxon p-value = 3.9E-05).

2.3. *Cardiac enhancers near cardiac GWAS SNPs are enriched for cardiac regulator motifs*

Cardiac transcription is primarily regulated by members of GATA, MEF2, STAT, NF-AT, Nkx, and FOX families of TFs.[19–22] Next, we tested whether predicted enhancers near GWAS SNPs are enriched for known cardiac TF binding motifs. We first constructed three SNP sets: (1) eSNPs: comprised of the top 500 SNPs in LD with a GWAS SNP ranked by the SVM score, (2) pSNPs: the top 500 SNPs in the LD with a GWAS SNP ranked by mean P300 tag density (using bigwig summary tool from UCSC) in human heart, (3) gSNPs: The GWAS SNPs themselves. For each SNP we extracted the 1kb genomic flanking region resulting in three sets of sequences. For each sequence we determined the binding sites corresponding to 981 vertebrate motifs in TRANSFAC[23] whose motif match score (using our own tool[24]) was in the top 95th percentile of scores achievable by that motif. We then determined the enriched motifs in one set of sequences relative to the other using Fisher Exact Test. Because enhancers have distinctive compositions which can bias motif enrichment, we normalized the two sequence sets for their GC composition via random sampling prior to motif enrichment analysis. When comparing SVM SNPs to the GWAS SNPs, 50 motifs were enriched with p-value ≤ 0.05, 11 of which corresponded to multiple representatives of GATA, STAT, NF-AT, Nkx families. When we compared the P300 SNPs with GWAS SNPs, among the 34 enriched motifs with GATA, Nkx and STAT families were represented by 4 motifs. Importantly, when we compare SVM SNPs directly to the P300 SNPs, GATA, FOX, MEF2 families of TF motifs were found to be enriched among the 32 enriched motifs. Figure 4 shows the top 50 motifs significantly enriched in SVM SNPs compared to GWAS SNPs or P300 SNPs. When we restrict the motif search to 20 bps flanking the SNP using same parameters, we still observe enrichment of NF-AT and STAT motifs in SVM SNPs relative to GWAS SNPs. However similar enrichment is also observed in P300 SNPs. It is possible that the SNP affect the formation of cis regulatory modules indirectly. Further investigation is required. In summary, all core cardiac TF families are enriched near eSNP loci, relative either to GWAS SNPs or to P300-bound regions. The overall conclusion was comparable when we used top 200 SVM scores and top 200 P300 score to be construct eSNP and pSNP sets. We note that because of small numbers, the p-values were modest and did not qualify a strict FDR threshold.



(a) SVM VS GWAS

(b) SVM VS P300

Fig. 4: Significantly enriched motifs in SVM SNPs. The size of each TF label is proportionsal to its significance. For instance, the p-value for GATA1 in (a) is 0.001 and in (b) is 0.004. The largest p-value is 0.05.

2.4. *Cardiac enhancers near cardiac GWAS SNPs are likely to regulate the nearby genes*

Next we tested whether the predicted enhancers are likely to regulate genes. While enhancers can in principle regulate non-neighboring genes, a majority of them do regulate nearby genes,[25] therefore, we focused only on the gene promoter closest to the SNP. For a SNP locus and a gene promoter, we estimated the likelihood of SNP locus to regulate the gene as the correlation between the DNase-I hypersensitivity (DHS) at the locus and the expression of the genes across 15 cell types in which DHS and RNA-seq was performed in parallel (see Methods); this approach to link a putative enhancer to a target genes is similar to Ref. 11. We constructed three comparison SNP sets. gSNP comprised of 229 GWAS SNPs. To construct eSNP set, we selected the SNP with highest SVM score in LD with each GWAS SNP as long as the SVM score was ≥ 0.8, resulting in 115 eSNP, all of which were intronic or intergenic. Similarly, to construct pSNP set, we selected the SNP with highest P300 mean tag density in LD with each GWAS SNP as long as the P300 tag density was ≥ 1, resulting in 58 pSNP. For each SNP we obtained the closest gene promoter. We then performed three pair-wise comparisons. For instance, when comparing eSNPs with gSNPs, we focused on genes that were closest to both an eSNP and a gSNP. Then we computed two DHS-expression correlations - between eSNP locus and the gene and between gSNP and the same gene. Given all such pairs of correlations we tested whether eSNP-gene correlation was greater than the gSNP gene correlation using paired one-side Wilcoxon test. We found that eSNP loci were more likely than gSNP loci to regulate the closest gene (based on 124 genes, p-value = 0.03), eSNP loci were more likely than pSNP loci to regulate the closest gene (based on 50 genes, p-value = 0.01), and pSNP loci were not more likely than eSNP loci to regulate the closest gene (based on 23 genes, p-value = 0.87). We also checked whether the distance of eSNPs from the closest gene promoter was shorter than that for gSNP or pSNP and we did not observe a statistical difference. The results suggest that SVM predicted enhancers are more likely to regulate the nearby genes relative to both the original GWAS SNPs and P300 predicted enhancers.

2.5. *Genes near cardiac enhancers are enriched for cardiac function*

Next we tested whether the genes uniquely closest to the eSNPs provide greater insight into the cardiovascular disease phenotype, relative to genes uniquely closest either to gSNPs or the pSNPs. We used the same criteria as above to obtain the closest gene lists, but unlike the expression analysis above we retained only the unique genes in each list. Unfortunately, the uniqueness requirement greatly reduced the number of genes with 94 for gSNP, 17 for eSNPs and only 2 for pSNPs. We then used ToppGene[26] to compare enrichment of disease categories in the three gene lists. ToppGene uses three sources for disease ontology terms - GWAS, Comparative Toxicogenomics Database, and OMIM. We excluded GWAS to avoid circularity. As expected, the pSNP gene list did not show any enrichment. At FDR ≤ 0.05 the genes near gSNP also did not show enrichment for any disease term. The 17 genes in the eSNP list include NOS3 and MYH7. NOS3 alone showed enrichment for 2 terms - “Hypertension, Pregnancy-Induced” and “Coronary Vasospasm”. MYH7 alone was enriched for 5 distinct terms from OMIM database, all immediately related to myopathy or cardiomayopathy. The results are based on very limited dataset and one cannot draw general conclusion but they suggest that SVM can uniquely lead to genes directly relevant to the phenotype.

3. Conclusion

Here we present a SVM model for human cardiac enhancers based on four epigenomic marks H3K4me1, H3K27me3, DHS and P300, each of which have previously shown to be associated with enhancers in various cell types. While P300 is known to bind to tissue specific enhancers,[12] and have been used as the gold standard for estimating accuracy of previous enhancer prediction approaches,[14,15,17] many P300 bound regions fail to exhibit enhancer activity.[12,13] Our SVM trained specifically on experimentally human cardiac enhancers validated in trangenic mouse, can not only predict other validated enhancers with high accuracy, it can also distinguish validated enhancers from the regions that were bound by P300 but failed to exhibit enhancer activity in transgenic mouse.

There are three prior approaches to predict enhancers. Narlikar et al. use clusters of known cardiac TF motifs as predictor of cardiac enhancers.[17] Lee at al. train a SVM model based on genomic features based on cardiac P300 bound regions.[14] Another SVM model for CD4+ T-cell enhancers based on epigenomic features, again, using P300-bound regions as the gold standard was proposed in.[15] We have demonstrated the ability of our SVM model to distinguish between active and inactive P300 bound sites. Additionally, direct comparison of prediction accuracy on novel validated cardiac enhancers of our SVM model with that of P300[14] and Narlikar et al.,[17] explicitly shows that active enhancers have specific epigenomic properties not captured just by P300 binding or by clusters of putative binding sites. Genomic regions bound by P300 may not be active. Therefore, use of additional features add the tissue specific context to the model. Furthermore, kernel transformation of feature space used by SVM builds a non-linear classifiers. Thus it captures a greater variety of enhancers by recognizing a wider combination of epigenetic factors.

It has been previously suggested that a better knowledge of context-specific enhancers can help interpret GWAS signals.[8] However, this reasonable assertion has not been tested explicitly on a specific disease area. Here we use our enhancer prediction tool to interpret GWAS studies related to cardiovascular phenotypes. We found an enrichment of high scoring cardiac enhancers near cardiac GWAS SNPs. Analysis of these putative enhancers suggest that (1) they are enriched for known core cardiac transcription factor binding sites, (2) they are likely to regulate nearby genes, and (3) they can uniquely point to certain genes involved with cardiac function and heart disease.

4. Methods

4.1. *Correlating DNase Hypersensitivity and Gene Expression*

To assess correlation of chromatin accessibility at a putative enhancer to expression level of a putative target gene, we extracted genome wide DHS as well as RNA-seq data from 15 cell types from a single study (GSE29692, GSE23316) representing a breadth of cell types HepG2, GM12878, A549, HeLa-S3, AG04450, BJ, NHLF, NHEK, HUVEC, h1-Hesc, HMEC, HSMM, K562, MCF-7, SK-N-SH_RA. For the enhancer region we extracted the DHS tag density in each of the 15 cell types using bigWigSummary tool. Correspondingly, for the putative target genes we obtained the gene expression (RPKM) in the same set of cell types. We then estimated the pearson correlation between DHS and gene expression as an indicator of interaction between the enhancer and the gene.

4.2. *More on SVM and grid search criteria*

There are several references available for SVM.[27–30] Here we give a brief review to appreciate our criteria for cross validation and grid search on the parameter space. In SVM, vector in original feature space is projected onto a higher dimensional feature space using kernel function (usually non-linear). Because of this the data which in original space is not linearly separable, becomes separable in transformed space, where the SVM tries to find a maximum margin hyperplane that separates the positive and negative set in the kernel space. SVM, employs a structural risk minimization (SRM) method[31,32] to obtain the hyperplane, which tries to balance complexity of the model while minimizing the empirical risk. Therefore, relative to traditional methods based on empirical risk minimization, SVM is better suited to handle the problem of overfitting. SVM chooses a maximum margin hyperplane by identifying subset of training data (called support vectors), which would be closer to the optimal separating plane. Support vectors are cases which are most difficult to classify as positive or negative. Therefore to ensure good performance of SVM classifier, it is necessary to have a set of extreme examples (in both positive and negative example in the training set) that would qualify as support vectors.

Our positive training set included 330 (80% of 415) regions while the negative training set included 1000 regions. We weighted the positive and negative examples to accommodate for the difference in sizes. An exhaustive search over the weight space was conducted to obtain best possible cross-validation result. The weight used for negative and positive set respectively was 1 and 1.2. Furthermore, we defined our criteria for grid search based on the observation that randomly sampled negative set may contain enhancer regions and therefore, it is not desirable to minimize false positive rate to extreme. In addition, we required that difference between two rates is below a fixed threshold. This is equivalent to maximizing the F-score, while keeping difference of true positive (TP) and true negative (TN) rate below a fixed threshold.

5. Acknowledgement

This work is support by NIH grant R01GM085226 to S.H. and UMD-UMB seed grant to S.H. and C.C.

References

1. G. A. Maston, S. K. Evans and M. R. Green, *Annu Rev Genomics Hum Genet* **7**, 29 (2006).
2. R. J. White, *Nat Rev Genet* **12**, 459 (2011).
3. S. Naranjo, K. Voesenek, E. de la Calle-Mustienes, A. Robert-Moreno, H. Kokotas, M. Grigoriadou, J. Economides, G. Van Camp, N. Hilgert, F. Moreno, B. Alsina, M. B. Petersen, H. Kremer and J. L. Gomez-Skarmeta, *Human genetics* **128**, 411 (2010).
4. L. A. Lettice, S. J. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill and E. de Graaff, *Hum Mol Genet* **12**, 1725 (2003).
5. L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio, *Proc Natl Acad Sci U S A* **106**, 9362 (2009).
6. D. J. Gaffney, J. B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai, G. E. Crawford, M. Stephens, Y. Gilad and J. K. Pritchard, *Genome Biol* **13**, p. R7 (2012).
7. V. Gotea, A. Visel, J. M. Westlund, M. A. Nobrega, L. A. Pennacchio and I. Ovcharenko, *Genome Res* **20**, 565 (2010).
8. J. Ernst and M. Kellis, *Nat Biotechnol* **28**, 817 (2010).
9. G. E. Zentner, P. J. Tesar and P. C. Scacheri, *Genome Res* **21**, 1273 (2011).
10. R. Birnbaum, E. Clowney, O. Agamy, M. Kim, J. Zhao, T. Yamanaka, Z. Pappalardo, S. Clarke, A. Wenger, L. Nguyen *et al.*, *Genome Research* **22**, 1059 (2012).
11. J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis and B. E. Bernstein, *Nature* **473**, 43 (2011).
12. A. Visel, M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin and L. A. Pennacchio, *Nature* **457**, 854 (2009).
13. M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young and R. Jaenisch, *Proc Natl Acad Sci U S A* **107**, 21932 (2010).
14. D. Lee, R. Karchin and M. A. Beer, *Genome Res* **21**, 2167 (2011).

15. M. Fernandez and D. Miranda-Saavedra, *Nucleic Acids Res* **40**, p. e77 (2012).
16. D. May, M. Blow, T. Kaplan, D. McCulley, B. Jensen, J. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright *et al.*, *Nature genetics* (2011).
17. L. Narlikar, N. Sakabe, A. Blanski, F. Arimura, J. Westlund, M. Nobrega and I. Ovcharenko, *Genome research* **20**, 381 (2010).
18. A. D. Johnson, R. E. Handsaker, S. L. Pulit, M. M. Nizzari, C. J. O'Donnell and P. I. de Bakker, *Bioinformatics* **24**, 2938 (2008).
19. N. Frey and E. N. Olson, *Annu Rev Physiol* **65**, 45 (2003).
20. S. Hannenhalli, M. E. Putt, J. M. Gilmore, J. Wang, M. S. Parmacek, J. A. Epstein, E. E. Morrisey, K. B. Margulies and T. P. Cappola, *Circulation* **114**, 1269 (2006).
21. I. Manukyan, J. Galatioto, E. Mascareno, S. Bhaduri and M. A. Siddiqui, *J Cell Mol Med* **14**, 1707 (2010).
22. J. Schlesinger, M. Schueler, M. Grunert, J. J. Fischer, Q. Zhang, T. Krueger, M. Lange, M. Tonjes, I. Dunkel and S. R. Sperling, *PLoS Genet* **7**, p. e1001313 (2011).
23. V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel and E. Wingender, *Nucleic Acids Res* **34**, D108 (2006).
24. S. Levy and S. Hannenhalli, *Mamm Genome* **13**, 510 (2002).
25. A. G. West and P. Fraser, *Hum Mol Genet* **14 Spec No 1**, R101 (2005).
26. J. Chen, E. E. Bardes, B. J. Aronow and A. G. Jegga, *Nucleic Acids Res* **37**, W305 (2009).
27. C. Burges, *Data mining and knowledge discovery* **2**, 121 (1998).
28. J. Suykens and J. Vandewalle, *Neural processing letters* **9**, 293 (1999).
29. B. Boser, I. Guyon and V. Vapnik, 144 (1992).
30. N. Cristianini and J. Shawe-Taylor, *An introduction to support Vector Machines: and other kernel-based learning methods* (Cambridge Univ Pr, 2000).
31. V. Vapnik, *The nature of statistical learning theory* (Springer-Verlag New York Inc, 2000).
32. C. Cortes and V. Vapnik, *Machine learning* **20**, 273 (1995).

Identification of Aberrant Pathway and Network Activity from High-Throughput Data

Rachel Karchin
Department of Biomedical Engineering and Institute for Computational Medicine
Johns Hopkins University
Baltimore, MD 21218, USA
Email: karchin{at}jhu.edu

Michael F. Ochs
Department of Oncology and Division of Oncology, Biostatistics and Bioinformatics
Sidney Kimmel Comprehensive Cancer Center
Johns Hopkins University
Baltimore, MD 21205, USA
Email: mfo{at}jhu.edu

Joshua M. Stuart
Biomolecular Engineering
University of California Santa Cruz
Santa Cruz, CA 95064, USA
Email: jstuart{at}soe.ucsc.edu

Trey Ideker
Departments of Medicine and Bioengineering
University of California, San Diego
9500 Gilman Drive, Mail Code 0688
La Jolla, CA 92093-0688

Joel S. Bader
Department of Biomedical Engineering and High-Throughput Biology Center
Johns Hopkins University
Baltimore, MD 21218, USA
Email: joel.bader{at}jhu.edu

January 3-7, 2012

Overview

Biology has become an information science, with an increasing capacity to generate data of great relevance to human disease. An important example is The Cancer Genome Atlas (TCGA) [1], which generates data on well-characterized oncology samples and provides a public portal for linking gene mutation and regulation to cancer therapies and outcomes. These types of well-characterized data sets provide an opportunity for researchers from many fields to contribute new ideas for computational analysis.

One theme represented in the 2013 Proceedings is analysis of such public data sets by algorithms known from computer science but less often applied in computational biology and bioinformatics. Previous types of algorithms have included support vector machines [2], graph diffusion [3, 4, 5], and Steiner trees [6, 7]. Algorithms represented this year include set cover (Przytycka and coworkers), color-coded paths (Kahveci and coworkers), and regularized regression (Gevart and Plevritis).

A second theme is using known biological networks and pathways to organize calculations. Perhaps the most prevalent example is Gene Set Enrichment Analysis (GSEA) [8]. Lussier and co-workers describe extensions of GSEA to data sets from individuals rather than groups, and Ritchie and coworkers use interactions to organize analysis of interaction terms in genome-wide association studies (GWAS).

New algorithms from computer science

Przytycka and coworkers extend a set-cover algorithm from genes [9] to modules. These cover algorithms work on bipartite graphs, here with one set of vertices representing disease cases, a second set of vertices representing features (genes or gene modules), and edges indicating that the gene or module is dysregulated in a specific disease case. The k-cover optimization problem is to identify the smallest number of features so that each case has edges to at least k features. The authors generalize this NP-hard problem by also assigning a cost for each module that is reduced when the genes within the module have concordant expression regulation. A fast, greedy forward selection adds modules incrementally, either from a pre-calculated set or by defining modules on the fly. The method is effective in recovering known subtypes of glioblastoma multiforme. This type of approach, based on support, recalls approaches such as the Apriori algorithm for itemset mining [10] and the Teiresias algorithm for pattern discovery [11].

Kahveci and coworkers investigate an algorithm to identify signaling pathways of defined length. For a pathway desired to have m steps, a possible algorithm explored is to color each vertex one of m colors, and then to search for paths that include one vertex of each color. It remains to be seen whether this method is competitive with other related approaches, such as prize-collecting Steiner trees [7] and flow-based methods [12] that have fast, optimal solvers. The restriction to length m paths is motivated by a requirement that signaling pathways include a membrane-bound receptor, cytoplasmic signaling proteins, and nuclear transcription factors; constraints based on this biology and directed interactions may also perform better than path length restrictions.

Gevart and Plevritis also describe methods motivated by TCGA data. This approach generally follows successful methods introduced by others that use genetic and epigenetic features (copy number variation, methylation) to suggest driver genes, and then build out downstream pathways using regularized regression [13, 14] or other network-based association tests [15]. While predictions of expression perform better than random for an ovarian cancer data set, the top drivers predicted for a gliobastoma multiforme data set perform no better than a random collection. These results point to the uncertainty of applying established algorithms to new data sets and the importance of randomization tests for unbiased assessment of performance.

Pathways as a guide to analysis

Lussier and coworkers investigate personalized RNA-seq data by generalizing a single-sample method they developed for microarray data [16]. The main idea is to generate pathway scores by comparing expression levels between pathway and non-pathway genes. The authors find that converting raw expression values to ranks improves performance for many tasks. While the method is assessed to be feasible, traditional analysis of sample groups still appears to out-perform single-sample analysis.

Ritchie and coworkers investigate interaction terms in genome-wide association studies. Gene-environment interactions are already addressed by conventional methods, but gene-gene interactions are more challenging for both computational and statistical reasons. Computing all gene-gene interactions, or more accurately SNP-SNP interactions, incurs a large computational cost. Furthermore, the large number of tests requires an interaction term to be large for adequate power. The method proposed by Ritchie and coworkers, and also explored by others previously, is to restrict tests to SNPs to pairs in genes that have prior evidence for participating in a shared biological process or pathway. The threshold for evidence is increased until the candidate pairs are reduced to an acceptably small number, for example equivalent to the number of single-SNP tests. One challenge with including interaction terms is that tests for marginal effects may actually have greater power even when the interaction term is non-zero. For example, dominant and recessive genetic models are equivalent to interaction terms at a single locus, and a one degree-of-freedom test of a linear model for phenotype versus allele dose can have greater power than a two degree-of-freedom test that includes the interaction term. In an application to a cataract phenotype, the authors test 57,376 two-SNP models, requiring a p-value of 8.7×10^{-7} for genome-wide significance. The best p-value is 3.4×10^{-6}, however, typical of other searches for that have failed to identify interactions with statistical significance. While it may be feasible to identify interaction terms with greater power from larger population sizes, the lack of significance sets an upper limit on the magnitude of interaction terms and hence a possible limit on the biological relevance. Furthermore, it remains unclear whether genes identified through interaction terms would have been missed by conventional marginal tests on individual SNPs.

Future perspective

The contributions to this Proceedings consider two types of network models: on the one hand pre-calculated modules or curated pathways, on the other modules or pathways discovered from biological data. An important future direction may be module searches that use high-throughput data but are biased by existing network models. Generative models, such as stochastic block models, may provide an appropriate framework for network analysis biased by empirical knowledge. These models have received increasing attention for both static module discovery and dynamic network evolution [17, 18, 19, 20].

A critical limitation of network biology is the limited amount of high-quality network data. High-throughput protein-protein interaction data sets are available for human [21] but are incomplete [22, 23, 24]. Interactions between transcription factors to regulated genes provide crucial links between protein signaling and gene regulation, but are even less well mapped for human. Experimental progress here could result in dramatic gains for computational methods that already exist but which have been limited by lack of data.

References

[1] International Cancer Genome Consortium, Thomas J Hudson, Warwick Anderson, Axel Artez, Anna D Barker, Cindy Bell, Rosa R Bernabé, M K Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, Alan Guttmacher, Mark Guyer, Fiona M Hemsley, Jennifer L Jennings, David Kerr, Peter Klatt, Patrik Kolar, Jun Kusada, David P Lane, Frank Laplace, Lu Youyong, Gerd Nettekoven, Brad Ozenberger, Jane Peterson, T S Rao, Jacques Remacle, Alan J Schafer, Tatsuhiro Shibata, Michael R Stratton, Joseph G Vockley, Koichi Watanabe, Huanming Yang, Matthew M F Yuen, Bartha M Knoppers, Martin Bobrow, Anne Cambon-Thomsen, Lynn G Dressler, Stephanie O M Dyke, Yann Joly, Kazuto Kato, Karen L Kennedy, Pilar Nicolás, Michael J Parker, Emmanuelle Rial-Sebbag, Carlos M Romeo-Casabona, Kenna M Shaw, Susan Wallace, Georgia L Wiesner, Nikolajs Zeps, Peter Lichter, Andrew V Biankin, Christian Chabannon, Lynda Chin, Bruno Clément, Enrique de Alava, Françoise Degos, Martin L Ferguson, Peter Geary, D Neil Hayes, Thomas J Hudson, Amber L Johns, Arek Kasprzyk, Hidewaki Nakagawa, Robert Penny, Miguel A Piris, Rajiv Sarin, Aldo Scarpa, Tatsuhiro Shibata, Marc van de Vijver, P Andrew Futreal, Hiroyuki Aburatani, Mónica Bayés, David D L Botwell, Peter J Campbell, Xavier Estivill, Daniela S Gerhard, Sean M Grimmond, Ivo Gut, Martin Hirst, Carlos López-Otín, Partha Majumder, Marco Marra, John D McPherson, Hidewaki Nakagawa, Zemin Ning, Xose S

Puente, Yijun Ruan, Tatsuhiro Shibata, Michael R Stratton, Hendrik G Stunnenberg, Harold Swerdlow, Victor E Velculescu, Richard K Wilson, Hong H Xue, Liu Yang, Paul T Spellman, Gary D Bader, Paul C Boutros, Peter J Campbell, Paul Flicek, Gad Getz, Roderic Guigó, Guangwu Guo, David Haussler, Simon Heath, Tim J Hubbard, Tao Jiang, Steven M Jones, Qibin Li, Nuria López-Bigas, Ruibang Luo, Lakshmi Muthuswamy, B F Francis Ouellette, John V Pearson, Xose S Puente, Victor Quesada, Benjamin J Raphael, Chris Sander, Tatsuhiro Shibata, Terence P Speed, Lincoln D Stein, Joshua M Stuart, Jon W Teague, Yasushi Totoki, Tatsuhiko Tsunoda, Alfonso Valencia, David A Wheeler, Honglong Wu, Shancen Zhao, Guangyu Zhou, Lincoln D Stein, Roderic Guigó, Tim J Hubbard, Yann Joly, Steven M Jones, Arek Kasprzyk, Mark Lathrop, Nuria López-Bigas, B F Francis Ouellette, Paul T Spellman, Jon W Teague, Gilles Thomas, Alfonso Valencia, Teruhiko Yoshida, Karen L Kennedy, Myles Axton, Stephanie O M Dyke, P Andrew Futreal, Daniela S Gerhard, Chris Gunter, Mark Guyer, Thomas J Hudson, John D McPherson, Linda J Miller, Brad Ozenberger, Kenna M Shaw, Arek Kasprzyk, Lincoln D Stein, Junjun Zhang, Syed A Haider, Jianxin Wang, Christina K Yung, Anthony Cros, Anthony Cross, Yong Liang, Saravanamuttu Gnaneshan, Jonathan Guberman, Jack Hsu, Martin Bobrow, Don R C Chalmers, Karl W Hasel, Yann Joly, Terry S H Kaan, Karen L Kennedy, Bartha M Knoppers, William W Lowrance, Tohru Masui, Pilar Nicolás, Emmanuelle Rial-Sebbag, Laura Lyman Rodriguez, Catherine Vergely, Teruhiko Yoshida, Sean M Grimmond, Andrew V Biankin, David D L Bowtell, Nicole Cloonan, Anna deFazio, James R Eshleman, Dariush Etemadmoghadam, Brooke B Gardiner, Brooke A Gardiner, James G Kench, Aldo Scarpa, Robert L Sutherland, Margaret A Tempero, Nicola J Waddell, Peter J Wilson, John D McPherson, Steve Gallinger, Ming-Sound Tsao, Patricia A Shaw, Gloria M Petersen, Debabrata Mukhopadhyay, Lynda Chin, Ronald A DePinho, Sarah Thayer, Lakshmi Muthuswamy, Kamran Shazand, Timothy Beck, Michelle Sam, Lee Timms, Vanessa Ballin, Youyong Lu, Jiafu Ji, Xiuqing Zhang, Feng Chen, Xueda Hu, Guangyu Zhou, Qi Yang, Geng Tian, Lianhai Zhang, Xiaofang Xing, Xianghong Li, Zhenggang Zhu, Yingyan Yu, Jun Yu, Huanming Yang, Mark Lathrop, Jörg Tost, Paul Brennan, Ivana Holcatova, David Zaridze, and Alvis... Brazma. International network of cancer genome projects. *Nature*, 464(7291):993–998, April 2010.

[2] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines.* And Other Kernel-Based Learning Methods. Cambridge Univ Pr, March 2000.

[3] S Brin and L Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks And Isdn Systems*, 30(1-7):107–117, 1998.

[4] Yan Qi, Yasir Suhail, Yu-yi Lin, Jef D Boeke, and Joel S Bader. Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic

interactions and co-complex membership from yeast genetic interactions. *Genome research*, 18(12):1991–2004, December 2008.

[5] Fabio Vandin, Patrick Clay, Eli Upfal, and Benjamin J Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 55–66, 2012.

[6] I Ljubic, R Weiskircher, U Pferschy, GW Klau, P Mutzel, and M Fischetti. An algorithmic framework for the exact solution of the prize-collecting Steiner tree problem. *Mathematical Programming*, 105(2-3):427–449, 2006.

[7] Marcus T Dittrich, Gunnar W Klau, Andreas Rosenwald, Thomas Dandekar, and Tobias Müller. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–31, July 2008.

[8] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[9] Yoo-Ah Kim, Stefan Wuchty, and Teresa M Przytycka. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Computational Biology*, 7(3):e1001095, March 2011.

[10] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc, September 1994.

[11] I Rigoutsos and A Floratos. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, 14(1):55–67, 1998.

[12] Esti Yeger-Lotem, Laura Riva, Linhui Julie Su, Aaron D Gitler, Anil G Cashikar, Oliver D King, Pavan K Auluck, Melissa L Geddie, Julie S Valastyan, David R Karger, Susan Lindquist, and Ernest Fraenkel. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature genetics*, 41(3):316–323, March 2009.

[13] Uri David Akavia, Oren Litvin, Jessica Kim, Felix Sanchez-Garcia, Dylan Kotliar, Helen C Causton, Panisa Pochanard, Eyal Mozes, Levi A Garraway, and Dana Pe'er. An Integrated Approach to Uncover Drivers of Cancer. *Cell*, 143(6):1005–1017, December 2010.

[14] Su-In Lee, Aimée M Dudley, David Drubin, Pamela A Silver, Nevan J Krogan, Dana Pe'er, and Daphne Koller. Learning a Prior on Regulatory Potential from eQTL Data. *PLoS Genetics*, 5(1):e1000358, January 2009.

[15] Kartik M Mani, Celine Lefebvre, Kai Wang, Wei Keat Lim, Katia Basso, Riccardo Dalla Favera, and Andrea Califano. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Molecular systems biology*, 4:169, 2008.

[16] Xinan Yang, Kelly Regan, Yong Huang, Qingbei Zhang, Jianrong Li, Tanguy Y Seiwert, Ezra E W Cohen, H Rosie Xing, and Yves A Lussier. Single sample expression-anchored mechanisms predict survival in head and neck cancer. *PLoS Computational Biology*, 8(1):e1002350, January 2012.

[17] Aaron Clauset, Cristopher Moore, and M E J Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, May 2008.

[18] J Hofman and C Wiggins. Bayesian approach to network modularity. *Physical Review Letters*, 2008.

[19] Joel S Bader and Yongjin Park. How networks change with time. *Bioinformatics*, 28(12):i40–i48, January 2012.

[20] Yongjin Park and Joel S Bader. Resolving the structure of interactomes with hierarchical agglomerative clustering. *BMC Bioinformatics*, 12 Suppl 1:S44, 2011.

[21] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S Goldberg, Lan V Zhang, Sharyl L Wong, Giovanni Franklin, Siming Li, Joanna S Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S Sikorski, Jean Vandenhaute, Huda Y Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E Cusick, David E Hill, Frederick P Roth, and Marc Vidal. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, October 2005.

[22] G Traver Hart, Arun K Ramani, and Edward M Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120, 2006.

[23] Hailiang Huang, Bruno M Jedynak, and Joel S Bader. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Computational Biology*, 3(11):e214, November 2007.

[24] Hailiang Huang and Joel S Bader. Precision and recall estimates for two-hybrid screens. *Bioinformatics*, 25(3):372–378, February 2009.

FROM UNCERTAIN PROTEIN INTERACTION NETWORKS TO SIGNALING PATHWAYS THROUGH INTENSIVE COLOR CODING

Haitham Gabr, Alin Dobra and Tamer Kahveci*

CISE Department, University of Florida,
Gainesville, FL 32611, USA
E-mail: {hgabr, adobra, tamer}@cise.ufl.edu*

Discovering signaling pathways in protein interaction networks is a key ingredient in understanding how proteins carry out cellular functions. These interactions however can be uncertain events that may or may not take place depending on many factors including the internal factors, such as the size and abundance of the proteins, or the external factors, such as mutations, disorders and drug intake. In this paper, we consider the problem of finding causal orderings of nodes in such protein interaction networks to discover signaling pathways. We adopt color coding technique to address this problem. Color coding method may fail with some probability. By allowing it to run for sufficient time, however, its confidence in the optimality of the result can converge close to 100%. Our key contribution in this paper is elimination of the key conservative assumptions made by the traditional color coding methods while computing its success probability. We do this by carefully establishing the relationship between node colors, network topology and success probability. As a result our method converges to any confidence value much faster than the traditional methods. Thus, it is scalable to larger protein interaction networks and longer signaling pathways than existing methods. We demonstrate, both theoretically and experimentally that our method outperforms existing methods.

Keywords: protein interaction networks; signaling pathways; color coding; chromatic polynomial

1. Introduction

Studying interactions between proteins has been of utmost importance in understanding how proteins work collectively to govern cellular functions.[1,2] Such collection of interactions among proteins is called a protein interaction network. The interactions are uncertain events. They may or may not take place depending on the internal factors, such as the size and abundance of the proteins, or the external factors, such as mutations, disorders and drug intake. Mathematically, a protein interaction network is often modeled as an edge-weighted undirected graph where each node denotes a protein and each edge represents an interaction between a pair of proteins. The weight of an edge denotes the level of confidence that this interaction truly exists.

Computational analysis of protein interaction networks has been essential in identification of signaling pathways. A signaling pathway is a series of proteins in which each protein participates in transmitting biological information by modifying its successor through an interaction. Thus, signaling pathways can be viewed as simple paths in protein interaction networks.[3] One outcome of the uncertainty of the interactions is that the pathway that transmits signals between two specific sets of proteins (e.g., from membrane receptors to transcription factors) may differ as the set of interactions change. Finding possible pathways in the presence of such uncertainty has great potential in numerous applications including identification of drug targets, studying complex diseases, drug-drug interaction and metabolic engineering.

The confidence value of an interaction between two proteins is often considered as the probability that a signal is transmitted between those two proteins. Scott *et al.* conjectured that a signal tends to move through the most probable pathway[4] (i.e., the pathway with the highest product of interaction confidence values). The following defines the *Minimum Weight Pathway Identification* problem which is identical to the problem of identifying the most probable pathway in a protein interaction network.

Problem. (Minimum Weight Pathway Identification) Consider a protein interaction network $G = (V, E, w)$ where V denotes the set of proteins and E denotes the set of interactions. Let us denote the confidence for each interaction in E with function $\lambda() : E \Rightarrow [0, 1]$. We define the function $w()$ on the edges as $w() = -\log \lambda()$. Assume that we are given a set of starting proteins $S \subseteq V$ and a set of target proteins $T \subseteq V$. Given a path length denoted by m, the problem is to find a path $\Phi = v_1 \rightarrow v_2 \rightarrow \ldots \rightarrow v_m$ with no repeating proteins, where $\sum_{i=1}^{m-1} w(v_i, v_{i+1})$ is the minimum among all paths with $v_1 \in S$, $v_m \in T$ and $v_i \in V$, $\forall i \in \{1, 2, \ldots, m\}$.

Scott *et al.* showed that the traveling-salesman problem is polynomial-time reducible to the problem above;[4] therefore it is NP-hard. They developed a method using the *color-coding* technique of Alon *et al.*[5] The idea of this method is to randomly assign each node in the graph one of m different colors. A pathway is *colorful* if and only if all of its nodes are in different color. They then search for an optimal colorful pathway. Finding a colorful path is computationally much cheaper than finding a path without assigning colors. The drawback is that the optimal path may not be colorful in a random color assignment, leading color coding to find a sub-optimal result. To deal with this, it repeats the coloring process for several iterations. The confidence in the optimality of the result monotonically increases with each iteration until it reaches a given level of confidence. As we elaborate later in Section 2, the confidence value depends solely on the pathway length m and does not capitalize on readily available information such as the network topology and color assignment. As a result, the method provides a theoretically correct but very conservative confidence value. Hence it requires many iterations in order to achieve a given confidence level, leading to an unnecessarily inefficient running time performance.

Gülsoy *et al.*[6] presented an enhanced color-coding technique called *k-hop coloring.* A colored network is k-hop colorable if the shortest path between all pairs of same-color nodes is more than k hops in length. This method exploits the network topology and the node colors to assign the network a maximal value k such that the network is k-hop colorable. This additional piece of information allows for higher success probability at each iteration, yielding fewer iterations than that by Scott *et al.*[4] However, subnetworks with high connectivity quickly diminish the ability to k-hop color the whole network for large values of k. For example, a network containing a clique of size m cannot be colored with $(m - 1)$-hop coloring using m colors.[6]

Our contribution. In this paper, we consider the problem of finding signaling pathways in protein interaction networks. We develop a new coloring method that overcomes the bottlenecks of existing coloring methods by Scott *et al.*[4] and Gülsoy *et al.*[6] Our contribution comes from a deeper understanding of the relation between network topology, random color assignment and confidence value. We assign a value that we call k_{max} to each node individually by studying the colors of all the nodes in the network. k_{max} value of a node v at an iteration is the maximal value of k such that there is no other node u that is reachable from v in k hops such that both u and v have the same color. We also study how this reflects on the resulting success probability for each iteration. Given different k_{max} values for each node on a pathway, we show how to obtain a bound on success probability. Based on these findings, we present a new method for detecting signaling pathways in protein interaction networks using an enhanced k-hop coloring technique. Given the parameter pathway length m, we start by randomly assigning one of m colors to each node in the graph, we then extract the optimal colorful pathway. We then calculate our new bound on success probability. We repeat this process until the cumulative success probability is at least equal to a given confidence level. Our experiments demonstrate that our method converges to high confidence values much faster than the existing methods including Scott *et al.*[4] This enables computational analysis of larger networks or longer pathways.

The rest of the paper is organized as follows. Section 2 discusses the background and related work. Section 3 describes our method in detail. Section 4 presents experiments evaluation. Finally, Section 5 concludes the paper.

2. Background

A number of methods have been developed so far to identify signaling networks from protein interaction networks. Kelley *et al.*[3] detected conserved signaling pathways between related organisms by performing global alignment between their protein interaction networks. Shlomi *et al.*[7] introduced QPath, a method for querying protein interaction networks for pathways using known homologous pathways as queries. Both Kelley *et al.*[3] and Shlomi *et al.*[7] are comparative methods. They require knowledge of multiple interaction networks.

Lu *et al.*[8] presented a divide-and-conquer algorithm to find signaling subnetworks in protein interaction networks. They scored the resulting subnetworks based on the similarity of expression profiles of their nodes to the given source and destination nodes. This method aims to detect paths whose proteins are highest in expression similarity, and thus it does not utilize the confidence in the interactions. Steffen *et al.*[9] used exhaustive search to list pathway candidates in protein interaction networks, and scored each one based on how similar the expression profiles of its genes are. Bebek *et al.*[10] presented a method for finding new signaling pathways using association rules of known ones. The time complexity of exhaustive graph search is exponential in terms of the network size, and hence is very inefficient. Gitter *et al.*[11] presented a method for discovering signaling pathways by adding edge orientation to protein interaction networks. They selected an optimal orientation of all edges in the network that maximizes the weights of all satisfied length-bound paths. They proved that this problem is NP-hard, and provided three approximation algorithms for it. As shown in their results, these methods do not scale well with increasing number of source and destination nodes and path length.

The closest studies to that presented in this paper are those by Scott *et al.*[4] and Gülsoy *et al.*[6] The former detected signaling pathways in protein interaction networks using color coding. The latter developed topology-aware color coding for network alignment. We describe both methods in Section 1. Both methods run multiple coloring iterations. Let us denote the probability that the coloring at an iteration is successful (i.e., true optimal path is colorful) with P_s. The probability that at least one out of r iterations is successful is $1 - (1 - P_s)^r$. Following from this, in order to ensure confidence of at least ϵ $(0 \leq \epsilon \leq 1)$, they run r iterations, such that $1 - (1 - P_s)^r \geq \epsilon$. Both methods calculate success probability as

$$P_s = \frac{m!}{N_c} \tag{1}$$

where N_c is the number of coloring assignments possible for the optimal pathway. They differ in the way they compute N_c. Scott *et al.*[4] calculated $N_c = m^m$. Gülsoy *et al.*[6] calculated $N_c \leq (m-k)^{m-k} \prod_{i=0}^{k-1}(m-i)$ where k is the value assigned to the network such that it is k-hop colorable. Notice that in Equation 1, smaller values for N_c are desirable. This is because small values for N_c increase success probability, and thus reduce the number of iterations needed to attain a given confidence level ϵ. *This paper develops a novel method that computes a much smaller upper bound on N_c than both of these approaches, leading to higher bound on P_s.*

3. Method description

This section describes our method in detail. Section 3.1 presents a high level description of our method. Section 3.2 makes key definitions needed by our method. Section 3.3 defines how we compute probability of success for our method. Section 3.4 theoretically shows why the performance of our method is better than or the same as that of existing methods.

3.1. *An overview of our method*

Consider a weighted undirected graph $G = (V, E, w)$, a path length m, a set of starting and target nodes S and T respectively, with $S, T \subseteq V$. Scott *et al.* has shown that it is possible to find the minimum weight path of a m nodes from S to T in G using dynamic programming.[4] In principle, our method follows the same steps. Algorithm 3.1 presents our method at a high level. The algorithm works iteratively. At each iteration we randomly color the network (Step 3). We then use dynamic programming to find the minimum weight colorful path (Step 4). The dynamic programming works as follows. Let us denote a coloring function with $c() : V \Longrightarrow C$. We dynamically tabulate the minimum weight of a colorful path colored only using C', starting within S and ending at v, using the following recurrence:[4]

$$W(v, C') = \min_{u:c(u)\in(C'\backslash\{c(v)\})} W(u, C'\backslash\{c(v)\}) + w(u, v), |C'| > 1 \tag{2}$$

where $W(v, \{c(v)\}) = 0$ if $v \in S$ and ∞ otherwise. Once we find the best colorful path in that iteration, we store it in a min-heap according to the weight of the path (Step 5). We then compute the probability that the current iteration was successful in finding the optimal path (i.e., minimum weighted path regardless of being colorful or not) (Step 6) and update our confidence in the best result seen so far (Step 7).

Algorithm 3.1 Compute the minimum weight path

```
Require: Input network $G = (V, E, w)$, starting and target node sets $S \subseteq V$ and $T \subseteq V$
Require: Color set $C = \{c_1, c_2, \ldots, c_m\}$
Require: Confidence cutoff $\epsilon$
1: $P \leftarrow 0$ {Initialize overall success probability}
2: while $P < \epsilon$ do
3:     Assign colors to the nodes in $V$ randomly from the set $C$
4:     $\Phi \leftarrow$ Find the minimum weight colorful path of length $m$ in $G$
5:     Store $\Phi$ in the min-heap of solutions observed so far if it is a new solution.
6:     Compute the probability of success $P_s$ for the current coloring iteration.
7:     $P \leftarrow 1 - (1 - P)(1 - P_s)$ {Update the overall success probability}
8: end while
```

As we noted earlier, Algorithm 3.1 is very similar to the method by Scott *et al.*[4] So, a legitimate question is what is the big challenge addressed in this paper? The answer lies in Step 6 of the algorithm where we compute the probability of success at each iteration. This step is missing in all the color coding methods to the best of our knowledge, including Scott *et al.*[4] among others.[5–7,12] All these existing methods precompute a probability of success prior to the iterations and use the same probability value throughout the iterations (see Equation 1 and Section 2). As a result, they make extremely conservative assumptions which have to hold regardless of which node gets which color. *Our contribution is to eliminate those worst case assumptions and recompute the probability of success at each iteration by carefully inspecting the colors of all the nodes.* We explain how we do this in the following sections.

3.2. *Basic definitions and model*

In this section, we build the mathematical model that will help us compute the probability of success in each iteration. Assume that we are given a protein interaction network similar to the one described in Section 1, denoted by $G = (V, E, w)$, where $w(u, v) = -\log \lambda(u, v)$. Also assume that the colors of the nodes are already assigned in the current iteration. We denote the set of possible colors with $C = \{c_1, c_2, \ldots, c_m\}$ and the color of node $v \in V$ with $c(v)$. We start by discuss several key concepts.

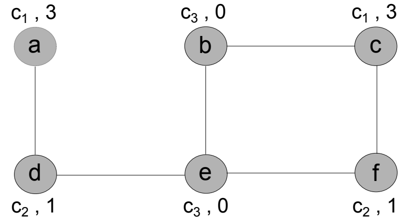


Fig. 1. A hypothetical protein interaction network with six nodes {a, b, c, d, e, f}. The network is colored using three colors $\{c_1, c_2, c_3\}$. Each node carries two labels. The label on the left denotes the color assigned to this node. The one on the right is the node's k_{max} value. For instance node d is assigned to color c_2 and its k_{max} value is 1 (i.e., there is no other node assigned to color c_2 within 1-hop of node d).

Definition 1. (Simple path) Given a network $G = (V, E)$, a *simple path* from u to v (u, $v \in V$) is an ordering $< v_1, v_2, \ldots, v_k >$, of a subset of the vertices of G such that $v_1 = u$, $v_k = v$, $(v_i, v_{i+1}) \in E$ and $v_i \neq v_j$ for all $i \neq j$.

Consider two nodes u and v in G. Let k be a positive integer. We say that v is *reachable* from u in k hops if there is a simple path from u to v that contains k edges.

Definition 2. (k neighborhood of a node). Let $v \in V$ be a node in G, and k be a nonnegative integer. We define the k neighborhood of node v as the set of nodes in $V \backslash \{v\}$ which are reachable from v in k hops or less. We denote this set using notation $\Psi_k(v)$.

Figure 1 shows an example of a colored network. In this example, $\Psi_1(a) = \{d\}$ because the node d is the only node that is reachable from the node a in 1 hop (or less). Similarly, $\Psi_1(f) = \{c, e\}$, $\Psi_2(a) = \{d, e\}$ and $\Psi_2(f) = \{c, e, b, d\}$. Following definition establishes the relationship between each node of the network and the rest of the network based on the colors assigned to all the nodes.

Definition 3. (k_{max} value of a node). Let $v \in V$ be a node in a colored network G. The k_{max} value of v, denoted with $k_{max}(v)$, is the maximal value of k such that the k neighborhood of v does not contain a node with the same color as v. i.e., $k_{max}(v) = \text{argmax}_k\{\forall u \in \Psi_k(v), c(u) \neq c(v)\}$.

Figure 1 shows the k_{max} values for the nodes in the network. For example, the colors of all the nodes in $\Psi_1(f) = \{c, e\}$ are different than the color of f. Expanding the neighborhood of f to two, we get $\Psi_2(f) = \{c, e, b, d\}$. In this set, $c(d) = c(f) = c_2$. Therefore $k_{max}(f) = 1$. Similarly, $k_{max}(a) = 3$ and $k_{max}(b) = 0$. Next definition characterizes a simple path of the network.

Definition 4. (k_{max} CONFIGURATION OF A PATH). Consider a simple path $\Phi = v_1 \to \ldots \to v_m$ of m nodes in G. The k_{max} configuration of Φ is the vector $[k_{max}(v_1), \ldots, k_{max}(v_m)]$.

As an example, in Figure 1, the k_{max} configuration of the path $\Phi = a \to d \to e \to f$ is [3, 1, 0, 1]. That for $a \to d \to e \to b$ is [3, 1, 0, 0].

3.3. *Bounding the probability of success tightly*

In this section, we focus on one coloring iteration and describe how we compute the probability of success in that iteration. Consider any colorful path with m nodes. The number of ways to assign colors to the nodes of that path while keeping it colorful is $m!$. Notice that this is equal to the numerator in Equation 1 for probability of success. The denominator in that equation, denoted by N_c, is the total number of ways to color that path regardless of whether it yields a colorful path or not.

Notice that there can be many different color assignments that yield the same k_{max} configuration for the same path. Also, as we will show later, the number of possible color assignments to the nodes of a path can be different for different k_{max} configurations. Indeed, the k_{max} configuration of a path describes the constraints imposed on all the nodes of that path about how many alternative colors can be assigned to them. Following from this observation, we first build a new undirected and unweighted graph, called the *constraint graph* from the k_{max} configuration. By utilizing the constraint graph we transform the problem of finding the number of possible colorings to the chromatic polynomial computation problem. Next, we describe how we build the constraint graph and how we utilize it to find the number of colorings.

Building the constraint graph. Assume that we are given a simple path $\Phi = v_1 \to v_2 \to \ldots \to v_m$ of m nodes along with its k_{max} configuration $[k_{max}(v_1), k_{max}(v_2), \ldots, k_{max}(v_m)]$. We build a constraint graph with m nodes $\{u_1, u_2, \ldots, u_m\}$. We denote the constraint graph as $G^\Phi = (V^\Phi, E^\Phi)$ where V^Φ is its set of nodes and E^Φ is its set of edges. For each pair of nodes u_i and u_j in V^Φ, we draw an undirected edge between them if the following condition holds:

$$j - i \le \max\{k_{max}(v_i), k_{max}(v_j)\}.$$

Notice that the indices i and j above show the positions of the nodes on the given path Φ. As a result, an edge between u_i and u_j in the constraint graph indicates that v_i and v_j can not be of the same color according to the underlying k_{max} configuration. Consider any coloring instance I that obeys the k_{max} configuration. Let v_i and v_j be any two nodes having the same color in I. Therefore $j - i > \max\{k_{max}(v_i), k_{max}(v_j)\}$.

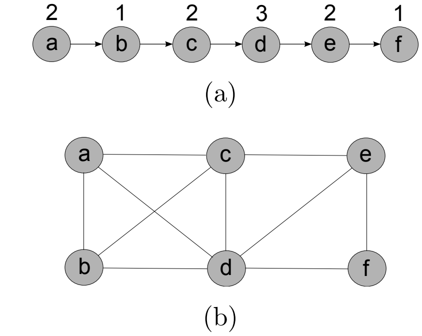


Fig. 2. (a) An example 6-node path with its k_{max} configuration shown above it. (b) The corresponding constraint graph G^Φ.

Therefore, the corresponding u_i and u_j in G^Φ are not adjacent. Hence, I also obeys the constraints of G^Φ. A similar argument can be made in reverse. Thus, each possible coloring of the given path Φ that obeys the k_{max} configuration corresponds to a chromatic coloring of the constraint graph G^Φ and vice versa. Figure 2 shows an example of a path, its k_{max} configuration and the corresponding constraint graph.

Computing the number of colorings. Formally, the value of the chromatic polynomial $A(G^\Phi, m)$ is equal to the number of ways of coloring G^Φ using m colors without any pair of adjacent nodes having the same color. Applying chromatic polynomials on the constraint graph of a path yields the number of possible colorings of that path. We use an edge-contraction recursive rule based on the fundamental reduction theorem.[13] To describe this, we first define two contraction operators on graph G^Φ. The first one removes one edge, (u, v) from the edge set of G^Φ. We denote this with $G^\Phi - (u, v)$. The second one merges two nodes, u and v, into a single node uv. To do this, we insert a new node uv to G^Φ. We also insert an edge between uv and all the nodes which are adjacent to either u or v. We then remove the nodes u and v along with all the edges incident to them. We denote this merge operation with $G^\Phi/\{u, v\}$. Using this notation, the chromatic polynomial is computed using the following recurrence relation

$$A(G^\Phi, m) = A(G^\Phi - (u, v), m) - A(G^\Phi/\{u, v\}, m) \tag{3}$$

Finally, an important question is: which path should we choose to use its corresponding k_{max} configuration as input to our method? Ideally, this path should be the optimal path that we don't know and are looking for. Instead, we use the optimal colorful path we find at each iteration. The main rationale behind this choice is that we expect that the local optimal path of a random coloring instance will have common nodes and edges with the overall optimal path. This is because the optimal path will contain edges with small weights. In Section 4.1 we empirically show that this indeed yields a good approximation to the value of P_s in practice.

Now we are ready to compute the probability of success, P_s, for a coloring instance of our method (i.e, Step 6 of Algorithm 3.1). At each iteration, we first build the constraint graph G^Φ of the best colorful path Φ found at that iteration. We compute the number of chromatic colorings of G^Φ as $A(G^\Phi, m)$ as described above. We then set $N_c = A(G^\Phi, m)$ and compute the probability of success using Equation 1 as $P_s = m!/N_c = m!/A(G^\Phi, m)$.

3.4. *Analysis of the probability of success*

One key question would regarding how we compute the probability of success is: Is it guaranteed to be better than existing methods including Scott *et al.*[4] and Gülsoy *et al.*[6]? In this section, we answer this theoretically. We start by defining a partial order between k_{max} configuration of a paths as follows: Consider two such configurations $\mathbf{x} = [x_1, x_2, \ldots, x_m]$ and $\mathbf{y} = [y_1, y_2, \ldots, y_m]$. We say that $\mathbf{x} \leq \mathbf{y}$ if and only if $\forall_i$, $x_i \leq y_i$.

Proposition 3.1. *Consider two k_{max} configurations $\mathbf{x}$ and $\mathbf{y}$ of two simple paths each having m nodes. Let us denote their corresponding constraint graphs G_x and G_y respectively. If $\mathbf{x} \leq \mathbf{y}$ then $A(G_x, m) \geq A(G_y, m)$.*

We omit detailed proof of Proposition 3.1 due to space limitation. However, briefly it follows from the observation that $\mathbf{x} \leq \mathbf{y}$ implies that every edge in G_x also appears in G_y. However, the opposite may not be true. In other words, G_x has only a subset of the constraints imposed by G_y. Thus, the chromatic polynomial $A(G_x, m)$ cannot be less than $A(G_y, m)$.

Proposition 3.1 has two important implications. First, traditional color coding method (such as Scott *et al.*[4]) computes $N_c = m^m$. This is the most conservative case in our model when the k_{max} configuration is [0, …, 0]. Clearly, this will yield the worst (i.e., largest) possible value for the chromatic polynomial since $[0, \ldots, 0] \leq \mathbf{y}$ for any k_{max} configuration $\mathbf{y}$. Second, let t be the smallest k_{max} value among all the nodes in the network. The formulation by Gülsoy *et al.*[6] corresponds to k_{max} configuration is $[t, \ldots, t]$. Let $\mathbf{y}$ be the k_{max} configuration of any m-node path in the same network. We have $[t, \ldots, t] \leq \mathbf{y}$ since all the entries of $\mathbf{y}$ have value t or more. *We conclude from these two implications that our method is guaranteed to produce less or same N_c value as the mentioned existing methods depending on the network topology and the color distribution. Smaller values for N_c implies larger success probability, and thus, faster convergence to the desired confidence value.*

As an example, our method computes the value of N_c for the path shown in Figure 2(a) is 5,760, while Scott *et al.*[4] and Gülsoy *et al.*[6] yield N_c = 46,656 and 18,750 respectively for the same example. According to Equation 1, such a decrease in the value of N_c leads 8.1 and 3.2 times larger success probability than the two above-mentioned methods respectively.

4. Experiments

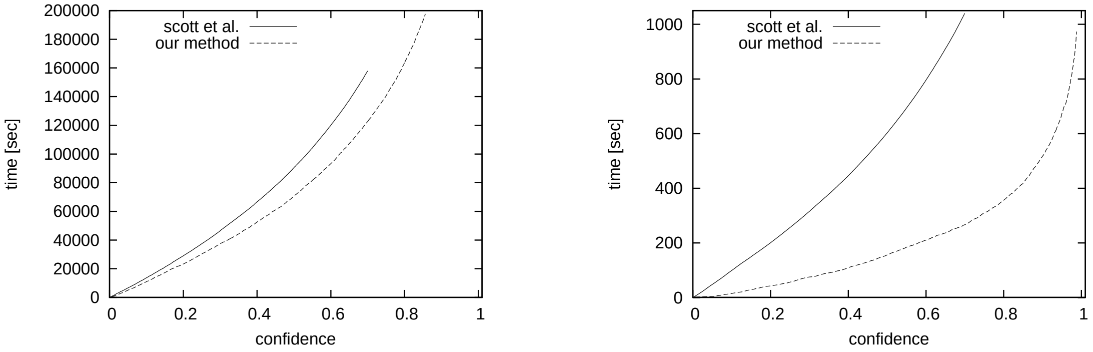
In this section, we evaluate our method on real protein interaction networks. We implemented our method in Java. We ran our experiments on Linux machines with 2.2-GHz dual AMD Opteron dual core processors and 3 GBs of main memory.

Datasets We used the protein interactions of *H. sapiens* and *R. norvegicus* taken from the MINT database.[14] The first one is a large dataset of 15,472 interactions among 6,122 proteins. The second one is a smaller dataset containing 806 interactions among 631 proteins. Each interaction is described by two interacting proteins and a reliability score between 0 and 1 that represents the level of confidence that this interaction exists. MINT calculates reliability scores of interactions from available evidence, such as the size and type of the experiment reporting the interaction, sequence similarity of ortholog proteins.[15]

We use the negative logarithm of MINT reliability scores as edge weights. In all experiments, we find pathways starting within the set of membrane proteins and ending within the set of transcription factors. We use the Gene Ontology database[16] to identify these sets. We identify membrane proteins as the ones annotated with the terms GO:0005886 and GO:0004872, and transcription factors as those with GO:0000988, GO:0001071 and GO:0006351.

4.1. *Performance assessment*

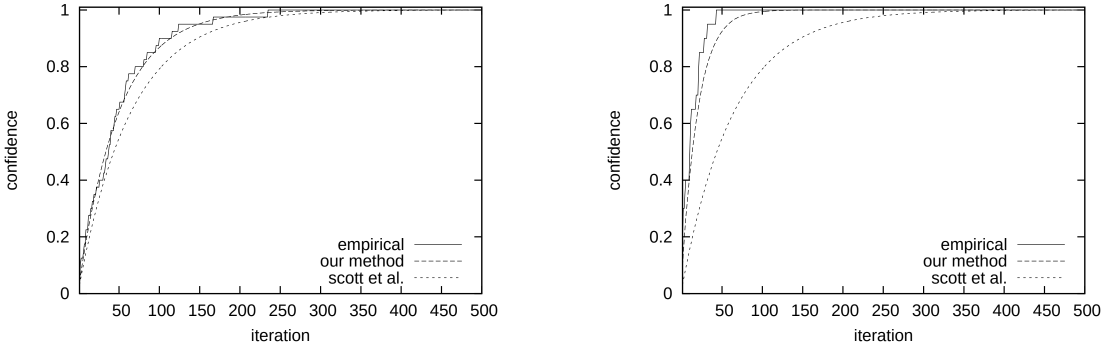
In Section 3.4, we have already shown theoretically that our method is guaranteed to be at least as fast as the traditional color coding methods. The gap however depends on the topology of the underlying protein interaction network. In this section, we experimentally evaluate how the performance of our method compares to Scott *et al.*[4] as a leading method. We run both methods on our datasets for 500 iterations. We repeat this experiment for pathway lengths = {4, 5, 6, 7, 8, 9}. We measure the total time taken and the confidence value computed by each method at each iteration. We run this process multiple times (at least 20 times) and report the



(a) *H.sapiens*

(b) *R.norvegicus*

Fig. 3. Total time needed to achieve a given level of confidence by our method and Scott *et al.* for *H.sapiens* and *R.norvegicus* for path length = 8.



(a) *H.sapiens*

(b) *R.norvegicus*

Fig. 4. Confidence level achieved after a given number of iterations using our method and to Scott *et al.* for *H.sapiens* and *R.norvegicus* when path length is fixed at 6. Empirical results denote the fraction of experiments in which the optimal path is found at or before a given iteration.

average of these runs. Below, we report a small subset of these experiments due to page limits.

Figure 3 shows the time it takes to reach to various confidence levels for path length = 8. Our method takes much less time than Scott *et al.* to achieve the same level of confidence. The gap between the two increases as the confidence level increases. We observe that the gap is significantly larger for the *R. norvegicus* dataset. This is mainly because this dataset is more sparse than the other one. As a result, it often produces very dense constraint graphs leading to high success probability values. Scott *et al.* is, on the other hand, oblivious to the density of the network. It produces the same conservative success probability for both datasets. As a result, as we can see in Figure 3, Scott *et al.* can only reach to around 70% confidence for both datasets after 500 iterations. Our method, on the other hand, reports 85% and more than 99% confidence for the *H. sapiens* and *R. norvegicus* datasets respectively after the same number of iterations. The difference between the largest confidence we report for the two datasets can be explained from the density of the two networks. As the network gets sparser, our method tends to gets larger confidence value. In Figure 3(a), we see that our method takes more time to complete

Table 1. Z-scores calculated for the optimal paths found by our method for *H.sapiens* and *R.norvegicus* for different path lengths. Here, μ is the mean of the weight of a random path in the same network with the same length. θ is the weight of the optimal path found by our method. Z is the Z-score of our method.

Dataset	Path Length								
	6			7			8		
	μ	θ	Z	μ	θ	Z	μ	θ	Z
H.sapiens	5.906	0.129	5.409	7.074	0.130	5.477	8.341	0.221	5.764
R.norvegicus	4.975	4.540	0.889	7.307	5.025	1.453	8.457	4.858	1.467

500 iterations than Scott *et al.* This is because it spends additional time to build constraint graph and solve a chromatic polynomial problem. Finally, we observed similar characteristics for other path lengths (results not shown). The main difference was that the performance gap between our method and Scott *et al.* further increases with larger path lengths.

In our next experiment, we evaluate whether our confidence computation is correct in practice. To do this, we computed an empirical confidence as follows. Recall that we repeated each experiment many times. At each iteration we computed the fraction of the experiments in which we were able to find the optimal result as the empirical confidence. Ideally, the theoretical value should not be larger than the empirical one; the closer the two values are the better. Figure 4 shows the empirical confidence value as well as the theoretical confidence value of our method and Scott *et al.*. The results demonstrate that the gap between the empirical results and our method is much smaller than that for Scott *et al.* This is because of the conservative way they use to calculate success probability of an iteration as discussed in section 2. This gap increases as the path length parameter increases (results not shown). Thus, we conclude that both Scott *et al.* and our method produces correct confidence values. Scott *et al.* is, however to conservative, and thus spends too many iterations to reach to the same confidence value.

4.2. *Validation Experiments*

So far we have shown that our method outperforms existing coloring strategies in terms of the running time performance. In this section, we evaluate the biological significance of the paths found using our method. It is worth mentioning that our method returns the same results as Scott *et al.*[4] when both of them are allowed to reach a high confidence value (such as 99% confidence). The main difference is that our method scales to larger networks and longer paths. Therefore, here we will only focus on the results obtained by our method.

4.2.1. *Statistical significance of the results*

In this section we assess the statistical significance of the paths found by our method. We use Z-score to measure statistical significance. Z-score indicates by how many standard deviations our optimal weight is better than the weight of an average random path, so higher values are better. For each dataset and path length m, we run our method to get the path with the minimum weight θ. We then generate 1000 random simple paths of length m, starting at a membrane protein and ending at a transcription factor. We compute the average weight μ of these random paths and their standard deviation σ. We then compute the Z-score as $Z = \frac{\mu-\theta}{\sigma}$.

Table 1 shows the results for *H.sapiens* and *R.norvegicus* for path lengths 6, 7 and 8. Our results are always better than the random paths. Particularly, for the *H.sapiens* network we obtain very significant results. The Z-score for *R.norvegicus* is less. This is mainly because the edge confidence values in this network have much less variation than those in *H.sapiens*. Our Z-score increases with increasing path length. This is not surprising because increasing the size of random selection leads to less chances of the selected path being better or closer to the optimal path. This implies that there is a great potential that methods that scale to large path length will yield important biological insights into signaling pathway identification.

4.2.2. *Biological significance of the results*

Another important question is: how biologically significant are our results? To answer this question, we validate our results using functional enrichment. We use the Gene Ontology[16] to compute functional enrichment of paths found at different iterations of our method. Let Φ be the path being tested, T be the universal set of GO terms, m be the path length, M be the total number of proteins in the dataset, G_i be the total number of proteins annotated with the Go term t_i in the dataset, and g_i be the number of proteins annotated with t_i in Φ. We compute functional enrichment of Φ as $\min_{t_i \in T} P(X \geq g_i | M, m, G_i)$ where X is a random variable under a hyper-geometric distribution with these parameters. Lower enrichment values indicate paths with common functions, and thus they are better.

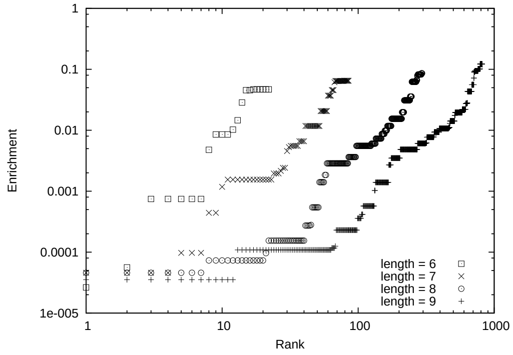


Fig. 5. Functional enrichment of best colorful paths found at different iterations of our method for *R.norvegicus* in sorted order. Smaller values are better.

Figure 5 plots the functional enrichment value of the best colorful paths found at different iterations of our algorithm in sorted order for the *R.norvegicus* network. We omit results for *H. sapiens* as it is very similar to those in Figure 5. We observe that as the distribution of the enrichment values follows power-law distribution. That is only a minority of the observed paths have very good enrichment while the majority tend to have bad ones. We observe that this behavior is consistent for all path lengths we tested. This suggests the following: (i) There can be multiple biologically interesting paths for the same start and end node sets. (ii) We need to have sufficiently high confidence in the result to avoid biologically meaningless paths since the enrichment drops quickly. (iii) Even long paths can be highly enriched. All of these observations show the importance of improving the running time performance of pathway discovery methods, and hence the importance of our contribution.

Next, we focus on a few of the most functionally enriched pathways our method finds on the *H. sapiens* network. Figure 6 shows three examples each having length of six. All the six genes in the path in Figure 6(a) regulate epidermal growth factor receptor signaling pathway. Among these the leftmost four genes appear in the ErbB signaling pathway. They also affect the development of various cancer types such as chronic myeloid leukemia, glioma and prostate cancer. In Figure 6(b), all the six genes are ephrin receptor binding. They affect cell growth and development and thus participate in cancer development. The five leftmost genes in Figure 6(c) negatively regulate the epidermal growth factor receptor signaling pathway. Notice that all the three pathways in this example overlap with each other, yet they also contain several genes that do not exist in others. For instance, the pathway in Figure 6(b) contains SRC unlike the other. SRC takes part in same pathways as most of the other genes in this figure, such as the ErbB signaling pathway. Thus, all of these significant paths reported by our method reveal different parts of the signaling networks through alternative paths.

5. Conclusion

In this paper, we presented an enhanced color-coding technique. We presented a novel way to calculate success probability for a single coloring iteration. We explained how to calculate the number of coloring possibilities for a path with a given k_{max} configuration. We also discussed the relation between configurations with different k_{max} values. We used the enhanced color-coding technique to find signaling pathways in protein interaction networks. We empirically showed that our method produces correct results, and that it needs less time than the leading method to produce these results. We also showed that the results of our method are of statistical and biological significance. Possible future extensions to the present work include extracting deregulated signaling pathways using a cancer gene expression dataset. The subject PPI network could be built from mRNA co-expression and high-throughput experiments.

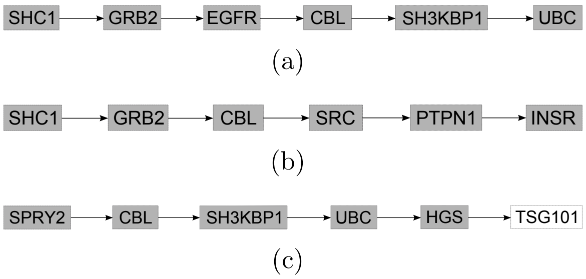


Fig. 6. Three sample pathways with functional enrichment value less than 10^{-11} found by our method in the *H.sapiens* dataset. The shaded nodes correspond to the genes which have common gene ontology term leading to the best functional enrichment. (a) The common term is GO:0042058. (b) The common term is GO:0046875.(c) The common term is GO:0042059.

References

1. B. Schwikowski, P. Uetz and S. Fields, *Nature Biotechnology* **18**, 1257 (December 2000).
2. P. Uetz, L. Giot and G. Cagney *et al.*, *Nature* **403**, 623 (February 2000).
3. B. P. Kelley, R. Sharan and R. M. Karp *et al.*, *PNAS* **100**, 11394 (September 2003).
4. J. Scott, T. Ideker, R. M. Karp and R. Sharan, Efficient algorithms for detecting signaling pathways in protein interaction networks, in *RECOMB*, (Springer-Verlag, Berlin, Heidelberg, 2005).
5. N. Alon, R. Yuster and U. Zwick, *J. ACM* , 844 (1995).
6. G. Gülsoy, B. Gandhi and T. Kahveci, Topology aware coloring of gene regulatory networks, in *ACM BCB*, (ACM, New York, NY, USA, 2011).
7. T. Shlomi, D. Segal, E. Ruppin and R. Sharan, *BMC Bioinformatics* **7**, p. 199 (2006).
8. S. Lu, F. Zhang, J. Chen and S.-H. Sze, *Algorithmica* **48**, 363 (August 2007).
9. M. Steffen, A. Petti and J. Aach *et al.*, *BMC Bioinformatics* **3**, p. 34 (2002).
10. G. Bebek and J. Yang, *BMC Bioinformatics* **8**, p. 335 (2007).
11. A. Gitter, J. Klein-Seetharaman, A. Gupta and Z. Bar-Joseph, *NAR* **39**, p. e22 (2011).
12. B. Dost, T. Shlomi and N. G. *et al.*, *Journal of Computational Biology* **15**, 913 (2008).
13. F. Dong, K. Koh and K. Teo, *Chromatic Polynomials And Chromaticity of Graphs* (World Scientific Pub., 2005).
14. A. Chatr-aryamontri, A. Ceol and L. M.-P. *et al.*, *Nucleic Acids Research* **35**, 572 (2007).
15. A. Ceol, A. Chatr Aryamontri and L. Licata *et al.*, *Nucleic Acids Research* **38**, D532 (2010).
16. M. Ashburner, C. A. Ball and J. A. Blake *et al.*, *Nature genetics* **25**, 25 (May 2000).

IDENTIFYING MASTER REGULATORS OF CANCER AND THEIR DOWNSTREAM TARGETS BY INTEGRATING GENOMIC AND EPIGENOMIC FEATURES

OLIVIER GEVAERT
Radiology, Stanford, 1201 Welch Road
Stanford, CA 94305
Email: olivier.gevaert@stanford.edu

SYLVIA PLEVRITIS
Radiology, Stanford, 1201 Welch Road
Stanford, CA 94305
Email: Sylvia.plevritis@ stanford.edu

Vast amounts of molecular data characterizing the genome, epigenome and transcriptome are becoming available for a variety of cancers. The current challenge is to integrate these diverse layers of molecular biology information to create a more comprehensive view of key biological processes underlying cancer. We developed a biocomputational algorithm that integrates copy number, DNA methylation, and gene expression data to study master regulators of cancer and identify their targets. Our algorithm starts by generating a list of candidate driver genes based on the rationale that genes that are driven by multiple genomic events in a subset of samples are unlikely to be randomly deregulated. We then select the master regulators from the candidate driver and identify their targets by inferring the underlying regulatory network of gene expression. We applied our biocomputational algorithm to identify master regulators and their targets in glioblastoma multiforme (GBM) and serous ovarian cancer. Our results suggest that the expression of candidate drivers is more likely to be influenced by copy number variations than DNA methylation. Next, we selected the master regulators and identified their downstream targets using module networks analysis. As a proof-of-concept, we show that the GBM and ovarian cancer module networks recapitulate known processes in these cancers. In addition, we identify master regulators that have not been previously reported and suggest their likely role. In summary, focusing on genes whose expression can be explained by their genomic and epigenomic aberrations is a promising strategy to identify master regulators of cancer.

1. Introduction

Technologies exist to rapidly and affordably profile the genome, epigenome and transcriptome of cancer. For example, advances in high throughput analysis allow quantification of global DNA variation, DNA methylation or RNA expression of biological samples (*1-4*). The current challenge is to integrate these layers of complex molecular biology information to produce a more comprehensive view of cancer (*5-9*). Successfully dealing with this complexity will allow determining how much each of the different types of genomic variations, i.e. mutation, copy number alteration or DNA methylation affect gene expression of key cancer drivers. Answering this question should provide a deeper understanding of cancer and insights on its initiation, progression and treatment response. Previous integration efforts have focused on how to distinguish driver genes from passenger genes. For example, Ciriello et al. developed a method to identify driver genes in glioblastoma based on mutual exclusivity by modeling copy number and mutation data (*10*). Vandin et al. developed a method to identify driver genes in cancer by focusing on pathways with a significant enrichment of approximately mutually exclusive genes (*11*). Several other investigators have identified driver genes through network analysis, such as Akavia et al. who used copy number data to filter potential regulators in a Bayesian module network analysis (*12*).

To identify master regulators of cancer and their targets, we built further on the network approach by filtering the candidate driver through a method that integrates copy number, DNA methylation, mutation and gene expression data. Our approach starts by generating a list of candidate driver genes based on the rationale that genes that are driven by multiple genomic events in a significant subset of samples are unlikely to be randomly deregulated. Examples of such genomic events for tumor suppressors are deletions, hyper-methylation or nonsense mutations. In the case of oncogenes, possible genomic or epigenomic events are amplification, hypo-methylation or a fusion with an active promoter region. Instead of using a statistical test on each genomic aberration separately, we developed a linear model that tests for concordance with three different types of genomic alterations simultaneously. We define these genes as candidate drivers because their expression can be significantly explained by the key mechanisms that drive oncogenesis: mutation, copy number alteration or DNA methylation. The second step of our algorithm selects the master regulators from these candidates and identifies their targets. This step applies a modified module networks analysis to computationally dissect the gene expression data into gene modules of co-expressed genes and assigning a regulatory program to each module (*12-14*). Our strategy has the advantage of using an informative way of selecting potential drivers and then focuses on those drivers that are likely to effect downstream targets.

We applied our algorithm to identify master regulators and their targets in glioblastoma multiforme and serous ovarian tumors from The Cancer Genome Atlas (TCGA). We found that the expression of the selected cancer drivers are greatly influenced by their copy number and to a much lesser extent by DNA methylation. In addition, for some drivers, we show synergy between genomic and epigenomic events. The second step of our algorithm selects a small set of potential cancer drivers as master regulators that explain much of the global gene expression in the reconstructed module network. Our results show that using candidate drivers from the first step improves the predictive performance on an independent test set of our models developed in the second step. This indicates that focusing on genes that are explained by their genomic and epigenomic profiles is a promising strategy to select master regulators of cancer.

2. Methods

2.1. *Algorithm*

We developed a biocomputational approach to identify key genes that drive human cancer. Our approach involves generating a list of candidate drivers (Step 1), followed by selecting the master regulators from the candidate drivers and their downstream targets (Step 2).

2.1.1. *Step 1: Identifying candidate drivers of cancer*

For a given gene to be considered a candidate cancer driver, we require that its gene expression be explained by its own genomic alterations, measured by its copy number, CpG DNA methylation and/or mutational variation. Our rationale is that cancer drivers whose expression can be explained by multiple genomic events are unlikely to be randomly deregulated. We used generalized linear models to predict the expression of each gene in terms of its own copy number, DNA methylation and mutation status. Our algorithm is initiated with a quality filter that removes copy number probes that are negatively correlated with gene expression and DNA methylation probes that are positively correlated with gene expression data. We reasoned that these probes have a higher chance of being associated with technical problems than a true underlying biological event. Next, we built a linear model to capture the effect of copy number, DNA methylation and mutation status on the expression level of a gene:

$$Exp_i = f(\beta_1 CGH_i + \beta_2 Methylation_i + \beta_3 Mutation_i) \qquad (1)$$

with ß$_i$ the coefficients of the three predictors (i.e. CGH, DNA methylation or mutation status). We used sequential feature selection when adding multiple predictors by including a predictor only when it increases the R-square statistic more than expected by chance, based on the chi-square distribution with one degree of freedom. This model building procedure was wrapped inside a 10-fold cross validation loop (10F-CV) to estimate the generalization performance of the model on unseen data. We required that a predictor – e.g. CGH status – was selected in all cross validation iterations. The performance of the model was estimated using the R-square statistic on unseen data in each cross validation loop. We used several thresholds on the R-square statistic ranging from 0.2 to 0.5 and evaluated the number of genes at each threshold. We focused on genes with high R-square values since for these genes the expression is significantly explained by their copy number, methylation or mutation status. Within this set of genes we identified genes that are identified as a transcription factor. We used several external sources of information to define a gene as a transcription factor such as HPRD, a census of human transcription factors (*15*) and Gene Ontology resulting in a final list of 3964 transcription factors. This results in a list of candidate drivers that will serve as input for step 2.

2.1.2. *Step 2: Identifying master regulators and their targets*

The second step of our algorithm involves identifying the master regulators (as a key subset of cancer drivers from step 1) and determining their downstream targets by reconstructing a regulatory module network. Our module network approach builds upon previous work (*13, 14*). The algorithm is initiated by clustering the gene expression data into gene modules of co-expressed genes and then assigns a regulatory program to each module. The regulatory program of each module is defined by a sparse linear combination of driver genes that predict the module's mean expression and are chosen from the list of transcription factors among the candidate drivers. The sparseness of the regulatory program is induced using elastic net regularization. We extended the module network framework in three ways: (a) first, we developed an approach to deal with auto-regulation, which is a situation where a regulator is selected in the regulatory program and is also a member of the same module. We allow this event to occur but relearn the regulatory program after removing the gene from the cluster. The regulator only stays in the regulatory program when it is also selected after removal of its expression in the module. (b) Second, we add a 10F-CV strategy that determines the regularization parameter for each module through minimization of the error. (c) Third, we use an iterative algorithm when adding regulators to the regulatory program by using the LARS-EN algorithm which has the advantage that it updates the elastic net solution sequentially (*16*) and thereby allows to stop adding regulators early.

After initial clustering of the data, the module network algorithm is run iteratively by learning the regulatory program and re-assigning genes to modules based on the updated regulatory program. Genes are reassigned to the module that they are closest to, based on Pearson correlation. We used k-means clustering with 100 clusters as the initial clustering algorithm. Next, our algorithm is run untill convergence corresponding to less than 1% of the genes being assigned to new modules. The module network is then interpreted using enrichment analysis using a hyper-geometric test to check for enrichment of gene sets in the gene modules to identify the key biological processes that are driven by the regulators. We used several databases of gene sets from MSigDB (*17*), GeneSetDB (*18*) and manually curated gene sets.

2.2. *Data*

We used data from The Cancer Genome Atlas (TCGA) on glioblastoma and ovarian cancer (data downloaded in May 2011). Gene symbols were used to map different technologies. Normal samples were removed. We used Level 3 Agilent G4502A gene expression data and Level 2 27K CpG methylation data. CpG sites were mapped to its closest gene transcription start site. The methylation probe level data was used since bi-modal signals were found for genes where multiple probes were present. Because averaging all probes for such a gene removed signal from the data, we defined methylation clusters based on a minimum Pearson correlation of 0.4 within a cluster. For the CGH data, two different platforms were used for the glioblastoma and ovarian project. For the glioblastoma project the CGH data was produced by the Agilent 244A platform and for the ovarian project the Agilent 1x1M platform was used. In both cases, we used the Level 3 CGH data. In the glioblastoma dataset, 251 patients had gene expression, DNA methylation and CGH data; these datasets were available for 511 ovarian cancer patients. For a limited number of patients, duplicate data was available however no averaging was done for these cases. We arbitrarily picked one case. When missing values were present, we estimated the missing value using 15-KNN (*19*). In most data sets a significant batch effect was observed and batch correction was done for all data sources using Combat (*20*). Mutation data was present for 140 glioblastoma patients and 324 ovarian patients through exome sequencing and we extracted all novel non-silent mutations. Gene expression data was present for 426 glioblastoma and 560 ovarian patients and used to generate the modules of the regulatory network. For all genes that had gene expression data, 14041 had also copy number, 9987 had DNA methylation and 8619 had both for ovarian cancer. For glioblastoma the overlap with gene expression data resulted in 13113 genes with copy number data, 9107 genes with DNA methylation data and 7510 with both.

3. Results

We used a two-step algorithm to identify the master regulators of cancer and their targets. We applied this algorithm on multi-dimensional TCGA ovarian cancer and glioblastoma datasets.

3.1. *Identifying candidate drivers for glioblastoma and ovarian cancer*

To identify candidate drivers of cancer, we developed a linear model to estimate the effect of copy number alterations, DNA methylation and mutation on gene expression levels (Step 1, Methods). Figure 1 shows the number of genes that is significantly explained by copy number, methylation or both at different R-square thresholds. More than 5000 genes have an R-square value for copy number alone of at least 0.20 in ovarian cancer compared to 1137 genes for glioblastoma reflecting the massive amount of copy number alterations that is present in serous ovarian cancer (*21*). Interestingly, DNA methylation is less informative when explaining gene expression data and much less genes are significant at each R-square threshold for both ovarian cancer and glioblastoma. For both glioblastoma and ovarian cancer, we found adding mutation data did not significantly change the results.

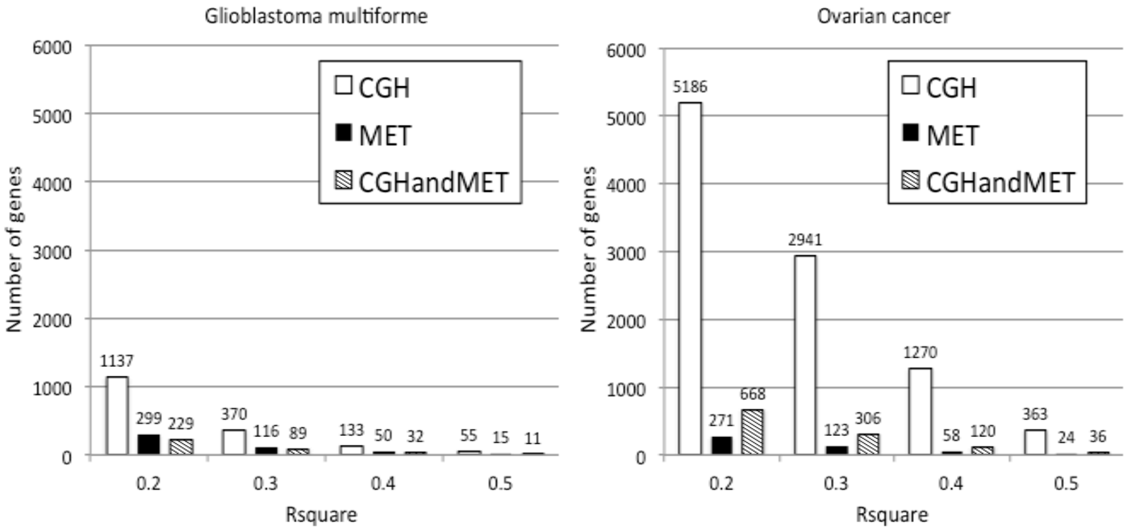


Figure 1 Number of genes whose expression is significantly explained by its own copy number, DNA methylation or both.

3.1.1. *Glioblastoma candidate drivers*

When focusing on the genes with high R-square values in glioblastoma, we verified that our algorithm discovered a number of interesting genes previously reported and validated on TCGA glioblastoma (*22*). For example the gene PDGFRA, part of the platelet-derived growth factor receptor, has an R-square of 0.38 when considering only its copy number, 0.29 when considering only its DNA methylation profile and 0.60 when considering both. This indicates that 60% of the expression of PDGFRA is explained by synergy between its copy number and DNA methylation. PDGFRA is a receptor tyrosine kinase and an important part of the RAS pathway. In addition, PDGFRA is also mutated in 3 out of 140 patients (2%). Other interesting examples for glioblastoma include the genes MGMT and GLI1. MGMT, well known for its association with glioblastoma sensitivity to alkylating agents (*23*), has an R-square value of 0.46 and is significantly explained by its DNA methylation and copy number profile. Similarly, GLI1, glioma associated oncogene homolog 1, has an R-square value of 0.46 and is mutated in 1 patient.

3.1.2. *Ovarian cancer candidate drivers*

The ovarian cancer candidate drivers also included interesting genes with high R-square values. The gene BRCA1 is known to be associated with ovarian cancer due to mutations (*21*). In our analysis using copy number and DNA methylation data, we found that BRCA1 has an R-square of 0.10 when considering only its copy number, 0.41 when considering only its methylation profile and 0.49 when considering both. This indicates that besides mutation, DNA methylation is an important mechanism driving BRCA1 expression. This finding was also shown in the original TCGA ovarian results demonstrating that our method is able to recapitulate previous results (*21*). BRCA2 gene expression on the other hand is only explained by its copy number and is not significantly epigenetically regulated. Other interesting examples for ovarian cancer are KRAS, mutated in 2 out of 324 cases, with an R-square of 0.60 solely based on its copy number, and RAB25 with an R-square of 0.82 solely based on its DNA methylation. RAB25 was shown to be highest ranked gene epigenetically silenced in the original TCGA ovarian results and this is also the case using our model (*21*).

3.1.3. *Gene set enrichment*

We used several databases with gene sets to investigate the enrichment of known pathways and biological processes in the set of driver genes. We looked at gene set enrichment of the gene lists at an R-square of 0.3. The glioblastoma driver genes explained only by copy number were enriched in genes identified in the TCGA glioblastoma results as part of significant copy number changes (*21*). In addition, gene sets related to copy number changes in many other cancers were also in the top enriched gene sets (*24-27*). Interestingly the genes explained significantly by their DNA methylation at this R-square threshold were enriched in extracellular matrix genes and genes related to cell migration. For the ovarian cancer genes only explained by their copy number we observed enrichment of proliferation pathways, genes related to a BRCA1/CHEK1 network (*28*) and an ovarian cancer survival signature (*29*). Next, the top genes explained only by DNA methylation are enriched in genes affected by methylation in other cancers (*30-32*) validating our approach.

3.2. *Identifying candidate master regulators of gene expression*

To identify the master regulators of the network and their downstream targets, we apply a module network approach (Step 2, Methods). Our module network analysis is based on linear regression with elastic net regularization using the key transcription factors from the candidate drivers identified in Step 1 (*13*). For both glioblastoma and ovarian cancer we built a module network to associate transcriptional driver genes with their downstream targets. We used the gene expression data of 426 glioblastoma and 560 ovarian cancer patients but used only the top half most varying genes in all further analysis. As potential regulators for the modules, we selected genes with a high R-square as significantly regulated by a combination of copy number and DNA methylation and intersected this list with known transcription factors. This resulted in 431 and 469 genes for glioblastoma and ovarian cancer respectively that are defined as a transcription factor, show high variance and are regulated significantly by genomic alterations.

Table 1 Master regulators for the ovarian and GBM network. Genes highlighted in the main text are in bold.

Glioblastoma network		Ovarian network	
Regulators	Nr Modules	Regulators	Nr Modules
ZNF300	10	**BATF**	13
TNFRSF1A	10	HTATIP2	9
PTRF	8	PML	9
WWTR1	8	NOD2	8
MYT1	7	JAK2	8
PYCARD	7	HMGA2	7
PATZ1	7	**TGFB3**	7
BASP1	6	KLF12	7
RAB32	6	AKAP8L	7
SATB1	6	YWHAH	6
ZMYND12	6	HLA-DQB1	5
CDC45	6	JARID2	5
ZNF217	6	RNF19A	5
KCNIP3	6	MORF4L1	5
ARNT2	6	SMAD4	5
BTF3L4	6	ZNF500	5
POGZ	6	**NFKB1**	5
TOB1	6	TRIM29	4
LGALS3	5	SPDEF	4
KCNH8	5	SREBF1	4

3.2.1. *Glioblastoma module network*

The master regulators in the glioblastoma network are TNFRSF1A, an important partner in the TNF and NF-kB pathway and ZNF300, both predicted as a regulator of 10 modules. We focused on the top DNA repair module in our network because this process is an important pathway in both glioblastoma and ovarian cancer. In the glioblastoma module network, module 22 is the top DNA repair module and is regulated by 6 regulators including DNMT1 and PARP1 and contains 81 genes. DNMT1 is a key player in regulation DNA methylation regulation and has been shown to be involved in inactivation of tumor suppressor genes and failure to maintain genomic stability (*33*). In addition PARP1 is known to regulate DNMT1 and forms a complex with DNMT1 (*34, 35*). Both are only explained by their copy number profile and are not driven by their own DNA methylation status.

3.2.2. *Ovarian cancer module network*

The master regulator for the ovarian cancer module network is BATF a transcription factor with unknown function. BATF is part of the regulatory program of 13 modules and its expression is significantly explained by its DNA methylation status. Other important regulators are NFKB1 and TGFB3. Similarly to the glioblastoma module network, we also focused on the most highly enriched DNA repair module for ovarian cancer: module 89. Module 89 contains 86 genes and has 10 regulators including EZH2, AURKA and CHAF1B. Interestingly CHAF1B was also predicted as a regulator of the top DNA repair module in glioblastoma. Other interesting regulators are CCNE1 and RAB25. CCNE was identified as a low frequency amplification in the original TCGA results and is predicted as the only regulator of a module enriched in the focal adhesion pathway. Next, RAB25 is the top methylated gene (*21*) and in our module network is part of the regulatory program together with MAML2, a member of the NOTCH pathway, a pathway also identified in the original ovarian TCGA results.

3.3. ***Algorithm Performance***

To evaluate the performance of our algorithm, we investigated how well our candidate drivers perform compared to random sets of transcription factors. We used an independent test set for both glioblastoma multiforme and ovarian cancer (*9, 36*) to estimate the generalized performance of each module on unseen data. To limit the computational power required and to facilitate comparison of the results, we ran the second step of our algorithm only once. This essentially corresponded to learning a regulatory program for the initial clustering. First, we established a baseline performance by incrementally and randomly adding transcription factors to the list of potential regulators. This was repeated 5 times for each number of potential transcription factors. Figure 2 shows how the performance evolves when adding more transcription factors. The performance is measured by averaging the R-square over all modules on the test set. Figure 2 shows that for glioblastoma and ovarian cancer, the performance plateaus after adding more than 600 potential regulators indicating that increasing the number of potential regulators beyond this point does not improve the predictive performance of the model on unseen data.

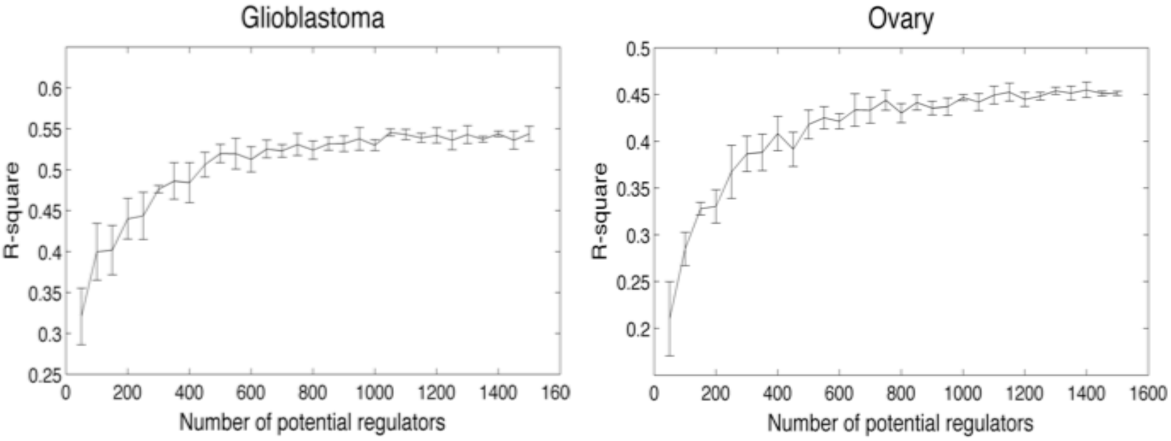


Figure 2 Generalized performance of module networks generated from randomly selected candidate drivers.

Finally, we investigated the performance of using transcription factors that are also driver genes. We reasoned that focusing only on candidate drivers as regulators would increase the performance of only a subset of modules and therefore focused on modules with a minimum R-square value on unseen data and compared the performance of these modules with random sets of transcription factors. For glioblastoma we saw an increase in performance independent of R-square threshold while for ovarian cancer the performance improved starting at a minimum R-square of 0.10. Figure 3 shows the average performance of modules with an R-square value of at least 0.20 on unseen data when adding incrementally regulators ranked by their own R-square value. Our results show for both glioblastoma and ovarian cancer that the generalization performance is comparable or better than random sets of transcription factors at several sizes of potential regulators (Figure 3).

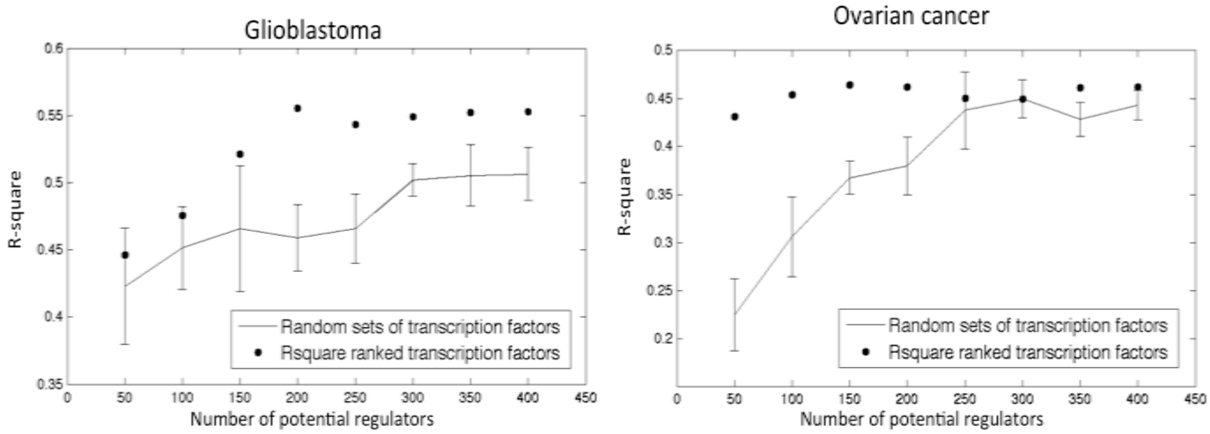


Figure 3 R-square performance of the module networks generated using candidate drivers vs. randomly drawn candidate drivers.

4. Discussion

To identify master regulators of gene regulation in cancer, we developed a biocomputational approach that first creates a select list of candidate cancer driver genes by integrating multiple genomic datasets. From this list, we select the master regulators and identify their targets when reconstructing a regulatory module network of cancer. For a gene to be on the list, we require that its expression be explained by known genomic and/or epigenomic aberrations, measured in terms of copy number variation, DNA methylation and mutational events. This requirement reduces the list of candidate drivers and improves the performance of the regulatory module network when applied to glioblastoma and ovarian cancer TCGA data.

For each candidate driver, we can determine which genomic aberration explains more of the gene expression. In the case of both GBM and ovarian cancer, the candidate drivers appear to be more influenced by their copy number variations than DNA methylation. DNA methylation appears to have a more subtle effect on gene expression. In addition, we identified many genes that showed synergy between their copy number and DNA methylation showing that a cancer cell can deregulate gene expression using both mechanisms. Besides copy number and DNA methylation, we also investigated the addition of mutation data to our linear model and investigated if mutation data has a significant effect on the amount of variance in gene expression that can be explained. As expected adding mutation data did not significantly change the results due to sparseness of mutation data. More importantly, only a subset of mutations will have an effect on gene expression because many missense mutations will not effect gene expression but may disrupt protein function. Determining this computationally requires dedicated methods that specifically model the mutation data and their impact on the final protein product to estimate which mutations have or do not have an effect on gene expression. For example, we observed that TP53 was not correlated with gene expression even though it is known to be an important tumor suppressor in ovarian cancer.

By focusing on candidate drivers as genes that are explained by their genomic and epigenomic profiles, we can identify more likely master regulators in the module networks analysis. We found that using transcription factors whose expression is determined by copy number or DNA methylation profile, had favorable performance on unseen data. In the context of the module networks generated from random sets of transcription factors, which were shown to plateau after adding more than 600 potential regulators, while our method provides a way of intelligently selecting regulators in module networks.

By the virtue of applying module network analysis, the master regulators are associated with downstream targets. The master regulators in both the glioblastoma and ovarian cancer network belong to known pathways affected in these cancers. In addition, we found several unknown genes that are important regulators in our module networks. For example, CHAF1B is predicted as a regulator for the top DNA repair module in both glioblastoma and ovarian cancer. CHAF1B is predicted to have a function in DNA repair and is part of a 4-gene signature predicting survival in glioma (*37*). Moreover, CHAF1B has been shown to be correlated to proliferation in several epithelial cancers (*38*). Interestingly we found that CHAF1B expression is dominated by its copy number in ovarian cancer and by DNA methylation in glioblastoma, showing the flexibility of our method. As more data comes available in the TCGA, such inter-cancer comparisons can be made with the potential to identify master regulators independent of cancer subtypes.

In summary, we developed a biocomputational approach for integrating multi-dimensional cancer data that allows to study how genomic and epigenomic features influence gene expression. Next, we used our method to identify master regulators of cancer and their downstream targets. Our approach has the potential to provide new insights in the molecular biology underlying cancer. Moreover, it associates drivers with their downstream targets, thereby enabling new insight into the biological mechanism underlying cancer progression.

5. Acknowledgements

This research was supported by the Center for Cancer Systems Biology (CCSB) at Stanford (U54 CA149145). OG is a fellow of the Fund for Scientific Research Flanders (FWO-Vlaanderen), an Honorary Fulbright Scholar of the Commission for Educational Exchange between the United States of America, Belgium and Luxembourg, and a Henri Benedictus Fellow of the King Bauduin Foundation and the Belgian American Educational Foundation (BAEF).

References

1. W. Pao *et al.*, *Clin Cancer Res* **15**, 5317 (2009).
2. C. Sotiriou, *Annals of Oncology* **20**, 10 (2009).
3. O. Gevaert, B. De Moor, *Expert Opinion on Medical Diagnostics* **3**, 157 (2009).
4. O. Gevaert, A. Daemen, B. De Moor, L. Libbrecht, *BMC Med Genomics* **2**, 69 (2009).
5. L. Chin, W. C. Hahn, G. Getz, M. Meyerson, *Genes & development* **25**, 534 (Mar 15, 2011).
6. O. Gevaert, S. Van Vooren, B. De Moor, *Annals of the New York Academy of Sciences* **1115**, 240 (2007).
7. A. Daemen, M. Signoretto, O. Gevaert, J. A. Suykens, B. De Moor, *PLoS ONE* **5**, e10225 (2010).
8. K. Leunen *et al.*, *Human mutation* **30**, 1693 (2009).
9. L. Gravendeel *et al.*, *Cancer research* **69**, 9065 (2009).
10. G. Ciriello, E. Cerami, C. Sander, N. Schultz, *Genome Res*, (Oct 12, 2011).
11. F. Vandin, E. Upfal, B. J. Raphael, *Genome Res*, (Jul 11, 2011).
12. U. D. Akavia *et al.*, *Cell* **143**, 1005 (Dec 10, 2010).
13. S.-I. Lee *et al.*, *PLoS genetics* **5**, e1000358 (2009).
14. E. Segal *et al.*, *Nature Genetics* **34**, 166 (2003).
15. J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, N. M. Luscombe, *Nat Rev Genet* **10**, 252 (Apr, 2009).
16. H. Zou, T. Hastie, *Journal of the Royal Statistical Society Series B-Statistical Methodology* **67**, 301 (2005).
17. A. Subramanian *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545 (2005).

18. A. C. Culhane *et al.*, *Nucleic Acids Res* **38**, D716 (Jan, 2010).
19. O. Troyanskaya *et al.*, *Bioinformatics* **17**, 520 (2001).
20. W. E. Johnson, C. Li, A. Rabinovic, *Biostatistics* **8**, 118 (Jan, 2007).
21. D. Bell *et al.*, *Nature* **474**, 609 (Jun 30, 2011).
22. R. McLendon *et al.*, *Nature* **455**, 1061 (2008).
23. M. Esteller *et al.*, *N Engl J Med* **343**, 1350 (Nov 9, 2000).
24. W. M. Lin *et al.*, *Cancer Res* **68**, 664 (Feb 1, 2008).
25. W. W. Lockwood *et al.*, *Oncogene* **27**, 4615 (Jul 31, 2008).
26. J. Greshock *et al.*, *Cancer Res* **67**, 3594 (Apr 15, 2007).
27. I. Osman *et al.*, *Clin Cancer Res* **12**, 3374 (Jun 1, 2006).
28. M. Pujana *et al.*, *Nature Genetics* **39**, 1338 (2007).
29. T. Bonome *et al.*, *Cancer Res* **68**, 5478 (Jul 1, 2008).
30. G. Heller *et al.*, *Cancer Res* **68**, 44 (Jan 1, 2008).
31. L. G. Acevedo, M. Bieda, R. Green, P. J. Farnham, *Cancer Res* **68**, 2641 (Apr 15, 2008).
32. N. Sato *et al.*, *Cancer Res* **63**, 3735 (Jul 1, 2003).
33. G. Rajendran *et al.*, *J Neurooncol* **104**, 483 (Sep, 2011).
34. M. Zampieri *et al.*, *Biochem J* **441**, 645 (Jan 15, 2012).
35. M. Zampieri *et al.*, *PLoS ONE* **4**, e4717 (2009).
36. R. W. Tothill *et al.*, *Clin Cancer Res* **14**, 5198 (Aug 15, 2008).
37. M. de Tayrac *et al.*, *Clin Cancer Res* **17**, 317 (Jan 15, 2011).
38. S. E. Polo *et al.*, *Histopathology* **57**, 716 (Nov, 2010).

MODULE COVER – A NEW APPROACH TO GENOTYPE-PHENOTYPE STUDIES[1]

YOO-AH KIM, RAHELEH SALARI[†], STEFAN WUCHTY, AND TERESA M. PRZYTYCKA

National Center for Biotechnology Information, NLM, NIH,
Bethesda, MD 2089, USA
Email: {kimy3, wuchtys,przytyck} ncbi.nlm.nih.gov ; rahelehs@cs.stanford.edu;

Uncovering and interpreting phenotype/genotype relationships are among the most challenging open questions in disease studies. Set cover approaches are explicitly designed to provide a representative set for diverse disease cases and thus are valuable in studies of heterogeneous datasets. At the same time pathway-centric methods have emerged as key approaches that significantly empower studies of genotype-phenotype relationships. Combining the utility of set cover techniques with the power of network-centric approaches, we designed a novel approach that extends the concept of set cover to network modules cover. We developed two alternative methods to solve the module cover problem: (i) an integrated method that simultaneously determines network modules and optimizes the coverage of disease cases. (ii) a two-step method where we first determined a candidate set of network modules and subsequently selected modules that provided the best coverage of the disease cases. The integrated method showed superior performance in the context of our application. We demonstrated the utility of the module cover approach for the identification of groups of related genes whose activity is perturbed in a coherent way by specific genomic alterations, allowing the interpretation of the heterogeneity of cancer cases.

1. Introduction

Complex diseases, such as cancer, are typically caused by a combination of genomic alterations, epigenetic and environmental factors, and different combinations of such factors may result in the same disease phenotype. In addition, signals that are associated with each individual genetic perturbation might be weak and difficult to separate from background noise. Collectively, these obstacles render the identification of subtle genotype-phenotype relationships extremely challenging.

Recently, pathway-centric methods have emerged as key approaches that empower studies on genotype-phenotype relationships. Such pathway-centric studies typically leverage large interaction networks inferred by high-throughput experiments. Projecting gene expression data on an interaction network, these approaches infer molecular activities on the level of biological pathways (subnetworks) rather than individual genes (*1*-*5*). Gene expression has been utilized to assess the activity of subnetworks (*6*), while genotypic data has lately been used to identify mutated subnetworks by exploring positions of mutated genes in interaction networks (*7*-*9*). An additional level of understanding of genotype-phenotype relationships can be obtained when both genotype and gene expression data are available. A recent study (*10, 11*) combined copy number alteration and gene expression data and applied a current flow approach to identify flow of information from potential genomic causes to differentially expressed disease genes.

[*] This work was supported in part by the Intramural Research Program of the Health, National Library of Medicine.

[†] Current address Computer Science Department, Stanford University Stanford CA 94305-5428

Generally, pathway-centric approaches are based on the premise that different genetic perturbations often dys-regulate the same pathway, leading to the same disease phenotype. Therefore, the identification of such dys-regulated pathways is important for the understanding of a disease, potentially guiding drug development efforts. However, complex diseases are usually vaguely defined, and typically what can be seen as a spectrum of diseases is annotated as one disease. In such a heterogeneous set, individual disease cases may be characterized by various combinations of dys-regulated pathways.

Set cover approaches have been proven useful in the determination of disease markers in heterogeneous datasets (1, 2, 5, 11). In a set cover, a gene is considered to cover a disease sample if the gene is dys-regulated in the sample. The underlying assumption of the set cover approach is that each disease case has some dys-regulated (thus covering) genes but if the disease is heterogeneous, different cases will typically have different covering genes. In particular, a multi set cover approach aims to find a set of genes so that each disease case is represented (covered) by at least a certain number of differentially expressed genes while the total number of selected genes is minimized (11). However, current set cover approaches do not consider several important issues: (i) if two different disease cases are covered by two different sets of genes this does not necessarily means that they are caused by a dys-regulation of different pathways (ii) signals of associations from an individual gene to genetic alterations may be weak and noisy.

Combining the strength of the set cover approach with the power and stability of network-centric methods, we designed a new technique that extends the concept of set cover from single genes to network modules. In contrast to previous “connected network cover” approaches which strived to identify one connected subnetwork covering most disease cases (*1, 2, 5*), our approach allows us to identify multiple subnetworks (modules), so that each disease case is covered by a number of modules while the total “cost” of modules is minimized. In addition to network information, the definition of a module involves a similarity measure between pairs of genes that is based on eQTL association profiles. While modules can be comprised of singleton genes, the trade-off between module granularity and similarity of genes in the module is controlled by a cost function.

Given the above definition of similarity, the module cover approach can be used to find covering subnetworks such that genes in each module are jointly regulated by the same genetic alterations. The problem of detecting subnetworks that are influenced by common genetic alterations has been recently approached with a variant of the LASSO method (*12*) and Bayesian partition methods (*13*) with different objectives in mind. In particular, none of the approaches was designed to deal with data heterogeneity while our set cover modules capture the heterogeneity of samples where each module covers a different subset of samples. In addition, the LASSO based method, GFlasso, in its current implementation does not scale to large datasets while the Bayesian approach does not utilize network information.

To solve the module cover problem, we developed an integrated method that simultaneously determines network modules and optimizes the cover of disease cases. For comparison, we also implemented a two-step method where we first determined candidate network modules and subsequently selected a subset of modules that cover disease cases. While the performance of the integrated method is superior to the two-step method, the two-step approach still performed better than a naïve method that was based on a single gene cover.

We applied the module cover approach to discover modules associated with genomic alterations in cancer patients, utilizing genomic alteration and gene expression data. Representing each gene by its eQTL (expression Quantitative Trait Loci) association profile our algorithms harness profile similarities between genes and identify modules of genes with highly correlated eQTL profiles that collectively cover all disease cases.

We start by introducing a mathematical formalization of the module cover problem and subsequently describe our two algorithms: ***Integrated Module Cover*** and ***Two-Step Module Cover***. Next, we introduce rigorous measures to compare the quality of the modules obtained by the two algorithms. Finally, we analyze the modules obtained by the Integrated Module Cover that was applied to Glioblastoma Multiforme (GBM) and ovarian cancer data. We conclude with a discussion of a broader spectrum of additional applications of the proposed approach.

2. Methods

2.1. *Introduction of the Module Cover Problem*

Here, we extended the concept of the minimum multi-set cover problem to a minimum multi-module cover problem. The classical minimum multi-set cover is formally defined as follows: Given a set of elements $E = \{e_1, e_2, \ldots, e_n\}$, a family of subsets $S = \{E_1, E_{2,\ldots,} E_m | E_i \subseteq E\}$ and a positive integer k, the goal is to select a subfamily of S so that each e_i is included at least k times. In our problem formulation, disease cases are the elements, and a subset of disease cases E_i corresponds to a gene where it is differentially expressed in those disease cases. More specifically, a gene g covers a disease case c ($cover(c, g)=1$) if the gene is differentially expressed in the given case, and $cover(c, g) = 0$ otherwise. To obtain the most prominent disease genes, we aim to select the smallest set of genes to cover all disease cases at least k times (*11*). Fig. 1A shows an example of a multi-set cover where disease cases are elements to be covered by selected genes. An edge between a gene and a case exists if the gene covers the case.

In the module cover approach, we select modules (instead of single genes) to cover disease cases (Fig. 1B). To ensure that genes in a selected module are coherent, the ‘cost’ of modules was defined so that we preferentially assigns low cost to modules with genes that are close to each other in the network and are coherent according to a given similarity measure, such as correlation of expression or eQTL association profiles. In eQTL analysis, gene expression is considered as a quantitative phenotype and controlled by genotypic information. Utilizing matching gene expression and copy number variation, we determined eQTL profiles of each gene by computing significance levels of associations of each gene to genomic alterations (See Section 5.3 for the details).

Let sim(g1, g2) be the eQTL similarity of the two genes, which is computed based on the correlation of their eQTL profiles. We assume that $0 \leq sim(g_1, g_2) \leq 1$. Let $distance(g_1, g_2)$ be the shortest distance between the two genes in the interaction network. We first adjust the similarity by the distance as

$$adjusted\_sim(g_1, g_2) = sim(g_1, g_2)^{1+(distance(g1,\, g2)-1)/(avg\_dist\ -1)} \qquad (1)$$

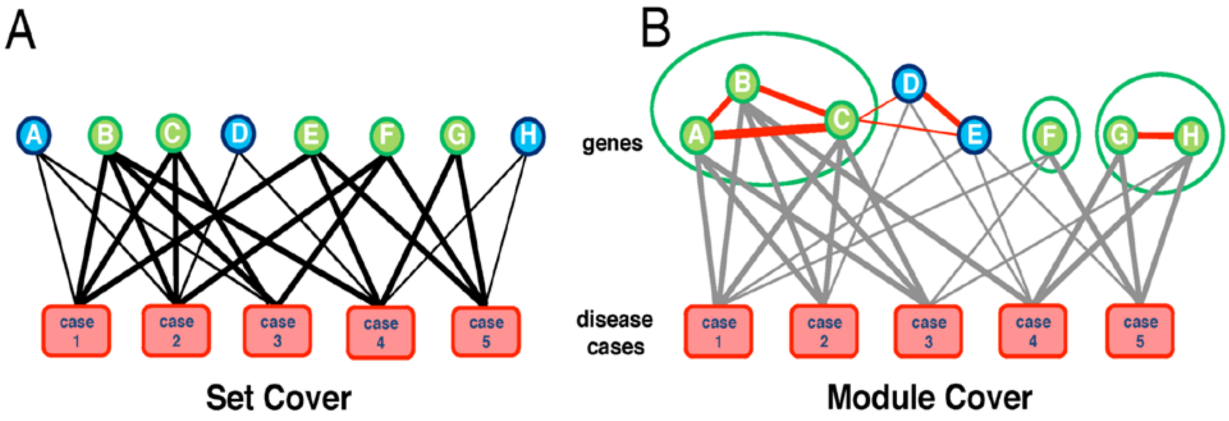


Figure 1. **Set Cover vs. Module Cover. (A)** In a classical set cover, an edge from a gene to a disease case exists if the gene is differentially expressed in the disease case (i.e. covering the case). Genes {B, C, E, F, G} are selected, and all cases are covered at least 3 times. **(B)** A module cover selects coherent modules. Red edges between genes represent the similarity between genes (e.g. based on the correlation coefficient of their eQTL profiles or gene expression patterns). In the example, modules {A, B, C}, {F}, {G, H} are selected, and all cases are covered at least 3 times.

where *avg_dist* is the average distance between all pairs of genes in the network. Since our weight function adjusts the similarity value with interaction information we obtain higher weight if two genes have more similar eQTL profiles and are in close proximity in the network. We define the weight function as follows:

$$w(g_1, g_2) = adjusted\_sim(g_1, g_2) - \theta \qquad (2)$$

where θ is a threshold parameter. The weight is positive (i.e. benefiting module cost) if the adjusted similarity is $>\theta$. Consequently, we define the cost of a module M as

$$Cost(M) = \alpha + |M| - \sum_{x \in M} \sum_{y \in M, y \neq x} w(x, y)/(|M| - 1) \qquad (3)$$

where α is the module initializing cost when a new module is created. We include this initial module cost to minimize the number of selected modules. With a larger α, a smaller number of modules with larger average size will be obtained, since costs increase when a new module is created. The objective of the second term (i.e. the number of genes) is to minimize the total number of selected genes. Finally, we subtract the cost computed as the sum of average weights of genes in the module, ensuring coherence of modules since the cost of a module decreases as the weights (and similarities) between genes increase.

Our goal is to find a minimum cost set of modules that cover all disease cases at least k times where the depth of coverage is a user defined parameter. More specifically, we search for a module set $S' = \{M_1, M_2, \ldots, M_t\}$ that minimizes $\sum_{Mi \in S'} Cost(Mi)$ with the constraint that $\sum_{Mi \in S'} \sum_{g \in Mi} cover(c, g) \geq k$ for each disease case c. The minimum module cover problem is NP-hard as it is a generalization of the minimum set cover, which is known to be NP-hard. In the following two subsections, we describe two different heuristic algorithms: *Integrated Module Cover* and *Two-Step Module Cover*. In the integrated module cover algorithm, we discover modules on the fly while we select genes to cover disease cases. In the two-step module cover algorithm, we first cluster genes based on their similarity to obtain a candidate sets of modules and subsequently select a subset of modules to cover disease cases.

2.2. *Integrated Module Cover*

In this algorithm, we greedily select genes to cover disease cases and simultaneously create modules of 'similar' genes. In each iteration, we consider all unselected genes and compute the cost of adding each of those genes, assuming two ways to add a gene:

1) add the gene as a separate module: the cost of adding the gene is simply α + 1.
2) add the gene to an existing module: To maintain the coherence of a module, we first check if for the candidate gene g the average weight $w(g, v)$ over all other genes in the module is positive. That is, we can add a gene g to a module M only if $\sum_{v \in M} w(g, v) > 0$. The increased cost resulting from adding gene g to module M is $Cost(M+\{g\}) - Cost(M)$.

To find the best extension of the cover we proceed as follows: Let $P(g)$ be the set of existing modules with a positive average edge weight with g as described in the case (2) The cost of adding a gene g is

$$IC(\mathrm{g}) = \min(\alpha+1, \min_{M_i \in P(g)} (Cost(M_i \cup \{g\}) - Cost(M_i)))$$

Since we want to cover disease cases to the largest degree, we also account for the 'benefit' of adding genes. Considering the set of disease cases C' that were covered less than k times by the end of the previous iteration we define the benefit by adding gene g as

$$Benefit\ (g) = \sum_{c \in C'} cover(c, g).$$

In each iteration, we greedily choose a gene with minimum $IC\ (g)/Benefit\ (g)$. If the minimum cost of gene g is obtained adding gene g to an existing module M, the module is updated as $M \cup \{g\}$. Otherwise a new module {g} is created.

2.3. *Two-Step Module Cover*

In the Two-Step heuristic, we first find a candidate set of modules by clustering genes based on their similarity and interaction data. Subsequently, we apply a covering algorithm to select the best set of modules. Specifically, we used Markov Cluster Algorithm (MCL), an unsupervised clustering algorithm based on simulation of stochastic flow in a network (*14*). Note, that a predefined set of modules/pathways may be used instead as well. Given a network of interacting genes, we weight each edge by the corresponding similarity value and obtain a candidate set of modules $\{M_1, M_2, \ldots, M_m\}$ using MCL. We then select modules with coherent/similar genes, covering as many samples as possible. The cost of selecting a module M is given by (3), and we define the benefit of selecting a module as the total coverage

$$Benefit(M_i) = \sum_{g \in Mi} \sum_{c \in C'} cover(c, g)$$

Where, as before, C' is the set of disease cases not covered k times by the end of the previous iteration. In each iteration, we greedily select a module with minimum $Cost\ (M)/\ Benefit\ (M)$.

3. Results

We applied our module cover algorithms to two data sets: the first dataset includes the data for 158 Glioblastoma Multiforme patients (GBM) and 32 non-tumor control samples. The data was collected by the NCI-sponsored Glioma Molecular Diagnostic Initiative (GMDI), which includes matching mRNA expression and copy number variation data for each patient (http://rembrandt.nci.nih.gov/). The second dataset includes 489 Ovarian Cancer data samples from TCGA (The Cancer Genome Atlas). The technical details of data processing are described in the Materials section.

3.1. *Analysis of Glioblastoma Multiforme Data from GMDI*

First, we wanted to estimate which of the two methods provides a better heuristic in the context of our application. Since our goal was to select modules whose members are associated in a coherent way with genotypic changes, we evaluated the two methods based on significance, strength, and coherence of the association.

3.1.1. *Comparison of the Module Cover approaches.*

We applied the integrated greedy module cover algorithm with $k = 300$ and $\alpha = 1$, allowing 5 samples (3%) to be covered less than k times to exclude outliers. We discuss the more detailed parameter selection in online Appendix Section 2. In particular, we found that the number of non-trivial modules (i.e. ≥ 3 genes) starts to level with k = 300, prompting us to choose this parameter value for our main analysis. We obtained 249 modules that contained a total of 513 genes including 41 non-singleton modules. The average distance between genes inside a module was 2.5.

For the two-step module cover, we applied MCL to the network of molecular interactions that have been weighted by correlating eQTL profiles of interacting genes. Using inflation parameter = 4 we obtained 3,401 candidate modules (see Appendix Table A1 and Figure A1 for details of parameter selection). The average size of the candidate modules was 3.21 and 2,677 modules were non-singleton. Subsequently, we greedily selected modules as described in Section 2.3. The two-step cover algorithm selected 801 genes in 454 modules. 233 modules (of which 171 modules are of size 2) were non-singleton. The average distance between genes inside a module was 1.1, indicating that the MCL cover provided more compact modules than the integrated module cover approach.

Testing which of the two approaches provided modules whose members were associated in a more coherent way with genotypic changes, we evaluated modules with respect to significance, strength and coherence of the association.

For each non-singleton module M, we first defined the significance of the association to each of tag loci as the average association significance of the genes in the module. Formally,

$$s_i(M) = \sum_{g \in M} s_i(g)/|M| \qquad (5)$$

where $s_i(g)$ represents $-\log_{10}$ p-value of the association provided by the linearly regressing between expression values of gene g and copy number variation of i-th tag locus (see Section 5.1 for more details).

The upper panel of Fig. 2A shows such association significance profiles of the 10 largest modules. We found strong associations with tag-loci on chromosome 7 and 10. These chromosomes carry signature alterations of GBM, coinciding with the genomic locations of GBM related genes such as EGFR and PTEN. In the lower panel of Fig. 2A, we show association significance profiles of the 10 largest modules selected by the two-step algorithm.

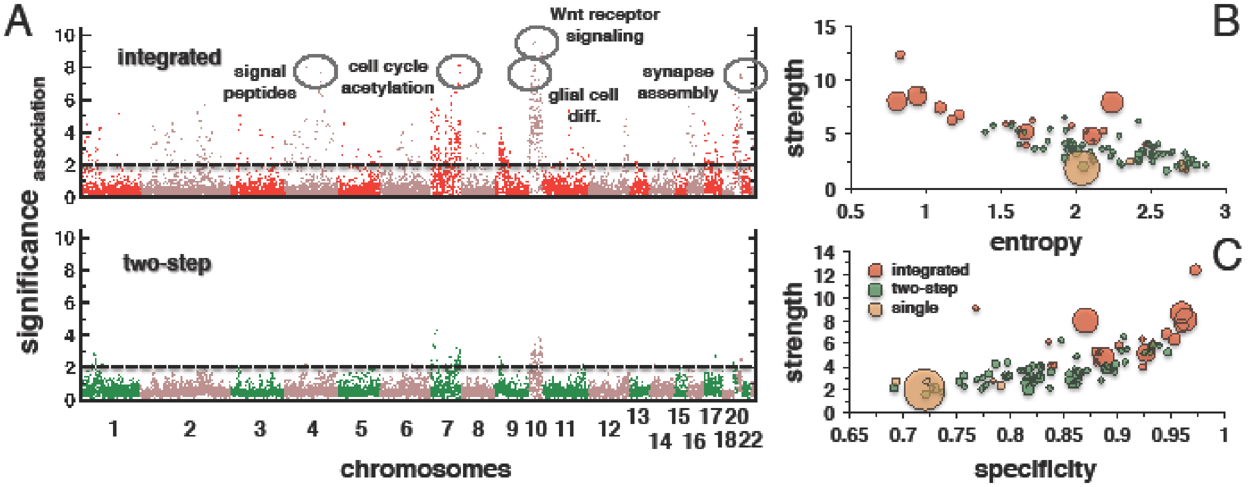


Figure 2: Comparison of module covers approaches in GBMs (A) Manhattan plots of module associations show average association significance for each tag-locus for the 10 largest modules we obtained with both methods. Modules obtained using the integrated method had more significant eQTL associations. In the upper panel, we also labeled associations that correspond to functionally coherent modules shown in Online Appendix Fig. A2. **(B, C)** Comparing the quality of modules, we observed that the Integrated method generated modules with higher strength, lower entropy and higher specificity Module size is indicated by the sizes of corresponding circles. The label "single" refers to modules we obtained using a set cover approach.

We observed that associations obtained by the two-step algorithm were weaker based on several different measures of quality introduced below.

To compare the approaches more quantitatively, first note that the total cost of modules selected by the integrated and two-step algorithms was 744 and 1439.05, respectively (Appendix, Table A1). The total weights between genes in modules (the third term in cost function (3)) were 18.63 and -184.05, showing that the modules selected by the integrated algorithm were much more coherent compared to the modules obtained by two-step algorithm.

To further quantify the quality of modules in terms of their association to genomic alterations, we devised several measures: The *strength* of association significance of a module was defined as the maximum significance of the associations of the given module over all loci:

$$\textit{Strength}\ (M) = \max_i s_i\,(M). \qquad (6)$$

We also computed the entropy of association profiles for each module. Since entropy measures the uncertainty of data, a good quality module (with only a few strong associations) is expected to have low entropy while entropy increases as data is more uniformly distributed. Formally, for each module M, we partitioned the range from 0 to strength *(M)* into 10 bins of equal sizes and assigned loci according to their significance. In each bin, we computed the percentage $\boldsymbol{p_j}$ of loci and defined the entropy as

$$\textit{Entropy}\ (M) = -\sum_{j\,\in\,bins} p_j\ \log_2 p_j \qquad (7)$$

For an association to be specific in a given module, only a few regulatory associations should have highly significant p-values while the remaining loci are expected to have insignificant p-values. Thus, we defined the specificity of a module M as the area of a cumulative histogram of association significance values. Specifically, we partitioned the range from *0 to strength (M)* into 10 bins of equal sizes and defined c_j to be the cumulative percentage of j-th bin. Then the specificity is defined as:

$$Specificity\ (M) = \sum_{j\,\in\,bins} c_j/|bins| \qquad (8)$$

Similar to entropy, specificity quantifies the distinction between significant associations and the remainder of the loci. However *specificity* approaches 1 only if a small number of significant loci exist whereas theoretically entropy can be low in the case when there is a few insignificant and many significant loci.

We found that the integrated module cover outperformed the two-step module cover approach based on all three measures (as summarized in Online Appendix Table A1). The average strength of modules (size ≥ 3) selected by the integrated module cover algorithm was 6.4, significantly outscoring an average of 3.6 of modules obtained by the two-step module cover algorithm ($P < 10^{-8}$, Wilcoxon test). Similarly, the average specificity for the integrated module cover was 0.9 while the average was 0.83 for the two-step cover ($P < 10^{-4}$, Wilcoxon test). The average entropy of modules selected by the integrated algorithm and two-step cover were 1.6 and 2.2, respectively ($P < 10^{-4}$, Wilcoxon test).

Fig. 2B,C presents a detailed comparison of the performance of the module cover approaches with respect to the mentioned measures. In addition, we included results obtained by the basic set cover algorithm labeled “single” in Figs. 2 B,C using the same parameter k = 300 and at most 5 outliers. In this case we defined the modules as the connected components of the subgraph spanned by the genes that were selected as the cover. We observed that modules of size ≥3 obtained by the integrated module cover approach were on average larger than modules found with the two-step approach. Specifically, modules identified by the integrated approach had significantly smaller entropy compared to modules obtained by the two-step approach (Fig. 2B, $P < 10^{-6}$, Kolmogorov-Smirnov test). In addition, these modules showed significantly higher strength (Fig. 2C, $P < 10^{-5}$, Kolmogorov-Smirnov test). However, the quality of modules obtained with both approaches was still superior to results of a single gene set cover, demonstrating general benefits of the module cover approach.

All alogrithms were implemented in Python and compute the solutions for the inputs of ~10,000 genes in a few minutes on NCBI linux machines.

3.1.2. *Analysis of GBM data*

We further analyzed modules provided by the integrated method. First, we determined enriched GO terms in modules using BINGO (*15*). Out of 21 modules with at least 3 genes, we found 14 modules having at least one GO term that they significantly enriched with (FDR < 0.05). In addition to modules enriched with typical cancer-related processes such as cell division, cell cycle, and immune response we also obtained more glioma-specific modules such as the WNT signaling pathway and glial cell differentiation. For example, only some subsets show dys-regulation of immune response or of WNT signaling while the cell cycle module is dys-regulated in almost all samples. Although our modules have been selected by using eQTL association profiles they allow us to recover GBM subtypes that previously were determined with expression profiles of single genes. Importantly, we observed that different modules were covering different sets of samples in a nonhierarchical (non-nested) way (Online Appendix, Fig. A2). This overlapping pattern of covering modules might explain why the number of GBM subtypes has been difficult to establish (*16, 17*).

3.1.3. *Analysis of Ovarian Cancer Data*

We also used the integrated module cover algorithm to analyze a set of 489 Ovarian Cancer samples from The Cancer Genome Atlas (TCGA). Applying the integrated module cover algorithm with k=70, $\alpha = 1$, and 25 outliers, we selected 485 genes grouped in 235 modules including 54 non-singleton modules. As in the analysis of GBM data, we choose k for which the number of nontrivial modules starts to level. Out of 12 modules of size at least 5, 9 modules were enriched with at least one GO terms significantly (FDR < 0.05).

To visualize the coverage of disease cases by modules of size ≥5, we counted the number of genes covering each sample (Fig 3A). Similarly to GBMs, we found that different modules are covering different subsets of samples. Note that a gene may cover a sample when it is either significantly upregulated or downregulated. In Fig 3B, we investigated the expression patterns of individual genes in the modules. Performing hierarchical clustering of the genes based on expression level, we obtained clusters consistent with the existing classification of cancer subtypes (*18*), in which the gene expression profile of ~1,000 selected genes was used to define 4 disease subtypes. Using only 185 genes in the 12 largest modules from our module cover, we successfully recovered these 4 subtypes (Fig 3B) despite the fact that these genes have not been selected explicitly to classify expression based subtypes. In the TCGA analysis (*18*), the authors attempted to identify genes whose differential expression helped to define each disease subtype. However, we found that our module-based analysis often provided a more informative picture. For example, in (*18*) one subgroup of the collagen gene family was found to support the Mesenchymal subtype while another subgroup of this family as well as the LUM gene which binds collagen fibrils was associated to the Differential subtype. In contrast, our approach grouped all these genes into “extracellular matrix organization” module, also containing several matrix metalloproteinase (MMP) genes. We found that genes in this module had very similar expression and were overexpressed in the Mesenchymal subtype.

4. Discussion

Uncovering modules that are associated with genomic alterations in a disease is a challenging task as well as an important step to understand complex diseases. To address this challenge we introduced a novel technique - module cover - that extends the concept of set cover to network modules. We provided a mathematical formalization of the problem and developed two heuristic solutions: the Integrated Module Cover approach, which greedily selects genes to cover disease cases while simultaneously detecting modules and a Two-Step approach that first detects modules and subsequently selects a cover.Using several quality measures, we established that the integrative approach outperformed the alternative two-step approach. However, both methods showed better performance than a naïve single gene based set cover approach. We also constructed modules utilizing gene expression rather than association profiles to define a similarity measure (data not shown). We observed that the modules obtained by the integrated approach based on gene expression showed lower association specificity/association strength than modules that were provided by eQTL profiles. However, expression based modules would be clearly preferred for uncovering expression patterns that occurs regardless of the association to genetic variations.

In general, the module cover approach is especially helpful in analyzing and classifying heterogeneous disease cases by exploring the way different combinations of dys-regulated of modules relate to a particular disease subcategory. Indeed, our analysis indicated that the gene set selected by module cover approach may be used for classification. Equally important, the selected module covers may help to interpret classifications that were obtained with other methods.

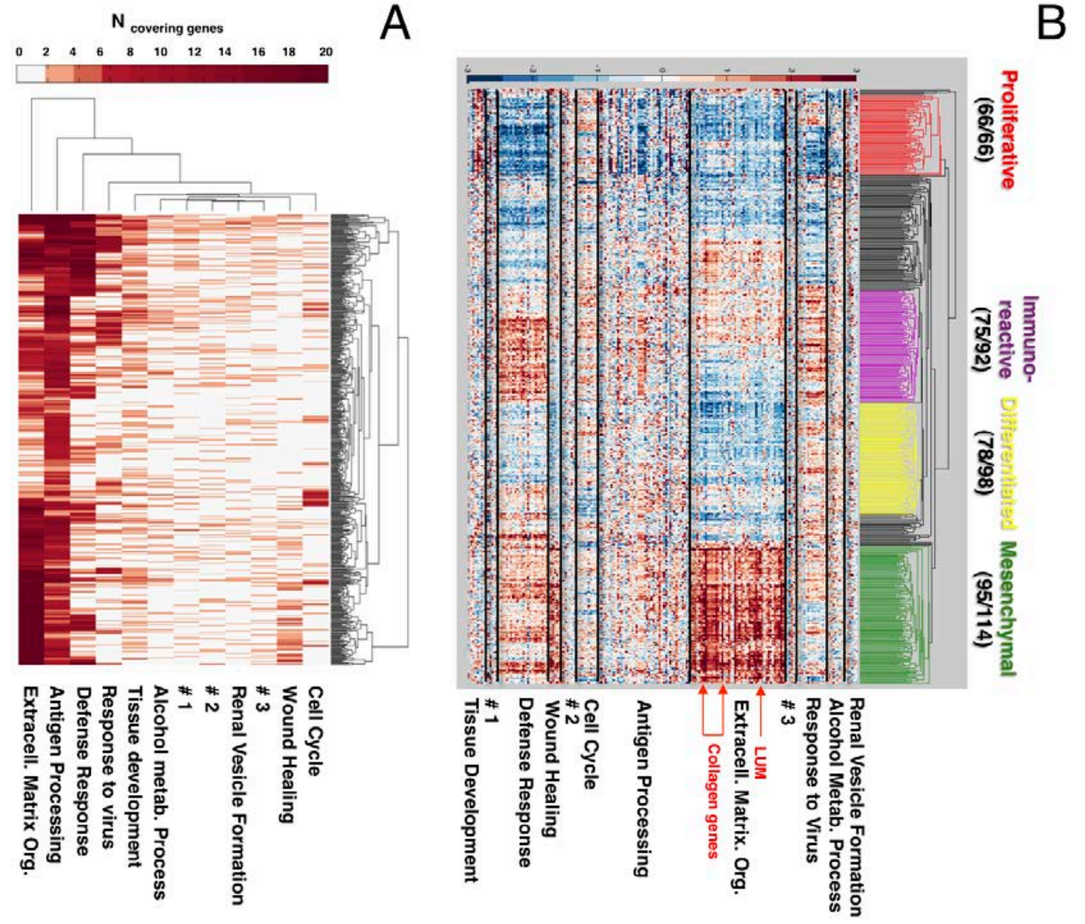


Figure 3: Modules in ovarian cancer obtained by the integrated module cover method. (A) For each disease case (y-axis) we displayed in the heat map the number of genes in each module that covered the sample **(B)** Expression based clustering of the genes in the modules provided clusters consistent with the existing classification of cancer subtypes. Arrows indicate genes of the extracellular matrix module discussed in the text. The fraction of genes assigned to a given cluster in (*18*) is shown next to the cluster name.

5. Materials

5.1 Data Treatment for Glioblastoma Multiforme Data from GMDI

Differentially Expressed Genes: Briefly, all samples were profiled using HG-U133 Plus 2.0 arrays that were normalized at the probe level with dChip (*16, 19*). Among probes representing each gene, we chose the probeset with the highest mean intensity in the tumor and control samples. We determined genes that are differentially expressed in each disease case compared to the non-tumor control cases with a Z-test. For a gene g and case c, we define cover(c, g) to be 1 if nominal p-value < 0.01 and 0 otherwise.

eQTL Profiles: To detect copy number alterations, samples were hybridized on the Genechip Human Mapping 100K arrays, and copy numbers were calculated using Affymetrix Copy Number Analysis Tool (CNAT 4). After probe-level normalization and summarization, calculated log2-tranformed ratios were used to estimate raw copy numbers. Using a Gaussian approach, raw SNP profiles were smoothed (> 500 kb window by default) and segmented with a Hidden Markov Model approach (*20-22*). We first performed local clustering, allowing us to obtain 911 tag loci (11). For each gene/tag-locus pair, we computed nominal p-values by linearly regressing gene expression and genomic alteration for all samples. We then define the eQTL significance profile for each gene, *g*, *as* $Assoc(g) = \{s_1(g), s_2(g), \ldots s_{911}(g)\}$, where $s_i(g)$ represents the $-log_{10}$ *p*-value of the association given by the linear regression between expression values of gene *g* and copy number variation of locus *i*. Using such profiles, we defined the similarity of two genes g_1 and g_2, $sim(g_1, g_2)$, as Pearson's correlation coefficient of $Assoc(g_1)$ and $Assoc(g_2)$.

Weights of Gene Pairs: We utilized human protein-protein interaction data from large-scale high-throughput screens (23-25) and several curated interaction databases (26-29), totaling 93,178 interactions among 11,691 genes. As a reliable source of experimentally confirmed protein-DNA interactions, we used 6,669 interactions between 2,822 transcription factors and structural genes from the TRED database (30). As for phosphorylation events between kinases and other proteins we found 5,462 interactions between 1,707 human proteins utilizing networKIN (31, 32) and phosphoELM database (33). Combining all interactions, the network contains 11,969 human proteins and 103,966 interactions. We computed the weights of each gene pairs using equation (1) with avg_distance = 3.6 and $\theta = 0.63$, a threshold that corresponds to the top 1%ile of weights of any pairs.

5.2 Data Treatment for Ovarian Cancer Data from TCGA

We utilized the unified expression data compiled in (*18*) based on expression values from three different expression platforms. Since there is no control (non-cancer data) in this dataset, we defined that a gene covers a sample if its expression in this sample was in the extreme 3% of the expression distribution. We then narrowed down the set of genes to 1,889 genes by considering genes that covered at least 5% of the samples. As for copy number variations, we used level 4 data obtained with GISTIC (*34*) and selected 1,923 genes with copy number alterations (calls = ±2) in at least 5% of all samples. For each differentially expressed gene we used linear regression to compute associations of the expression of this gene with copy number variation of each of the 1,923 genes. We used p-values of these associations to compute association profiles as explained in 5.1. Edge weights in interaction graph were calculated as described in 5.1 with θ = 0.58, a threshold corresponding to the top 5% ile.

References

1. I. Ulitsky, R. Karp, R. Shamir, in *Research in Computational Molecular Biology*. (2008), pp. 347-359.
2. S. A. Chowdhury, M. Koyuturk, *Pac Symp Biocomput*, 133 (2010).
3. H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, T. Ideker, *Mol Syst Biol* **3**, 140 (2007).
4. E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, D. Lee, *PLoS Comput Biol* **4**, e1000217 (Nov, 2008).
5. I. Ulitsky, A. Krishnamurthy, R. M. Karp, R. Shamir, *PLoS One* **5**, e13367 (2010).
6. T. Ideker, O. Ozier, B. Schwikowski, A. F. Siegel, *Bioinformatics* **18 Suppl 1**, S233 (2002).
7. F. Vandin, E. Upfal, B. J. Raphael, *J Comput Biol* **18**, 507 (Mar, 2011).
8. Vandin F., P. Clay, E. Upfal, R. B.J., in *Pacific Symposium on Biocomputing*. (2012).
9. I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, E. M. Marcotte, *Genome Res* **21**, 1109 (Jul, 2011).
10. Y. A. Kim, J. H. Przytycki, S. Wuchty, T. M. Przytycka, *Phys Biol* **8**, 035012 (Jun, 2011).
11. Y. A. Kim, S. Wuchty, T. M. Przytycka, *PLoS Comput Biol* **7**, e1001095 (Mar, 2011).
12. S. Kim, K. A. Sohn, E. P. Xing, *Bioinformatics* **25**, i204 (Jun 15, 2009).
13. W. Zhang, J. Zhu, E. E. Schadt, J. S. Liu, *PLoS Comput Biol* **6**, e1000642 (Jan, 2010).
14. A. J. Enright, S. Van Dongen, C. A. Ouzounis, *Nucleic Acids Res* **30**, 1575 (Apr 1, 2002).
15. S. Maere, K. Heymans, M. Kuiper, *Bioinformatics* **21**, 3448 (Aug 15, 2005).
16. A. Li *et al.*, *Cancer Res* **69**, 2091 (Mar 1, 2009).
17. R. Shen *et al.*, *PLoS One* **7**, e35236 (2012).
18. *Nature* **474**, 609 (Jun 30, 2011).
19. C. Li, W. H. Wong, *Proc Natl Acad Sci U S A* **98**, 31 (Jan 2, 2001).
20. Y. Kotliarov *et al.*, *Cancer Res* **66**, 9428 (Oct 1, 2006).
21. R. C. Gentleman *et al.*, *Genome Biol* **5**, R80 (2004).
22. J. Fridlyand, A. M. Snijders, D. Pinkel, G. Albertson, A. N. Jain, *Journal of Multivariate Analysis* **90**, 132 (2004).
23. R. M. Ewing *et al.*, *Mol Syst Biol* **3**, 89 (2007).
24. J. F. Rual *et al.*, *Nature* **437**, 1173 (Oct 20, 2005).
25. U. Stelzl *et al.*, *Cell* **122**, 957 (Sep 23, 2005).
26. A. Chatr-aryamontri *et al.*, *Nucleic Acids Res* **35**, D572 (Jan, 2007).
27. S. Kerrien *et al.*, *Nucleic Acids Res* **35**, D561 (Jan, 2007).
28. L. Matthews *et al.*, *Nucleic Acids Res*, (Nov 3, 2008).
29. S. Peri *et al.*, *Nucleic Acids Res* **32**, D497 (Jan 1, 2004).
30. C. Jiang, Z. Xuan, F. Zhao, M. Q. Zhang, *Nucleic Acids Res* **35**, D137 (Jan, 2007).
31. R. Linding *et al.*, *Cell* **129**, 1415 (Jun 29, 2007).
32. R. Linding *et al.*, *Nucleic Acids Res* **36**, D695 (Jan, 2008).
33. F. Diella, C. M. Gould, C. Chica, A. Via, T. J. Gibson, *Nucleic Acids Res* **36**, D240 (Jan, 2008).
34. R. Beroukhim *et al.*, *Proc Natl Acad Sci U S A* **104**, 20007 (Dec 11, 2007).

Online appendix: http://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/Modue_Cover/Appendix

NEXT-GENERATION ANALYSIS OF CATARACTS: DETERMINING KNOWLEDGE DRIVEN GENE-GENE INTERACTIONS USING BIOFILTER, AND GENE-ENVIRONMENT INTERACTIONS USING THE PHENX TOOLKIT*

SARAH A. PENDERGRASS

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 503 Wartik Lab
University Park, PA 16802, USA
Email: sap29@psu.edu

SHEFALI S. VERMA

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: szs14@psu.edu

EMILY R. HOLZINGER

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: Emily.R.Holzinger@vanderbilt.edu

CARRIE B. MOORE

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: ccb12@psu.edu

JOHN WALLACE

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: jrw32@psu.edu

SCOTT M. DUDEK

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: sud23@psu.edu

WAYNE HUGGINS

RTI International
Research Triangle Park, NC,USA
Email: whuggins@rti.org

TERRIE KITCHNER

Marshfield Clinic
Marshfield, WI, USA
Email: Kitchner.Terrie@mcrf.mfldclin.edu

CAROL WAUDBY

Marshfield Clinic
Marshfield, WI, USA
Email: WAUDBY.CAROL@mcrf.mfldclin.edu

RICHARD BERG

Marshfield Clinic
Marshfield, WI, USA
Email: Berg.Richard@mcrf.mfldclin.edu

CATHERINE A. MCCARTY

Essential Rural Health
Duluth, MN, USA
Email: CMcCarty@eirh.org

MARYLYN D. RITCHIE

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: Marylyn.ritchie@psu.edu

Investigating the association between biobank derived genomic data and the information of linked electronic health records (EHRs) is an emerging area of research for dissecting the architecture of complex human traits, where cases and controls for study are defined through the use of electronic phenotyping algorithms deployed in large EHR systems. For our study, 2580 cataract cases and 1367 controls were identified within the Marshfield Personalized Medicine Research Project (PMRP) Biobank and linked EHR, which is a member of the NHGRI-funded electronic Medical Records and Genomics (eMERGE) Network. Our goal was to explore potential gene-gene and gene-environment interactions within these data for 529,431 single nucleotide polymorphisms (SNPs) with minor allele frequency > 1%, in order to explore higher level associations with cataract risk beyond investigations of single SNP-phenotype associations. To build our SNP-SNP interaction models we utilized a prior-knowledge driven filtering method called Biofilter to minimize the multiple testing burden of exploring the vast array of interaction models possible from our extensive number of SNPs. Using the Biofilter, we developed 57,376 prior-knowledge directed SNP-SNP models to test for association with cataract status. We selected models that required 6 sources of external domain knowledge. We identified 5 statistically significant models with an interaction term with p-value < 0.05, as well as an overall model with p-value < 0.05 associated with cataract status. We also conducted gene-environment interaction analyses for all GWAS SNPs and a set of environmental factors from the PhenX Toolkit: smoking, UV exposure, and alcohol use; these environmental factors have been previously associated with the formation of cataracts. We found a total of 288 models that exhibit an interaction term with a p-value $\leq 1\text{x}10^{-4}$ associated with cataract status. Our results show these approaches enable advanced searches for epistasis and gene-environment interactions beyond GWAS, and that the EHR based approach provides an additional source of data for seeking these advanced explanatory models of the etiology of complex disease/outcome such as cataracts.

[*] This work supported by the following grants: U19 HL0659625, R01 LM010040, U01 HG006389

1. Introduction

DNA biobanks coupled to electronic health records (EHR) have become a valuable resource for investigating the genetic architecture of complex traits, as EHR contain a wide array of medical information including billing codes and clinical laboratory measurements, often yielding a large sample size. Through carefully defining phenotypes, and using deployable algorithms that combine multiple sources of information in the EHR, cases and controls can be defined for association studies, such as defining age-related cataract cases and controls [1,2]. The Marshfield Personalized Medicine Research Project Biobank (Marshfield PMRP) and linked EHR, used for the study described herein, is one such resource [3]. The Marshfield PMRP is a member of the NHGRI-funded electronic Medical Records and Genomics (eMERGE) Network, a network of similar Biobanks coupled with EHR based data [4].

Cataracts are a leading cause of blindness globally [5], and are believed to arise from a combination of age, environmental factors, and heritable factors [6]. Thus, understanding the genetic etiology of cataracts, coupled with the effect of environment as a modifier, could have a profound impact on human health. For our study, algorithms proven for age-related cataract case identification [2] were deployed in the Marshfield PMRP EHR to identify 2580 cataract cases and 1367 controls, with further study details presented in Table 1. A total of 529,431 single nucleotide polymorphisms (SNPs) were available after PMRP genotyping coupled with quality control filtering and selection for SNPs with a minor allele frequency > 1%.

Table 1. Marshfield Cataract Study Description

Age	$\geq$ 50
Ancestry	European American
Total Sample Size	3947
# Controls	1367
# Cases	2580
After QC & Allele Frequency Cutoff	
# Controls	1364
# Cases	2576
% Women Controls	59
% Men Controls	41
% Women Cases	59
% Men Cases	41

Single SNP-phenotype associations are a dominant study design used in most genome-wide association studies (GWAS), however, more complex models that include interactions may more accurately describe the relationship between genetic variation and complex outcomes. Investigating all gene by gene (GxG), and in extension, all SNP by SNP (SNPxSNP) pairwise models is possible depending on the number of SNPs that have been genotyped. Unfortunately, the multiple hypothesis testing burden and risk of Type I error is inflated when investigating all pairwise models. A different approach can be used, utilizing prior biological knowledge methods directing model development. Thus, to investigate more complex models beyond single SNP-Phenotype associations for the Marshfield PMRP cataract dataset, we used the prior knowledge accessible through Biofilter 1.0 (a new implementation of Biofilter after the original description in [7]) to direct the investigation of pairwise GxG interaction models based on the following resources: the Kyoto Encyclopedia of Genes and Genomes (KEGG) [8], Reactome [9], Gene Ontology (GO) [10], the Protein families database (Pfam) [11], NetPath [12], Biological General Repository for Interaction Datasets (BioGrid) [13], and the Molecular INTeraction Database (MINT) [14]. Using the Biofilter, we developed 57,376 prior-knowledge directed SNPxSNP models to test for association with cataract status.

In addition, for this study we investigated gene-environment interactions (GxE), as there are clearly known environmental exposures that increase cataract risk, and when incorporated into analyses, may provide new models for the contribution of both environment and genetic architecture to cataracts. The Marshfield PMRP collected standardized Phenotypes and eXposures (PhenX) measures as a member of the PhenX Real-world, Implementation, SharingING (PhenX RISING) project. PhenX has the goal of defining standard phenotypic measures through a framework of measurement protocols via a web-based toolkit [15]. Environmental exposures such as smoking, sun exposure, and alcohol use, have been associated with increased cataract rates [16]. Thus we used 12 PhenX defined environmental exposures to investigate GxE interactions for the Marshfield PMRP cataract data focused on smoking, UV exposure, and alcohol use measures.

Through integrating EHR data, advanced bioinformatics tools, and PhenX, we can pursue advanced searches for epistasis and gene-environment interactions in genome-wide studies of common disease.

2. Methods

2.1. *Marshfield EHR and Age-Related Cataract Case Identification*

The Marshfield PMRP is a population based biobank with ~20,000 subjects, aged 18 years and older, enrolled in the Marshfield Clinic healthcare system in central Wisconsin [3]. DNA, plasma, and serum samples are collected at the time the enrollee completes a written informed consent document, with allowance for ongoing access to the linked medical records. PMRP participants also complete questionnaires, including responses regarding smoking history, occupation, and diet.

To identify cataract surgery cases aged 50 years and older within the PMRP, Current Procedural Terminology (CPT) codes in the Marshfield Clinic EHR were used. A research coordinator manually abstracted additional information to identify the eye affected, the type and severity of the cataract, and the level of visual acuity prior to the cataract surgery. This was also done to remove any cases with non-age related cataracts.

To identify individuals with diagnosed cataracts but without surgery, and to identify the type of cataract, International classification of diseases, 9th revision (ICD-9) codes and CPT codes were used, coupled with Natural Language Processing (NLP) and Intelligent Character Recognition (ICR) of free-text in the EHR. NLP and exclusion criteria were used to identify individuals with congenital and traumatic cataracts for omission from the study. Further details of the identification of cataract cases and controls and the efficacy of the EHR defined phenotyping can be found in Waudby et al., 2011 [2].

All total, the procedures used on the EHR identified 2580 cataract cases and 1367 controls in the Marshfield PMRP data.

2.2. *Genotyping*

The eMERGE network and the Center for Inherited Disease Research (CIDR) at Johns Hopkins university performed the genotyping of the Marshfield PMRP samples, using the Illumina Human660W-Quadv1 A platform with total of 560,635 SNPs and 96,731 intensity-only probes. Bead Studio version 3.3.7 was used by CIDR for the genotyping calls. The total cohort genotyped included 3947 samples from the Marshfield PMRP, 21 blind duplicates, and 85 HapMap controls. The HapMap concordance rate was 99.8% and the blind duplicate reproducibility rate was 99.99%. For quality control and data cleaning the eMERGE quality control (QC) pipeline developed by the eMERGE Genomics Working Group [17] was used. Any SNPs with a minor allele frequency > 1% , SNP call rate > 99%, Sample Call Rate > 99% were used in further analysis. After QC and allele frequency filtering using PLINK [18], a total of 529,431 SNPs were used for further analysis.

2.3. *PhenX*

The standardized phenotypic and environmental consensus measures for Phenotypes and eXposures (PhenX) [15] were used to capture the environmental variables used in this study. The PhenX Toolkit (https://www.phenxtoolkit.org/) offers high-quality, well-established, standard measures of phenotypes and exposures for use in epidemiological studies.

The Marshfield PRMP is part of the PhenX RISING consortium, which is comprised of seven groups funded by the National Human Genome Research Institute (NHGRI) and the Office of Behavioral and Social Sciences Research (OBSSR) to incorporate PhenX (https://www.phenxtoolkit.org/) measures into existing population-based genomic studies.

Table 2. The PhenX measures used for this study

PhenX Measure	Survey Question
PX030301 Alcohol 30Day Frequency	During the past 30 days, on how many days did you drink one or more drinks of an alcoholic beverage?
PX030301 Alcohol 30Day Quantity	During the past 30 days, how many drinks did you usually have each day?
PX030602 Cigarette Smoking 100	Have you smoked at least 100 cigarettes in your entire life?
PX030602 Cigarette Smoking Current	Do you now smoke cigarettes every day, some days, or not at all?
PX030602 Cigarette Smoking Everyday 6Month	Have you EVER smoked cigarettes EVERY DAY for at least 6 months?
PX030802 Everyday Smoker Quantity 1Day	On the average, about how many cigarettes do you now smoke each day?
PX030802 Someday Smoker Days 1Month	On how many of the past 30 days did you smoke cigarettes?
PX030802 Someday Smoker Quantity 1Day	On the average, on those days, how many cigarettes did you usually smoke each day?
PX030802 Former Smoker Smoking 6Month	Have you EVER smoked cigarettes EVERY DAY for at least 6 months?
PX030802 Former Smoker Quantity 1DayA	When you last smoked every day, on average how many cigarettes did you smoke each day?
PX030802 Former Smoker Quantity 1DayB	When you last smoked fairly regularly, on average how many cigarettes did you smoke each day?
PX061301 Weekend Sun Hours Last Decade	On a typical weekend day in the summer, about how many hours did you generally spend in the mid-day sun in the past ten years?

For this initiative, Marshfield PRMP subjects with GWAS data who were alive with known, non-institutionalized addresses and who had given consent for re-contact were mailed a 32-page self-administered questionnaire that contained 35 PhenX measures across a range of phenotypic domains including alcohol and tobacco use questions (McCarty et al. 2012, *in preparation*).

For this study, we considered 12 of these measures, shown in Table 2.

2.4. *BioFilter 1.0*

For the SNPxSNP analysis, Biofilter 1.0 was used. Biofilter has been upgraded from the initial Biofilter 0.5 [7], with the addition of more data sources, improved the handling of data, and the development of an eternal database for prior knowledge called the Library of Knowledge Integration (LOKI). Biofilter 1.0 and LOKI are freely available for non-commercial research institutions. For full details see: *http://ritchielab.psu.edu/ritchielab/software*.

Biofilter 1.0 utilizes prior biological knowledge through accessing the data of several publically available biological information databases, all compiled within the LOKI database developed specifically for Biofilter. The data sources selected for Biofilter contain information on networks, connections, and/or pathways that establish relationships between genes and gene products. Biofilter is a "gene based" approach, thus all the region information (such as genes) and position information (such as SNPs) are mapped to genes within LOKI.

The following sources that are compiled within LOKI were used for the Biofilter model building: the Kyoto Encyclopedia of Genes and Genomes (KEGG) [8], Reactome [9], Gene Ontology (GO) [10], the Protein families database (Pfam) [11], and NetPath [12], Biological General Repository for Interaction Datasets (BioGrid) [13], and the Molecular INTeraction Database (MINT) [14]. The database source used in LOKI solely for the purpose of mapping SNPs to genes is the National Center for Biotechnology (NCBI) dbSNP [19] database.

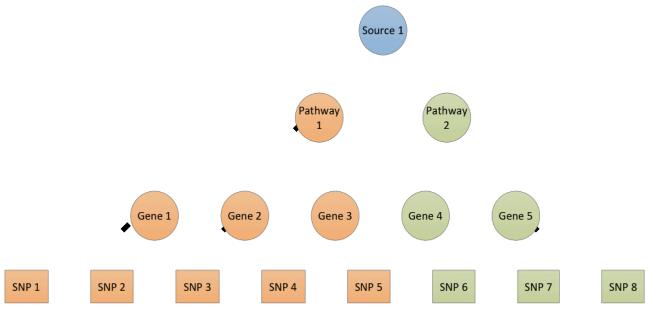


Figure 1. Simplified model for one Biofilter 1.0 database source with 2 pathways, 5 genes, and 8 SNPs

The following process was used within Biofilter 1.0 to develop the SNPxSNP models used in prior knowledge directed association testing. Figure 1 shows a simplified example of how the Biofilter 1.0 model generation process works. First, the input list of SNPs are mapped to genes within Biofilter. Next, comprehensive pairs of genes that are all terminal leaves of the graph for Pathway 1 in Source 1, and Pathway 2 in Source 1 are generated, only for genes that contain SNPs in the input list of SNPs.

Implication scores are used in Biofilter to give each pairwise model a "score" indicating how many sources have that connected pair of genes represented, the higher the implication score, the more sources have indicated a connection between a pair of genes. The implication index is a measure of the number of data sources providing evidence of an interaction, a sum of the number of data sources supporting each of the two genes and the connection between them. In the case of our simplified example, for Genes 1-5, that all contain SNPs within the input list, the following pairwise Gene-Gene models would result, each with an implication score of 1:

Gene1 – Gene 2
Gene1 – Gene 3
Gene 2 – Gene 3
Gene 4 – Gene 5

This process continues through all other sources used for Biofilter. Each time a pairwise combination of genes is found in another source (such as the pair Gene1-Gene2), the implication score for that pairwise model will be increased by 1. Lastly, the G-G models are broken into all pairwise combinations of SNPs across the genes, *within P1 or P2*. The SNP-SNP models would look like the following:

SNP1-SNP3
SNP1-SNP4
SNP1-SNP5
SNP2-SNP3
SNP2-SNP4
SNP2-SNP5
SNP3-SNP5
SNP3-SNP4
SNP6-SNP7
SNP6-SNP8

This same process was used within Biofilter 1.0 to develop the SNPxSNP models used for our prior knowledge directed association testing. First, the 529,431 SNPs were mapped to their corresponding genes. Next, the genes corresponding to the SNPs of the dataset were mapped to the gene-relationship graphs for each LOKI source used. After this mapping process, gene pairs were exhaustively generated for each occurrence of two genes within a single pathway and single source. Implication scores were calculated for the pairwise models. After the gene-gene models were developed in Biofilter, the models were divided into exhaustive SNP-SNP pairs for association testing.

Table 3 indicates the number of models that were found at each implication score cutoff. An implication index cutoff of 4 actually incorporates all possible pairwise models for all SNPs we had for this study, a total of 603,032 models. We found an implication score cutoff of 6 resulted in a balance between a large group of models for exploration (57,376 models), but still maintained a very computationally feasible set of associations to investigate, limiting our type 1 error rate more than using all exhaustive pairs of SNP-SNPs or some of the less stringent implication score cutoffs. With a requirement for an implication index of 6, as we had in this study, the gene-gene relationship or known interaction had to be found in nearly all of the data sources we used within LOKI.

Table 3. Number of Resulting Models for Each Implication Score Cutoff

Implication Index Cutoff	Number of Models
4	603032
5	337113
6	57376
7	2479

2.5. *Statistical Analysis*

For the SNPxSNP models generated through the use of Biofilter, PLATO [20] was used to determine the significance of the interaction term (via a t-test), and the significance of the overall model (via an F test). The full model was: SNP1 + SNP2 + SNP1*SNP2, for all of the pairwise sets of SNPs generated by Biofilter with an implication index of 6.

For the GxE (SNPxE) models, the full model was: SNP1 + ENV1 + SNP1*ENV1, for all the possible unique SNPxE pairs, from the set of 529,431 SNPs and the PhenX variables described earlier in methods. Again, the outcome was case control status for cataracts. PLINK [18] was used, and results were maintained for further inspection that had an P-interaction term < 0.05. The GGPlot2 [21] package in R was used for Figure 2.

3. Results

3.1. *GxE Results*

Figure 2 shows a Manhattan plot of the association results for the PhenX GxE models that had interaction p-values ≤ $1x10^{-2}$, a total of 288 models exhibited an interaction term with a p-value ≤ 1x10-4 associated with cataract status. The top five GxE interaction results for each PhenX measure are also presented in Table 4, sorted by chromosome to highlight results similar across SNPs and regions for multiple PhenX measures. The measurement "PX030802 Former Smoker Smoking 6Month" a survey question asking "Have you ever smoked cigarettes every day for at least 6 months?" with the SNP rs2058131, near the genes *ZNF471* and *ZFP28* on Chromosome 19, that had an association interaction term p-value of $2.72x10^{-7}$, was the most significant interaction term p-value found when compared to the other 12 PhenX measurements we used in our GxE analysis.

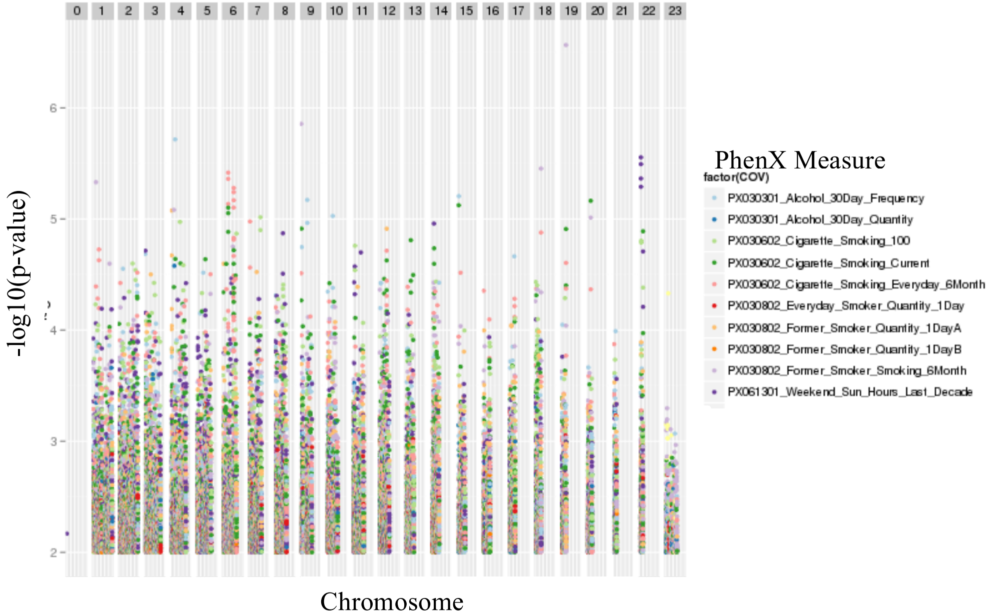


Figure 2. Manhattan plot of the association results for the GxE interaction models. Displayed are the results for the 10 PhenX measures that had interaction p-values < $1x10^{-2}$, two PhenX measures did not have an interaction p-value less than $1x10^{-2}$.

Table 4. Five most significant GxE results for each PhenX measurement, sorted by chromosome and gene

Chr	BP	RSID	PhenX variable	P-value	Gene
1	38802669	rs4568792	Former Smoker Smoking 6 Month	4.67x10^{-6}	
1	233192767	rs7412124	Alcohol 30 Day Quantity	0.00022	
2	52547408	rs6726893	Everyday Smoker Quantity 1 Day	0.00047	
3	69335296	rs12494107	Someday Smoker Days 1 Month	0.019	*FRMD4B*
3	99983201	rs13091236	Alcohol 30 Day Quantity	0.00028	*ST3GAL6*
3	100016640	rs13059624	Alcohol 30 Day Quantity	0.00025	*DCBLD2*
3	100092139	rs13084692	Alcohol 30 Day Quantity	0.00028	*DCBLD2*
4	8423988	rs747580	Former Smoker Quantity 1 Day A	2.13x10^{-5}	*ACOX3*
4	8480717	rs2631731	Former Smoker Quantity 1 Day A	8.40x10^{-6}	*ACOX3*
4	37295333	rs2048257	Former Smoker Smoking 6 Month	8.29x10^{-6}	*RELL1*
4	42991353	rs17457584	Alcohol 30 Day Quantity	2.64x10^{-5}	
4	53149495	rs346005	Alcohol 30 Day Frequency	1.92x10^{-6}	
4	114190823	rs1026975	Cigarette Smoking 100	1.06x10^{-5}	*ANK2*
5	123324046	rs2250107	Everyday Smoker Quantity 1 Day	0.00049	
5	123343048	rs2546839	Everyday Smoker Quantity 1Day	0.00044	
6	4098136	rs653674	Former Smoker Quantity 1Day A	2.53x10^{-5}	
6	66776318	rs6899720	Cigarette Smoking Current	8.87x10^{-6}	
6	66776318	rs6899720	Cigarette Smoking Everyday 6 Month	3.83x10^{-6}	
6	66900502	rs12528760	Cigarette Smoking Current	7.93x10^{-6}	
6	66900502	rs12528760	Cigarette Smoking Everyday 6 Month	4.36x10^{-6}	
6	135370485	rs6929661	Cigarette Smoking Everyday 6 Month	6.78x10^{-6}	*HBS1L*
6	135376293	rs1014021	Cigarette Smoking Everyday 6 Month	5.74x10^{-6}	*HBS1L*
6	135407511	rs6569990	Cigarette Smoking Everyday 6 Month	5.26x10^{-6}	*HBS1L*
6	170414666	rs3012437	Cigarette Smoking 100	1.36x10^{-5}	*LOC285804*
7	97704227	rs3735258	Former Smoker Quantity 1 Day A	3.01x10^{-5}	*DKFZP434B0335*
7	144435178	rs10254774	Cigarette Smoking 100	1.26x10^{-5}	
9	2449444	rs1006575	Former Smoker Smoking 6 Month	1.39x10^{-6}	
9	6256440	rs2026991	Cigarette Smoking Current	1.22x10^{-5}	
9	79252094	rs10116050	Alcohol 30 Day Frequency	1.09x10^{-5}	*GNA14*
9	79255890	rs4745639	Alcohol 30 Day Frequency	6.74x10^{-6}	*GNA14*
10	72815798	rs4747150	Alcohol 30 Day Frequency	9.41x10^{-6}	
12	96988908	rs11109339	Former Smoker Quantity 1Day A	1.22x10^{-5}	
13	59162465	rs1379518	Former Smoker Quantity 1Day B	0.0011	*DIAPH3*
14	36400549	rs1325530	Weekend Sun Hours Last Decade	1.10x10^{-5}	*SLC25A21*
15	20596568	rs3812923	Cigarette Smoking Current	7.53x10^{-6}	*NIPA1*
15	22652931	rs752873	Alcohol 30 Day Frequency	6.23x10^{-6}	*SNRPN*
15	48790334	rs10519284	Someday Smoker Days 1 Month	0.018	*SPPL2A*
15	48796548	rs12898588	Someday Smoker Days 1 Month	0.018	*SPPL2A*
15	48797199	rs7165492	Someday Smoker Days 1 Month	0.018	*SPPL2A*
15	59933890	rs17238096	Someday Smoker Quantity 1 Day	0.022	*VPS13C*
16	50786246	rs2245948	Former Smoker Quantity 1 Day B	0.0019	
16	50794864	rs7200614	Former Smoker Quantity 1 Day B	0.0019	
16	64857219	rs233546	Someday Smoker Days 1 Month	0.018	
16	76582490	rs6564494	Former Smoker Quantity 1 Day B	0.0021	
16	83259294	rs4783043	Former Smoker Quantity 1 Day B	0.0020	
17	2685811	rs9747501	Everyday Smoker Quantity 1 Day	0.00055	*GARNL4*
17	40022727	rs16970865	Someday Smoker Quantity 1 Day	0.021	
17	40035928	rs9904409	Someday Smoker Quantity 1 Day	0.020	
17	40057582	rs10451262	Someday Smoker Quantity 1 Day	0.019	
18	74848306	rs612829	Former Smoker Smoking 6 Month	3.52x10^{-6}	*SALL3*
19	61733570	rs2058131	Former Smoker Smoking 6 Month	2.72x10^{-7}	
20	45927980	rs8121494	Cigarette Smoking Current	6.86x10^{-6}	

21	33511140	rs9978523	Everyday Smoker Quantity 1 Day	0.00059	
22	22564245	rs738807	Weekend Sun Hours Last Decade	5.14x10^{-6}	
22	22564493	rs5751759	Weekend Sun Hours Last Decade	2.79x10^{-6}	
22	22565198	rs4822443	Weekend Sun Hours Last Decade	3.21x10^{-6}	*MIF*
22	22567862	rs2000466	Weekend Sun Hours Last Decade	4.33x10^{-6}	*MIF*
22	26292517	rs723184	Cigarette Smoking 100	1.63x10^{-5}	
22	26298208	rs5762257	Cigarette Smoking 100	1.57x10^{-5}	
22	32107269	rs5998902	Someday Smoker Quantity 1 Day	0.022	*LARGE*

Table abbreviations:
Chr = Chromosome
BP = Base pair location of SNP
RSID = SNP ID
P-Value = P-value of the interaction
Gene = Gene symbol of gene the SNP is within (blank if not within a gene)

3.2. *GxG Results*

The top Biofilter 1.0 derived GxG models are presented in Table 5. A total of 5 models had both a p-value for the interaction term and the overall model p-value < 0.05. A total of 7 genes were in the five models. Of the five models, the most significant was for a model with *EGF*, which is epidermal growth factor, and *EGFR*, which codes for the cell surface receptor that binds to epidermal growth factor.

Table 5: The 5 SNPxSNP models with an interaction p-value < 0.05 and a total model p-value < 0.05 after association testing of the Biofilter derived pairwise models. Presented in Table 5 are the effect coefficients for the main effects (β_1 and β_2) and the interaction term (β_3), as well as the P-value for the interaction term (P_{ixn}) and the P-value for the model (P_{mod}).

Gene 1	SNP1	Allele /MAF	Gene 2	SNP2	Allele /MAF	β_1	β_2	β_3	P_{ixn}	P_{mod}
EGF	rs2298999	0.41	EGFR	rs17172446	0.23	0.039	0.087	-0.086	3.37x10^{-6}	3.93x10^{-5}
LCP2	rs2338872	0.40	VAV1	rs12979659	0.37	0.036	0.058	-0.072	8.28x10^{-6}	6.18x10^{-5}
FYN	rs1327200	0.34	DOCK1	rs10829597	0.40	-0.084	-0.049	0.067	4.53x10^{-5}	2.56x10^{-5}
FYN	rs1327200	0.34	DOCK1	rs2050305	0.37	-0.080	-0.046	0.066	7.56x10^{-5}	4.24x10^{-5}
DOCK1	rs9418709	0.47	SRC	rs6063022	0.15	0.019	0.061	-0.083	5.74x10^{-5}	0.00036

4. Discussion

The results presented herein are an exploration of the use of multiple novel approaches for investigating gene and phenotype associations within EHR based data. We performed an analysis with PhenX derived measures, seeking GxE interaction models for the Marshfield Cataract data set. The majority of the significant interactions were found for smoking related measures. We did find some highly correlated PhenX measures with significant interactions for SNPs within similar regions, such as the results on chromosome 6 for SNPs rs6899720 and rs12528760, for smoking related phenotypes. Through searches in the NCBI catalog [22], as well as the National Center for Biotechnology (NCBI) dbSNP [19], these two SNPs, as well the SNP in our most significant GxE model, did not show previous GWA level significant associations for any phenotypes.

We also performed an exploratory analysis with Biofilter 1.0, an updated and improved implementation of the originally published Biofilter. The results are intriguing, and provide the basis for hypotheses that can be investigated further, highlighting how Biofilter results have a biological context that provide additional information for resulting models. For instance, the most significant model contained *EGF* and *EGFR*, which are known to have a biological interaction as the gene product of *EGFR* is a receptor for the gene product of *EGF*. Epidermal growth factor is found in human tears [23], and ocular effects have been found after the administration of EGFR inhibitors administered to patients [24]. Interestingly, three of the models that passed our significance cutoff contained two of the same genes, *FYN,* a member of the protein-tyrosine kinase oncogene family implicated in cell growth, and *DOCK1,* dedicator of cytokinesis 1. These models as a whole implicate genes related to cell growth, the cell cycle, and epidermal growth.

We are currently developing Biofilter 2.0 which will be include additional database sources and allow for the use of other position and region based information beyond SNPs and genes, such as copy number variation (CNV) data, evolutionary conserved regions, and regulatory regions, allowing users to incorporate additional sources of prior knowledge as well as utilize other sources of genetic variation measurement data, with a more user-friendly interface.

Our results provide more complex models for an association between genetic variation and cataract outcome, moving beyond the more standard SNP-phenotype associations. The models found we intend to investigate further and warrant additional investigation of the environment and genetic variables contributing to these more complex models. These bioinformatics approaches can be used with other datasets, to expand the investigation of the relationship between genetic architecture and phenotypic outcome. With these approaches that consider the complexity of the data and harness the power of novel bioinformatics tools, we will elucidate the missing heritability of complex traits.

Acknowledgments

This work was supported by the following grants: U19 HL0659625, R01 LM010040, U01 HG006389

References

1. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, et al. (2012) Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. Journal of the American Medical Informatics Association : JAMIA 19: 225-234.
2. Waudby CJ, Berg RL, Linneman JG, Rasmussen LV, Peissig PL, et al. (2011) Cataract research using electronic health records. BMC Ophthalmol 11: 32.
3. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD (2005) Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. Personalized Medicine 2: 49-79.
4. Pathak J, Pan H, Wang J, Kashyap S, Schad PA, et al. (2011) Evaluating Phenotypic Data Elements for Genetics and Epidemiological Research: Experiences from the eMERGE and PhenX Network Projects. AMIA Summits Transl Sci Proc 2011: 41-45.
5. Michael R, Bron AJ (2011) The ageing lens and cataract: a model of normal and pathological ageing. Philos Trans R Soc Lond B Biol Sci 366: 1278-1292.

6. Hammond CJ, Duncan DD, Snieder H, de Lange M, West SK, et al. (2001) The heritability of age-related cortical cataract: the twin eye study. Invest Ophthalmol Vis Sci 42: 601-605.
7. Bush WS, Dudek SM, Ritchie MD (2009) Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput: 368-379.
8. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29-34.
9. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, et al. (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37: D619-622.
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 25: 25-29.
11. Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28: 405-420.
12. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, et al. (2010) NetPath: a public resource of curated signal transduction pathways. Genome Biol 11: R3.
13. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 39: D698-704.
14. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, et al. (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40: D857-861.
15. Stover PJ, Harlan WR, Hammond JA, Hendershot T, Hamilton CM (2010) PhenX: a toolkit for interdisciplinary genetics research. Curr Opin Lipidol 21: 136-140.
16. Abraham AG, Condon NG, West Gower E (2006) The new epidemiology of cataract. Ophthalmol Clin North Am 19: 415-425.
17. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, et al. (2011) Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet Chapter 1: Unit1 19.
18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81: 559-575.
19. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311.
20. Grady BJ, Torstenson E, Dudek SM, Giles J, Sexton D, et al. (2010) Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. Pac Symp Biocomput: 315-326.
21. Wickham H (2009) ggplot2: elegant graphics for data analysis: Springer New York.
22. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106: 9362-9367.
23. Ohashi Y, Motokura M, Kinoshita Y, Mano T, Watanabe H, et al. (1989) Presence of epidermal growth factor in human tears. Invest Ophthalmol Vis Sci 30: 1879-1882.
24. Saint-Jean A, Sainz de la Maza M, Morral M, Torras J, Quintana R, et al. (2012) Ocular Adverse Events of Systemic Inhibitors of the Epidermal Growth Factor Receptor: Report of 5 Cases. Ophthalmology.

INTERPRETING PERSONAL TRANSCRIPTOMES: PERSONALIZED MECHANISM-SCALE PROFILING OF RNA-SEQ DATA

ALAN PEREZ-RATHKE[†]

Department of Medicine, University of Illinois at Chicago
Chicago, IL 60612, USA
Email: perezrat@uic.edu

HAIQUAN LI[†]

Department of Medicine, University of Illinois at Chicago
Chicago, IL 60612, USA
Email: haiquan@uic.edu

YVES A. LUSSIER*,[†]

Departments of Medicine & Bioengineering, University of Illinois at Chicago
Chicago, IL 60612, USA
Email: ylussier@uic.edu

Despite thousands of reported studies unveiling gene-level signatures for complex diseases, few of these techniques work at the single-sample level with explicit underpinning of biological mechanisms. This presents both a critical dilemma in the field of personalized medicine as well as a plethora of opportunities for analysis of RNA-seq data. In this study, we hypothesize that the "Functional Analysis of Individual Microarray Expression" (FAIME) method we developed could be smoothly extended to RNA-seq data and unveil intrinsic underlying mechanism signatures across different scales of biological data for the same complex disease. Using publicly available RNA-seq data for gastric cancer, we confirmed the effectiveness of this method (i) to translate each sample transcriptome to pathway-scale scores, (ii) to predict deregulated pathways in gastric cancer against gold standards (FDR<5%, Precision=75%, Recall =92%), and (iii) to predict phenotypes in an independent dataset and expression platform (RNA-seq vs microarrays, Fisher Exact Test $p<10^{-6}$). Measuring at a single-sample level, FAIME could differentiate cancer samples from normal ones; furthermore, it achieved comparative performance in identifying differentially expressed pathways as compared to state-of-the-art cross-sample methods. These results motivate future work on mechanism-level biomarker discovery predictive of diagnoses, treatment, and therapy.

[†] This work was supported in part by the Cancer Research Foundation, the Medical Scientist Training Program of the University of Illinois (2T32 GM079086-06) and by the University of Illinois Cancer Center.

* Corresponding author

1. Introduction

Interpreting differentially expressed genes at the biological scale using enrichment statistics (Enrichment) or Gene-Set Analyses (GSA) has become routine for microarray and RNA-Seq studies. By design, these analyses require group assignment as well as derived mechanisms (e.g., Kyoto Encyclopedia of Genes and Genomes, i.e. KEGG pathways [1]) to reference differences of expression between these groups. While biologists are well served with such studies, evaluating individual patients in clinic necessitates single patient measures. Indeed, conventional single molecule biomarkers are popular because of their crisp thresholds that are interpretable as normal or abnormal. FDA-approved biomarkers are often required to reveal clinically interpretable biological mechanistic information useful in diagnosis of disease and prognosis of therapeutic response. While gene expression classifiers (signatures) have been shown as accurate predictors, they paradoxically are not comprised of “driver genes” (known mechanisms of diseases) or therapeutic response [2]. When developed using different datasets, there is poor genetic concordance between signatures. In contrast, we have shown mechanistic overlap at the protein interaction level between signatures predictive of clinical outcome in breast cancer [3] and in prostate cancer [4]. The lack of mechanistic underpinning prohibits in part the wide adoption and FDA approval of expression classifiers [5]. Indeed, MammaPrint® microarray [6] and of OncotypeDX [7] are both classifiers derived from mechanisms (wound healing signature from animal models, and curated breast cancer driver genes,).

Few genome-wide methods have been developed using gene-sets for imputing biological mechanisms (most have been for microarrays measuring RNA expression). In these studies, scoring mechanisms by the median or mean expression of their corresponding gene-set were shown to be capable of generating classifiers but at a lower accuracy than single-transcript RNA expression-level signatures [8, 9]. More accurate mechanism classifiers can be derived from methods comparing phenotypic group assignments between samples to identify principal components (PCA) [10, 11] or by the expression of key genes to represent the whole pathway such as in CORG [12] and LLR [13]. We developed “Functional Analysis of Individual Microarray Expression” (FAIME), a weighted rank method that can impute mechanism-scores on each expression array sample and eliminate the group assignment requirement [14]. We have shown FAIME’s accuracy in generating classifiers predictive of outcome in independent expression array datasets of head and neck [14] and lung cancers [15]. We have also experimentally validated FAIME for predicting microRNA targets within cell lines and animal models [16]. We have additionally demonstrated that while the genetic overlap of RNA-level classifiers across three head and neck cancer datasets was ~3% at False Discovery Rate (FDR) <5%, more than 46%-61% of the FAIME-anchored KEGG pathways classifiers overlapped in the same datasets (FDR<5%) [14]. We have also demonstrated that FAIME can be employed on continuous phenotypes such as survival in cox-regression [12]. These studies [10-14] transcend those using conventional gene enrichment or gene set enrichment analyses (GSEA) that cannot provide individual measurements of mechanisms on a single sample and require comparison between multiple samples groups (in distinct categorical phenotypes) to infer gene-set-level predictions. Recently, related work in mass spectrometry protein complexes (derived from interaction networks) were shown to be more accurate for designing classifiers than single proteins [17]. However, to our knowledge, no mechanism-level methodology has yet been designed specifically for interpreting individual RNA-sequencing samples. Such a methodology is a requirement to develop RNA-seq based, clinically predictive mechanism-level classifiers. To our knowledge, no method of mechanism imputation has been developed for RNA-seq at the single sample level.

We hypothesized that the FAIME weighted rank-based method we developed for expression arrays would be more accurate than the simpler 'median expression' and 'mean expression' methods. To confirm this for each method, we systematically compared the different false discovery rate thresholds for accuracy and for biological reproducibility across transcriptomic measurements using (i) proxy gold standards in the same datasets and (ii) validating in independent datasets (RNA-Seq vs array expression).

2. Methods

2.1. *Data preparation and databases*

All datasets were obtained from the Gene Expression Omnibus (GEO) [18]. To demonstrate the feasibility of the FAIME technique on RNA-seq data, the Asian gastric cancer dataset GSE36968 [19], consisting of 24 gastric cancers and 6 normal stomach samples, was used. GSE36968 was sequenced with Life Technologies SOLiD™ sequencing platform. This dataset was already in Reads Per Kilobase of exon model per Million mapped reads (RPKM) format [20]. Since RPKM is a widely accepted standard for RNA-seq normalization by biologists, no additional pre-processing was performed. To validate and show concordance among RNA-seq and microarray data, the Asian gastric cancer microarray dataset GSE13861 [21], consisting of 71 gastric cancer and 19 normal samples, was used. This dataset was already quantile normalized [22] and $\log_2$ transformed.

2.2. *Microarray platform annotation*

Microarray platform annotation was downloaded from the GEO website (http://www.ncbi.nlm.nih.gov/geo/) for the GSE13861 dataset using Illumina HumanWG-6 v3.0 expression beadchip.

2.3. *KEGG pathway annotations*

KEGG pathway annotations are embedded in Bioconductor database KEGG.db [23] version 2.7.1. The 229 KEGG pathways with more than 3 annotated genes are studied.

2.4. *FAIME pathway scoring of each sample*

From the methodologies in [1], to quantitatively assign a mechanism's "expression deregulation" via its gene members, whose expression is measured in RPKM, all expressed genes (set G) in each sample are sorted in a descending order according to their expression levels, and then, as shown in Eq. (1), an exponential decreasing weight (w) is assigned to the ordered genes. The resultant weighted expression values are used to prioritize relatively highly expressed genes as in the first step of Bioconductor package *OrderedList* [24, 25]. Specifically, let $r_{g,s}$ be the expression rank for each gene $g \in G$ in a sample s, let $|G|$ be the total number of distinct genes measured and the weight assigned to each gene per sample ($w_{g,s}$) is calculated as follows:

$$w_{g,s} = (r_{g,s}) \cdot (e^{\frac{r_{g,s} - |G|}{|G|}}) \tag{1}$$

A Normalized Centroid (*NC*) is defined as the uni-dimensional average of the weighted expression values of a gene-set. Specifically, the sum of the weighted expression of gene element in a gene-set is normalized according to its cardinality. For every KEGG pathway, there is a gene-set *KEGGi* in which genes satisfy $g \in KEGG_i$ and a complement gene-set ($G/KEGG_i$) comprised of all available measured genes that are not annotated to this KEGG pathway. Thus we calculate the normalized centroid of each gene-set $KEGG_i$ in each sample s and that of its complement gene-set as follows:

$$NC(KEGG_{i,s}) = \frac{1}{|KEGG_i|} \sum_{g \in KEGG_i} (w_{g,s}) \tag{2a}$$

$$NC(G/KEGG_{i,s}) = \frac{1}{|G/KEGG_i|} \sum_{g \in G/KEGG_i} (w_{g,s}) \tag{2b}$$

$$where \quad G/KEGG_i = \{g : g \notin KEGG_i \cap g \in G\}$$

Furthermore, Eq. (3) calculates the Functional FAIME Score (*F* in equations) of each gene-set of a KEGG pathway in every sample as the difference between the normalized centroid of its gene-set and that of its complement gene-set. We define functional scores as functional biological mechanisms of the gene-set associated with a KEGG pathway in a given example.

$$F_{KEGG_{i,s}} = F(KEGG_{i,s}) = NC(KEGG_{i,s}) - NC(G/KEGG_{i,s}) \tag{3}$$

Eq. (4) calculates for a sample s, the FAIME Profile "FP_s" defined as the set of all FAIME scores of sample s, $F_{KEGGi,s}$, assigned to every term.

$$FP_s = \{F_{KEGG_{1,s}}, \ldots, F_{KEGG_{i,s}}, \ldots, F_{KEGG_{n,s}}\} \tag{4}$$

where n is the total number of KEGG pathways.

In this way, patient-specific FAIME profiles of KEGG pathways are generated for each sample. Each sample has a continuous effective value for each category term which is the group difference between the genes annotated by the KEGG pathway and their individual complementary set of genes [16].

Calculations were performed using the latest FAIME R package which has been imrpoved to compute scores concurrently and allow for custom transformations (available: https://bitbucket.org/lussierlab/faime-opensource). Experiments were made with alternate transformations such as uniform-weighted rank and median selection, but we found that the original methodology performed the most consistently.

2.5. *Simpler methods for scoring each sample pathways*

To evaluate FAIME against alternative single-sample pathway scoring methods, we defined two unranked and two ranked methods. The unranked methods, *RPKM mean* and *RPKM median*, compute a sample's pathway score as either the mean or median of the RPKM values of the pathway's gene set respectively. Analogous rank-based methods, *Mean of Ranked RPKM* and *Median of Ranked RPKM*, first convert a sample's RPKM values into ranks and then score each pathway as the mean or median of the constitutive ranks respectively.

2.6. *Unsupervised hierarchical clustering (Figure 1)*

As seen in **Figure 1**, FAIME scores for all 229 KEGG pathways were used in generating the unsupervised hierarchical clustering of RNA-Seq dataset GSE36968. Similarly, other ranked methods (RPKM mean, RPKM median, mean of ranked RPKM and median of ranked RPKM) were employed for clustering as comparison. The clustered heat map was generated using the *heatmap* function of R with Ward's method as the distance criterion.

2.7. *Predicting deregulated pathways between two sets of samples using Wilcoxon parametric test (Figure 2&3, Table 1)*

In sections 2.4 and 2.5, we have described five methods (FAIME, RPKM mean, etc) that transform genome-wide RNA-seq or microarray-level measures of expression of a sample into pathway scores for this sample. Comparing samples of gastric cancer to normal gastric tissue, we calculate the deregulated pathways using the non-parametric Wilcoxon statistic and adjust for multiple comparisons using FDR. Thus, a set of deregulated pathways at different FDR thresholds can be imputed form the same dataset for each pathway scoring method. These can be compared to methods that calculate deregulated pathways directly from the gene-level expression such as GSEA and Enrichment studies (See section 2.8, ROC: Receiver Operating Characteristic).

2.8. *Evaluating pathway-scoring methods using ROC curves and proxy gold standards operating on the same RNA-seq dataset (Figure 2)*

Since it is unfeasible to biologically validate all predicted KEGG pathways, accuracy was determined using alternatively (i) GSEA [26] or (ii) conventional enrichment of differentially expressed genes (R package for SAM [27] analysis at FDR<5%) as proxy gold standards. At a given FDR, the set $positives_{GSEA}$ was calculated as the set of KEGG pathways found significantly differentially scored between cancer versus normal under GSEA (gene-set permutation); the set $positives_{FAIME}$ was calculated as the set of KEGG pathways found significantly differentially scored between cancer versus normal by running SAM [27] on the FAIME scores of each sample (Wilcoxon-statistic); the set $positives_{Enrichment}$ was calculated by first using SAM to identify significantly differentially expressed genes (Wilcoxon-statistic, fixed gene level FDR < 5%) and then performing hypergeometric enrichment on those genes for the KEGG pathways at the given FDR cutoff for pathways. Using GSEA as a proxy gold standard (**Figure 2, Panel A&B**), $positives_{GSEA}$

was fixed at FDR < 25% as recommended by the authors. Then, at various maximum FDRs ranging from 0% to 35%, the set of *true positives* for FAIME was calculated as $positives_{GSEA} \cap positives_{FAIME}$, the set of *false positives* as the set difference $positives_{FAIME} - positives_{GSEA}$, the set of *false negatives* as the set difference $positives_{GSEA} - positives_{FAIME}$, and the set of *true negatives* as the set difference $KEGG_{ALL}$ - (*true positives* ∪ *false positives* ∪ *false negatives*). With these values, we could then create a receiver-operating characteristic (ROC) curve for FAIME by plotting the *true positive rate* according to Eq. (A.1), versus the *false positive rate* according to Eq. (A.2). To compare with FAIME, a similar procedure was used to create the ROC curve for hypergeometric enrichment **(Figure 2, Panels C&D)**. To allow comparison of GSEA and FAIME, hypergeometric enrichment at FDR < 5% was instead used as a proxy gold standard and the corresponding ROC curves were created.

2.9. *Evaluating pathway-scoring methods in an independent dataset using concordance of prediction (Table1) and clustering (Figure 3)*

For each of the five pathway-scoring methods (see 2.4-2.6; FAIME, RKPM mean, etc), the R package for SAM [27] was successively used to prioritize pathways deregulated between gastric tumors and normal gastric tissue at FDR<2.5% and at FDR<5% in RNA-seq dataset GSE36968. The corresponding FAIME scores of those pathways in independent microarray dataset GSE13861 were then used as the basis for hierarchical clustering in **Figure 3 (**R's *heatmap* function with Ward's method as the distance criterion). Similarly, differentially expressed pathways imputed from dataset GSE13861 at FDR 2.5% and 5% were used to hierarchically clustering samples in RNA-seq dataset GSE36968 and reported in **Table 1**. Furthermore, these analyses were successively conducted on the four other pathway-scoring methods: RPKM mean, RPKM median, mean of ranked RPKM, and median of ranked RPKM. The reciprocal study was conducted as well: prioritizing pathways for each method in the microarray studies and clustering the RNA-seq samples using the pathway scores of each RNA-seq sample corresponding to those prioritized pathways. Clustering accuracies of each method are reported in **Figure 3.** Further, an additional evaluation was conducted: the Fisher Exact Test (FET) and odds ratio of the concordance between the prioritized pathways derived independently over microarrays and RNA-seq are reported in **Table 1**.

3. Results and Discussion

To our knowledge, we present the first study of mechanism imputation at the single sample level for RNA-seq. This experiment differs from our previous ones in that we systematically also include as control intermediate geneset methods of computations such as mean, median, etc. In order to evaluate the feasibility to impute valid pathway scores at the individual sample level, we evaluated five distinct pathway-scoring methods in each of the following four experiments: (i) clinical phenotype clustering of individual RNA-seq samples by their pathways scores, (ii) concordance between pathways predicted at the RNA-seq single-sample level against those predicted at the cohort-wide level (such as in GSEA), (iii) the predictive power of prioritized pathways in one dataset as classification

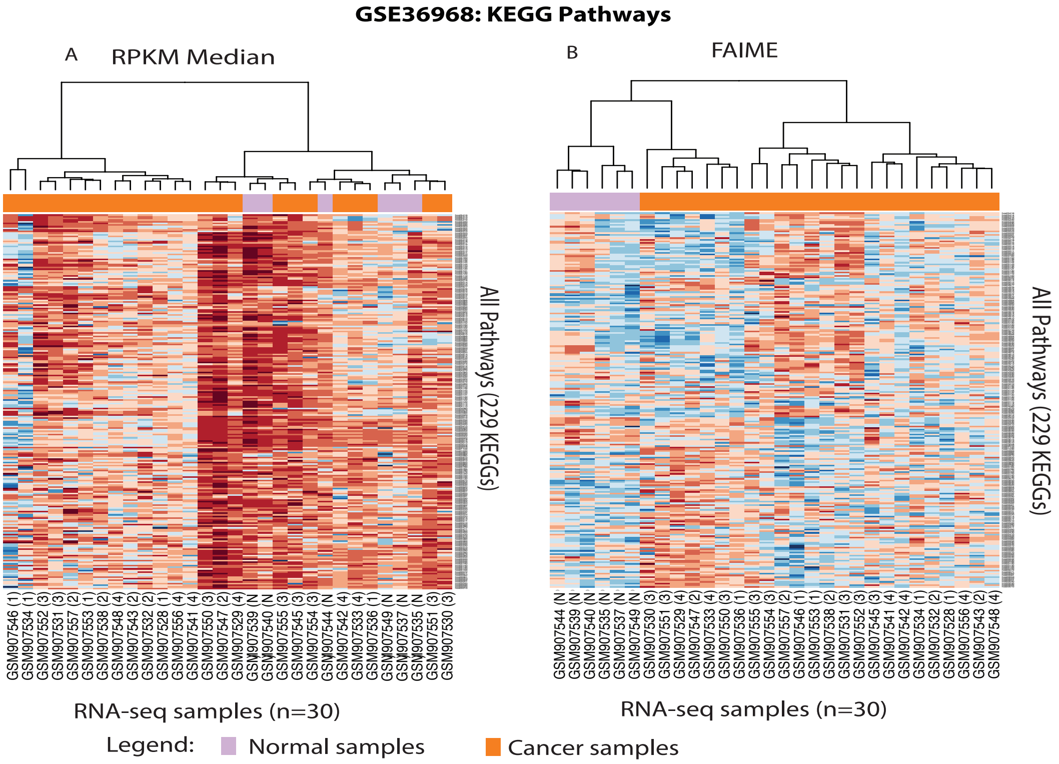


Figure 1. Unsupervised hierarchical clustering of all KEGG pathway-level scores imputed from RNA-seq RPKM of individual samples. ***Panel A:*** clustering RNA-seq dataset GSE36968 by "RPKM median" measure on individual sample ("RPKM means" - not shown - is equally inaccurate). ***Panel B:*** clustering RNA-seq dataset GSE36968 by FAIME scores imputed from individual samples (every other ranked-based method provided equally good clustering, not shown). This illustrates the pathway level clustering possible with pathway scoring at the single sample level (note: GSEA and Enrichment are not designed for this purpose). **Legend:** up-regulated pathways in cancer are blue and down-regulated ones are red. columns=30 samples; rows=229 KEGG pathways (formatted for reading at high magnification).

features for another dataset, and (iv) the concordance between pathway predictions conducted in two independent datasets (**Figure 1, Figure 2, Figure 3** and **Table 1,** respectively). In **Figure 1 Panel B**, FAIME scores for the entire KEGG ontology (229 pathways[1] were used to perform unsupervised hierarchical clustering. Using Ward's method [28] as the distance criterion, all normal samples were found within the same cluster, as were gastric cancer samples in RNA-seq dataset GSE36968. Other rank-based methods (mean of ranked RPKM, median of ranked RPKM) achieved similar clustering results but unranked methods (RPKM mean, RPKM median) failed to cluster accurately (**Figure 1, Panel A**). Note that cross-sample methods GSEA and Enrichment cannot work on single-sample level. Top panels in **Figure 2** demonstrate ROC curves for the KEGG pathways using GSE [26]as the proxy gold standard. For up-regulated pathways (**Figure 2 Panel A**), FAIME ROC performance compares favorably to hypergeometric enrichment. For down-regulated pathways (**Figure 2 Panel B**), FAIME and hypergeometric enrichment performed similarly. Bottom panels in **Figure 2** demonstrate ROC curves for the KEGG pathways using hypergeometric enrichment as the proxy gold standard. For both up-regulated (left) and down-regulated (right) pathways, FAIME ROC as the proxy gold

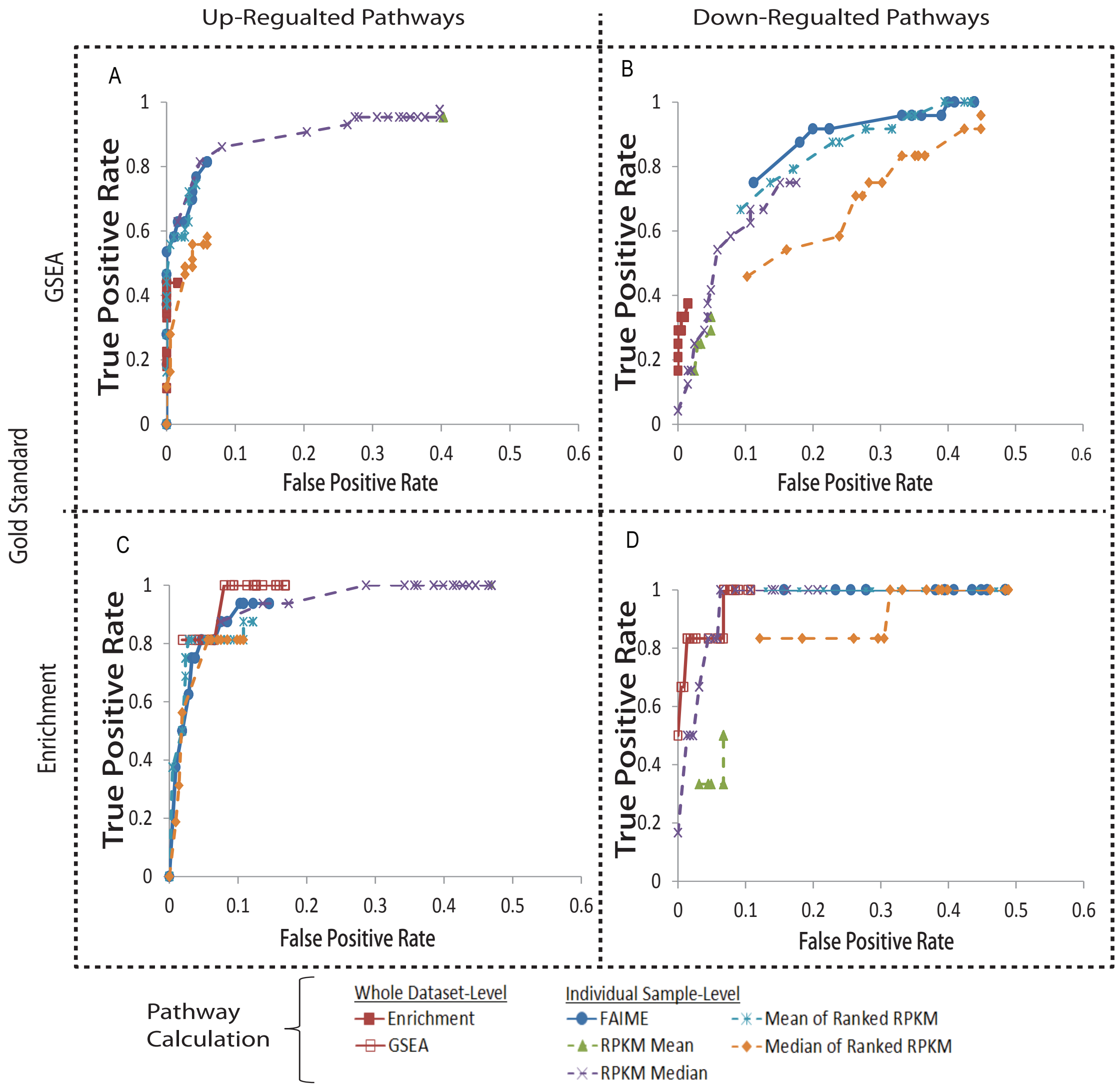


Figure 2: ROC curves of FAIME methods in identifying differentially expressed pathways as compared to GSEA, Enrichment, RPKM mean, RPKM median, mean of ranked RPKM and median of ranked RPKM. Panel A and B: ROC curves using differentially expressed pathways of GSEA as a proxy gold standard (FDR<25%). Panel C and D: ROC curves using differentially expressed pathways by Enrichment as a proxy gold standard (FDR<5%). Up- and down- regulated pathways vary at each accuracy threshold for each method and calculated is available at: http://lussierlab.org/publications/FAIME-rnaseq.

standard. For up-regulated pathways (**Figure 2 Panel A**), FAIME ROC performance is comparable to GSEA. We also compared the FAIME ROC performance with simpler, single-sample measures such as RPKM mean, RPMK median, mean of ranked RPKM and median of ranked RPKM (dashed lines) for both down-regulated pathways and up-regulated pathways, using either GSEA or enrichment method as benchmark. FAIME yields either superior or similar ROC performance as compared to these single-sample methods. The exception is the RPKM median method which surpasses ranked methods as well as RPKM mean.

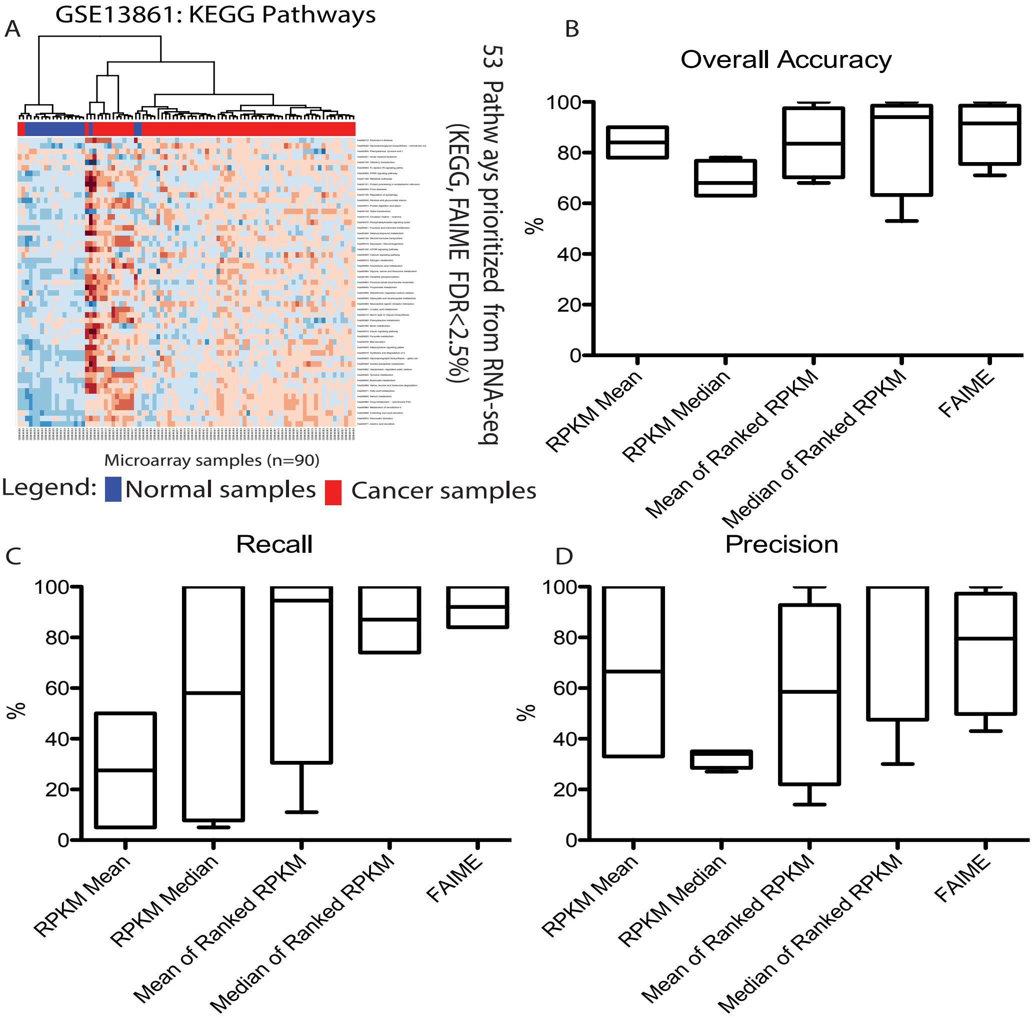


Figure 3: Unsupervised hierarchical clustering of gastric cancer datasets using sample-level scores of differentially expressed KEGG pathways learned from another independent dataset. Panel A: Clustering of microarray dataset GSE13861 by 53 significant different expressed FAIME pathways (FDR<0.025) learned from GSE36968 (large figure at http://lussierlab.org/publications/FAIME-rnaseq). **Panel B, C, D:** As described in **Methods (Section 2.6)**, deregulated pathways were prioritized in one dataset and their classification accuracies evaluated in an independent one (and vice-versa) producing accuracy scores reported here. Rank-based methods (mean of ranked RPKM, median of ranked RPKM, and FAIME) achieved overall better predictive performance across datasets as compared to unranked mean and median methods.

Figure 3a demonstrates hierarchical clustering of microarray dataset GSE13861 with 53 significantly differentially expressed FAIME features (FDR < 0.025) found in RNA- seq dataset GSE36968. 84 out of 90 (93.3%) samples are classified correctly. In a second set of experiments, reciprocal clustering of RNA-seq dataset GSE36968 using 122 and 140 differently expressed FAIME pathway features of microarray dataset GSE13861 (FDR <0.025 and FDR < 0.05 respectively). The overall accuracy, precision, and recall are shown in **Figures 3b**, **3d** and **3c** respectively. As shown from the three panels, RPKM median and RPKM mean methods achieved the worst results as compared to rank-basedmethods (mean of ranked RPKM, median or ranked RPKM, and FAIME).

Table 1. Pathway prediction concordance between the independent RNA-seq and microarrays datasets for each pathway-scoring method (Sub-table A). **Sub-table B** shows the stringent concordant subset of deregulated pathways prioritized by three techniques in both dataset (intersection): GSEA, Enrichment and FAIME that respectively predicted 29, 10 and 12 upregulated KEGG pathways and 21, 31 and 46 downregulated ones. Pathways known involved in gastric cancer are highlighted in blue (e.g. gemcitabine[5-FU], a pyrimidine analog, is a standard combination in treatment of gastric cancer). Detailed at: http://lussierlab.org/publications/FAIME-rnaseq

Sub-table A		↓*pathways**		↑*pathways**	
	Method	**odds ratio**	**FET pvalue**	**odds ratio**	**FET pvalue**
Dataset-wide metrics	GSEA	40	$5x10^{-13}$	60	$<2x10^{-16}$
	Enrichment	14	$6x10^{-3}$	106	$2x10^{-11}$
Single Sample Metrics	Mean RPKM	12	$2x10^{-4}$	17	$1x10^{-13}$
	Median RPKM	14	$5x10^{-3}$	14	$4x10^{-14}$
	Mean of ranked RPKM	15	$4x10^{-15}$	*no overlap*	
	Median of Ranked RPKM	6	$2x10^{-8}$	*no overlap*	
	FAIME	19	$<2x10^{-16}$	∞	$1x10^{-6}$

Type	KEGG ID	Sub-table B **KEGG Pathways**
↑	04110	Cell cycle
↑	04115	P53 signaling pathway
↑	00240	Pyrimidine metabolism
↑	03040	Spliceosome
↑	03013	RNA transport
↑	03008	Ribosome biogenesis in eukaryotes
↓	00982	Drug metabolism – cytochrome P450
↓	00980	Metabolism of xenobiotics by cytochrome P450

Legend: * ↓↑respectively down- and up- regulated pathways in gastric cancer;
Odds ratio from the intersection between RNA-seq & array predictions; ∞: infinite (division by zero).

Evaluations conducted on the same dataset with proxy gold standards demonstrated that each method could produce modest to good accuracies - with the RPKM-mean method dominating. Paradoxically, the RPKM-mean was the worst method in term of recall and modest in terms of precision. This demonstrates that RPKM-mean is a volatile metric. In addition, the rank-based methods failed to identify up-regulated pathways in either GSE13861 or GSE36968 (Table 1). The FAIME method (which is a weighted rank-based method) achieved the most overall stable performance in reflecting the uniform underlying mechanisms across distinct types of datasets of the same gastric cancer diseases.

3.1. Future Studies and Limitations

While many studies have been completed in large RNA-seq datasets – they largely remain unavailable (either embargoed or simply not deposited in GEO). We are completing additional studies to corroborate the findings of this report in (i) other cancers, (ii) other diseases, and (iii) for predicting response to therapy. Identifying key genes in each pathway would merit to be evaluated in RNA-seq as well (e.g. CORG, [12]). Finally, other type of gene-sets beyond KEGG pathways and curated pathways should be considered. Co-expression modules derived from large scale studies of multiple disease conditions have provided insight in new biology and could be utilized as non-curated gene-sets. Protein complexes, that worked well in mass spectrometry [17], could also be utilized as gene-sets for pathway discovery in RNA-seq.

Further, we are exploring other pathway scoring approaches at the single-sample level that would conserve the inherent vectorial structure of pathway expression, without the requirement of cross-sample analyses. We are also evaluating FAIME in a prospective clinical trial in predicting therapeutic response to recurrent head and neck cancer.

Additionally, FAIME exploits an exponential transformation algorithm that weights better highly expressed genes and thus rectifies (i) the saturation of microarray probes at high dynamic range and (ii) the high relative and absolute error rate (noise) on low expression measurements. Only the latter bias remains salient for RNA-seq. However, RPKM may not be the optimal metric for correcting biases of oversampling longer gene in next-gen seq. Moreover, most RNA-seq datasets are measured after reverse transcription on DNA-seq platforms, adding another potentially biased step to model. Thus, improving on mechanism-scoring methods for requires integrating modeling of new biases of specific RNA-seq platforms (e.g. adjustments for RNA fragment length that vary between platforms, gene length biases, reverse transcription, etc).

4. Conclusion

To demonstrate the feasibility of single-sample classification, we performed an entirely unsupervised hierarchical clustering of RNA-seq dataset GSE36968. This clustering does not rely on differentially expressed features found by a tool requiring multiple samples such as SAM [27] or GSEA. Instead, the FAIME scores for all KEGG pathways are used. Figure 1 demonstrates the success of this approach with 100% of normal samples being contained within the same cluster.

Accurate pathway-scoring techniques could conceivably be used as a single sample analysis mechanism whereby clinicians could establish a patient's pathway profile [14] as a diagnostic and prognostic utility. Identifying pathways with exceptionally high or low scores could also serve as a means to elucidate individualized drug targets. This could then allow for a personalized drug regimen based on transcriptomic analysis. However as shown with Mammaprint® and OncotypeDX®, the technologies adoption is complex and requires more than technical prowess.

Software availability

We provide a package allowing for high-throughput analyses of the five studied pathway-scoring methods on individual samples (https://bitbucket.org/lussierlab/faime-opensource).

Appendix

The *true positive rates* and *false positive rates* used in the ROC plots for FAIME, GSEA, and hypergeometric enrichment were calculated as follows:

$$\textit{true positive rate} = \frac{|\,\textit{true positives}\,|}{|\,\textit{true positives} \cup \textit{false negatives}\,|} \tag{A.1}$$

$$\textit{false positive rate} = \frac{|\,\textit{false positives}\,|}{|\,\textit{false positives} \cup \textit{true negatives}\,|} \tag{A.2}$$

References

1. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 1999. **27**(1): p. 29-34.
2. Massagué, J., *Sorting Out Breast-Cancer Gene Signatures.* New England Journal of Medicine, 2007. **356**(3): p. 294-297.

3. Chen, J., et al., *Protein interaction network underpins concordant prognosis among heterogeneous breast cancer signatures.* Journal of Biomedical Informatics, 2010. **43**(3): p. 385-396.
4. Chen, J.L., et al., *Protein-network modeling of prostate cancer gene signatures reveals essential pathways in disease recurrence.* Journal of the American Medical Informatics Association, 2011. **18**(4): p. 392-402.
5. Kulasingam, V., M.P. Pavlou, and E.P. Diamandis, *Integrating high-throughput technologies in the quest for effective biomarkers for ovarian cancer.* Nat Rev Cancer, 2010. **10**(5): p. 371-378.
6. Knauer, M., et al., *The predictive value of the 70-gene signature for adjuvant chemotherapy in early breast cancer.* Breast Cancer Research and Treatment, 2010. **120**(3): p. 655-661.
7. Paik, S., et al., *A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer.* New England Journal of Medicine, 2004. **351**(27): p. 2817-2826.
8. Guo, Z., et al., *Towards precise classification of cancers based on robust gene functional expression profiles.* BMC Bioinformatics, 2005. **6**(1): p. 58.
9. Abraham, G., et al., *Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context.* BMC Bioinformatics, 2010. **11**(1): p. 277.
10. Bild, A.H., et al., *Oncogenic pathway signatures in human cancers as a guide to targeted therapies.* Nature, 2006. **439**(7074): p. 353-357.
11. Chen, X. and L. Wang, *Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer.* Journal of Computational Biology, 2009. **16**(2): p. 265-278.
12. Lee, E., et al., *Inferring Pathway Activity toward Precise Disease Classification.* PLoS Comput Biol, 2008. **4**(11): p. e1000217.
13. Su, J., B.-J. Yoon, and E.R. Dougherty, *Accurate and Reliable Cancer Classification Based on Probabilistic Inference of Pathway Activity.* PLoS ONE, 2009. **4**(12): p. e8161.
14. Yang, X., et al., *Single Sample Expression-Anchored Mechanisms Predict Survival in Head and Neck Cancer.* PLoS Comput Biol, 2012. **8**(1): p. e1002350.
15. Yang, X., et al. *Towards Mechanism Classifiers: Expression-anchored Gene Ontology Signature Predicts Clinical Outcome in Lung Adenocarcinoma Patients.* in *AMIA 2012 Annual Symposium*. 2012. Chicago.
16. Lee, Y., et al., *Network Modeling Identifies Molecular Functions Targeted by miR-204 to Suppress Head and Neck Tumor Metastasis.* PLoS Comput Biol, 2010. **6**(4): p. e1000730.
17. Goh, W.W.B., et al., *Proteomics Signature Profiling (PSP): A Novel Contextualization Approach for Cancer Proteomics.* Journal of Proteome Research, 2012. **11**(3): p. 1571-1581.
18. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Research, 2002. **30**(1): p. 207-210.
19. Kim, Y.H., et al., *AMPKα Modulation in Cancer Progression: Multilayer Integrative Analysis of the Whole Transcriptome in Asian Gastric Cancer.* Cancer Research, 2012. **72**(10): p. 2512-2521.
20. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Meth, 2008. **5**(7): p. 621-628.
21. Cho, J.Y., et al., *Gene Expression Signature–Based Prognostic Risk Score in Gastric Cancer.* Clinical Cancer Research, 2011. **17**(7): p. 1850-1857.
22. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-193.
23. Marc Carlson, S.F., Herve Pages and Nianhua Li *KEGG.db: A set of annotation maps for KEGG.*
24. Lottaz, C., et al., *OrderedList—a bioconductor package for detecting similarity in ordered gene lists.* Bioinformatics, 2006. **22**(18): p. 2315-2316.
25. YANG, X., et al., *SIMILARITIES OF ORDERED GENE LISTS.* Journal of Bioinformatics and Computational Biology, 2006. **04**(03): p. 693-708.
26. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-15550.
27. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proceedings of the National Academy of Sciences, 2001. **98**(9): p. 5116-5121.
28. Ward, J.H., Jr., *Hierarchical Grouping to Optimize an Objective Function.* Journal of the American Statistical Association, 1963. **58**(301): p. 236-244.

PERSONALIZED MEDICINE: FROM GENOTYPES AND MOLECULAR PHENOTYPES TOWARDS COMPUTED THERAPY

OLIVER STEGLE
Max Planck Institutes Tübingen, 72076 Tübingen, Germany
Email: oliver.stegle@tuebingen.mpg.de

STEVEN E. BRENNER
Department of Plant & Microbial Biology, 111 Koshland Hall, University of California, Berkeley 94720-3102
Email: brenner@compbio.berkeley.edu

QUAID MORRIS
University of Toronto, Donnelly Centre,160 College Street, Toronto, ON M5S 3E1, Canada
Email: quaid.morris@utoronto.ca

JENNIFER LISTGARTEN
Microsoft Research, 110 Glendon Avenue, Suite PH1, Los Angeles, CA
Email: jennl@microsoft.com

Introduction

Sequencing, genotyping, and large-scale phenotyping are currently available for a number of important patient cohorts and will soon be available as a result of routine medical practice. These molecular data, in conjunction with electronic medical records and rich, on-line resources, are setting the stage for truly personalized medicine. Personalized medicine promises to yield better disease classification, enable patient-specific treatment, and also allow for improved preventive medical screening. This session explores technical challenges and new opportunities that arise from the application of genome-scale experimentation for personalized genomics and medicine.

Realizing the promises of personalized medicine requires robust analysis approaches that handle a breadth of data, addressing key statistical challenges, and understanding how to leverage the wealth of information that is available. Examples of some of these challenges include hidden structure within the data that can confound analysis results and lead to loss of power; missing or incomplete information; data heterogeneity and limitations; and the burden of multiple testing.

While these challenges are not new, per se, the scale of genomic datasets comes with added difficulties, but also offers new opportunities for methodological innovation. For example, genome-wide association studies (GWAS) generate millions of hypotheses, requiring special consideration to reduce the burden of multiple testing so that the rate of false discoveries can be controlled [1] while retaining sufficient statistical power to detect true genetic associations, for example with single nucleotide polymorphisms (SNPs). One can begin to tackle these issues by incorporation of prior information (e.g. Lee et al. [2] and Sun et al. [3]), or using multivariate modeling [4]. Tied in with these techniques are also methods that combine groups of candidate features (e.g., SNPs) in such a way as to obtain higher power, thereby attributing larger effect sizes, and uncovering a more complete picture of the underlying sources of heritability (e.g. Yang et al. [5] and Tatonetti et al. [4]). These challenges are magnified as personal genomics moves to using genome sequence data.

Statistical genomics is further complicated by the fact that, in real world settings, multiple confounders are intertwined, affecting the data in ways which require complex models and the need for heterogenous data to be analyzed together rather than independently. For example, when relating genotype to phenotype in a GWAS, population structure and family relatedness can reduce power to detect true associations and cause spurious associations [6]. Most molecular phenotypes, such as gene expression, are additionally contaminated with experimental artifacts or environmental influences. Such confounding factors, sometimes termed *expression heterogeneity*, have been shown to severely corrupt results when naïve analyses are performed [7-8,12]. When seeking the genetic underpinnings of gene expression, such as in an expression quantitative trait loci analysis, problems of population structure, family relatedness and expression heterogeneity can be jointly present, and therefore models that address all of them simultaneously are required [12]. Additionally, individual readings of high-dimensional cellular phenotypes cannot be considered as independent, and thus hypothesizing and learning hidden regulatory causes of co-expression, such as cell type or transcription factor activity, has been shown to shed light on otherwise incomprehensible expression patterns [13]. The trend we see in the problems and solutions just described is that large-scale data sets, while potentially problematic, also support analysis strategies not available on smaller datasets. In particular, they allow for us to deduce and then model hidden confounders from high-dimensional measurements, by way of Principal Components Analysis (e.g. Eigenstrat. [6]), Factor Analysis [7-8], and Linear Mixed Models [9-11], for example. All of these approaches leverage high data dimensionality, assuming that confounders act similarly on a large fraction of SNPs or phenotypes, which allows these factors to be reconstructed solely from the observed data.

Ultimately, personalized medicine needs to make its way into the clinic--results of statistical inference need to be communicated to both clinicians and patients. In such a setting, how knowledgeable do end-users need to be about statistics, molecular genetics, and machine learning in order to interpret results in a way that is useful to that user? Should software come with user-friendly tutorials on overfitting, multiple testing issues, p-values, false discovery rates and the 'winner's curse'? Although physicians and patients may be interested in inferences about health and disease, what they require assistance in acting on these inferences to guide medical and lifestyle decisions that maximize expected benefit to the patient.

Session contributions

Our session explores these challenges within the context of personalized medicine.

The keynote lecture will be from **Atul Butte**, who has extensively demonstrated how comprehensive information about impacts of genetic variation have an important role in the interpretation of individual genomes, with strong implications for the clinic.

In **Province et al.,** a statistical method is developed to allow for robust combination of analyzed data sets for meta-analysis. In particular, the authors develop a framework for combining the results from different genome-wide scans when hidden dependency structures (may) couple together the various data sets. For example, when the same individuals appear in multiple data sets, these data sets are not completely independent and should not be treated so. Similarly, if siblings appear across data sets, these data sets are not completely independent. The authors use the reported p-values from each data set to estimate the full pairwise correlation matrix between all data sets that are to be combined, and then use this correlation matrix to correct for the dependency structure. With increasing data set sizes, relatedness of individuals will become an even more pervasive problem than it currently is; the methodology introduced in this paper will enable more general meta-analyses of such data sets.

Identifying clinical risk factors related to difficult-to-diagnose diseases remains a daunting problem. Such risk factors are important for early diagnosis, prognostics and preventative care. Using a case-study for one such disease, Alzheimer's, **Li et al.**, present a strategy to identify novel clinical markers using a manually curated database containing patient phenotype data and genome-wide associations. The author's driving hypothesis is that traits that share genetic underpinning with Alzheimer's, as inferred by shared GWAS results, could serve as clinical risk factors. They find six clinical traits significantly associated with Alzheimers, of which one was not previously known as a clinical risk factor. This newly discovered association was then validated using electronic medical records, suggesting that it could be used as a new and effective prognostic marker.

Although genome-wide association scans are now routinely turning up important and reproducible associations, finding the actual causal variants responsible for disease generally requires further genotyping. **Crawford et al.** describe the properties of a custom content BeadChip designed for fine-mapping metabolic diseases and traits. Through application of this chip to 360 HapMap samples of European, African, Asian and Mexican descent, they explore the allele frequency distribution of these SNPs in these populations, and overall population differentiation. Also, they were able to identify, by way of pathway enrichment, a single SNP which indicates a difference in the functional properties of glutathione and drug metabolism through cytochrome P450 between the European and Mexican populations.

In addition to direct genetic factors, the state of microbiomes has also been shown to be predictive of phenotype and can help to understand patient well-being. To this end, it is necessary to extract useful information from metagenomic data, for example originating from the human gut. **Biswas**

et al. develop a hierarchical dictionary-based model to discover metagenomic units from pooled DNA-sequencing reads. The authors consider various likelihood models, including negative-binomial models, which are well suited for overdispersed count data. The resulting model is able to outperform several state-of-the-art assembly methods, both on simulated data and human gut metagenome datasets.

Several genomic analyses on health-related data require clustering of molecular data such as gene expression profiles. A key challenge in this context is to make an appropriate choice of the number of clusters. **Huang et al.** propose an efficient clustering approach that is suitable for heterogeneous molecular datasets as from disease studies. The developed approach is substantially faster than previous methods and does not require setting the number of clusters *a priori.* As a result, the approach yields clusterings that are better enriched for interpretable GO categories when applied to cancer genome data sets.

Finally, once molecular patterns indicative of disease have been identified, the next step is to understand the mechanisms that lead to disease. **Flores et al.** consider mutations in telomerase complexes, which can disrupt either nucleic acid binding or catalysis, thereby causing numerous human diseases. The authors tackle the underlying process by building a partial model of the human telomerase complex. Several predictions can be made from the model, elucidating disease-associated mutations.

References

1. JD. Storey, R. Tibshirani, *Proc Nat Acad Sci* **16**, 9440 (2003).
2. S.-I. Lee, A.M. Dudley, D. Drubin, P.A. Silver, N.J. Krogan, D. Pe'erand D. Koller, *PLoS Genet* **5**, e1000358 (2009).
3. L. Sun, RV. Craiu, AD. Paterson, SB. Bull, *Genet Epidemiol* **6,** 519-30 (2006).
4. NP. Tatonetti, JT. Dudley, H. Sagreiya, AJ. Butte, RB. Altman, *BMC Bioinf* **11**, S9 (2010).
5. J. Yang, B. Benyamin, BP. McEvoy, S. Gordon *et al.*, *Nat Genet* **42**, 565–569 (2010).
6. AL. Price, NA. Zaitlen, D. Reich, N. Patterson, *Nat Rev Genet* **11**, 459–463 (2010).
7. JT. Leek, JD. Storey JD., PLoS Genet **3**, e161 (2007).
8. O. Stegle, L. Parts, R. Durbin, J. Winn, PLoS Comp Biology **6**, e1000770 (2010).
9. HM. Kang, J.H. Sul *et al.*, *Nat Genet* **42**, 348–354 (2010).
10. Z. Zhang, Z. Ersoz, CQ. Lai, *et al., Nat Genet* **42**, 355–360 (2010).
11. C. Lippert, J. Listgarten, Y. Liu, C. Kadie, R. Davidson, D. Heckerman, *Nat Methods* **8**, 833-835 (2011).
12. J. Listgarten, C. Kadie, EE. Schadt, D. Heckerman D., *Proc Nat Acad Sci* **107**, 16465 (2010).
13. L. Parts, O. Stegle, J. Winn, R. Durbin, *PLoS Genet* **7**, e1001276 (2011).

AMP: ASSEMBLY MATCHING PURSUIT

S. BISWAS

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA

V. JOJIC*

Department of Computer Science, University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
E-mail:* ***vjojic@cs.unc.edu*

This paper is submitted to the Pacific Symposium on Biocomputing 2013 session **Personalized medicine: from genotypes and molecular phenotypes towards therapy**. The paper contains contains original, unpublished results, and is not currently under consideration elsewhere. All co-authors concur with the contents of the paper.

AMP: ASSEMBLY MATCHING PURSUIT*

S. BISWAS

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
E-mail: sbiswas@live.unc.edu

V. JOJIC*

Department of Computer Science, University of North Carolina at Chapel Hill,
Chapel Hill, North Carolina, USA
**E-mail: vjojic@cs.unc.edu*

Metagenomics, the study of the total genetic material isolated from a biological host, promises to reveal host-microbe or microbe-microbe interactions that may help to personalize medicine or improve agronomic practice. We introduce a method that discovers metagenomic units (MGUs) relevant for phenotype prediction through sequence-based dictionary learning. The method aggregates patient-specific dictionaries and estimates MGU abundances in order to summarize a whole population and yield universally predictive biomarkers. We analyze the impact of Gaussian, Poisson, and Negative Binomial read count models in guiding dictionary construction by examining classification efficiency on a number of synthetic datasets and a real dataset from Ref. 1. Each outperforms standard methods of dictionary composition, such as random projection and orthogonal matching pursuit. Additionally, the predictive MGUs they recover are biologically relevant.

1. Introduction

Advances in bioinformatics, refinements in DNA amplification, and the proliferation of computational power have greatly aided the analysis of DNA sequences recovered from environmental microbiomes. Early metagenomic studies focused on the sequencing of the 16S-rRNA sequence in an attempt to discover trends at a genus level.[2] Most reported large species diversity even between related hosts, and it is now becoming clear that metagenomic correlations may be better studied in other units such as genes or functional groups.[3] This requires the study of the full metagenome, a more complex task than 16S sequence study. In general, comparative metagenomics examines how the microbial composition of metagenomic samples correlates with host properties. If we can identify bacterial taxa, genes, or operons that are consistently predictive of disease, then these biological signatures could be used to build models that aid in diagnosis and treatment. For example, such approaches would have medical implications for diseases such as Inflammatory Bowel Disease (IBD) that may be treated via modulation of the gut microbiota.[4]

Our approach to summarization of metagenomic datasets is based on adaptive dictionary learning. In this framework, a signal (e.g. a set of DNA sequencing reads) is succinctly represented in terms of a small number of dictionary elements, sometimes called atoms or words. The history of dictionary learning is rich and varied, tracing back to projection pursuit.[5] Famous dictionaries, such as the Fourier basis and wavelets, have been successfully used to

*Code and supplemental material available from: `http://www.cs.unc.edu/~vjojic/amp` decompose and denoise a variety of signals.[6] Algorithms for efficient discovery of sparse representations in such dictionaries have swept through the statistical, machine learning, signal processing, and computer vision communities.[7–9] The advent of locality sensitive hashing[10] and random projection[11] have additionally made the task of handling large datasets feasible, if not trivial. Indeed, the name of the game is random projection as any projection of the data seems to be informative. However, as we show, pure random projections do not always work efficiently.

Here we demonstrate how short-read metagenomic sequencing data can be decomposed into a sequence-based dictionary assembled on the fly. The dictionary contigs composed from reads prioritized by our simple probabilistic models turn out to be discriminative. Patient-specific dictionaries can then be merged together and processed to discover a short universal dictionary that is predictive of phenotype across a population. Finally, we contrast the performances of our Assembly Matching Pursuit algorithms with the performance of a standard Random Projection method[11] and the popular short-read assembler, SOAPdenovo.[12]

1.1. *Notation and primitive sequence operations*

We will denote the ℓ_2 norm as $\|x\|_2 = \sqrt{\sum_i x_i^2}$. Given a matrix D and an index set of its columns, I, we will use D_I to denote a matrix consisting of only those columns. Similarly, for a vector w and an index set of its coordinates I we will use w_I to denote a vector composed only of those coordinates. Finally, $D_{i,:}$ denotes the i^{th} row of D.

We define `kmers(Seq,k)` as the set of all k long contiguous substrings of `Seq`; we assume that k is set ahead of time and simply use `kmers(Seq)`.

We define `overlap(Seq,Kmers,m)` the subset of k-mers in the ordered set, `Kmers`, that overlaps with at least `m` letters of either terminus of the sequence, `Seq`. A k-mer is also included in this overlapping set if its reverse and complement overlaps with `Seq`. We assume that `overlap(EMPTY,Kmers,m)` returns `Kmers`, and that `overlap(Seq,K,m)` returns `EMPTY` when no k-mer in `K` overlaps with `Seq`.

We define `count(Kmers`$_i$`, Seq)` as the number of times the i^{th} k-mer $\in$ `Kmers` occurs in `Seq`.

We define `extend(Seq,Kmer)` to be the sequence constructed by appending `Kmer` – either as given or reversed and complemented – to the overlapping end of `Seq`. This is done by removing the overlapping segment of `Seq`, and concatenating the remaining part of `Seq` with `Kmer`. We assume that `extend(Seq,EMPTY) = Seq` and `extend(EMPTY,Kmer)= Kmer`.

1.2. *Dictionaries for metagenomic read sets*

We introduce a representation of the read sets in terms of dictionaries meant to capture the k-mer profile of the sample. Given k and bound on genome length of any given microbiome's member, S, we can construct an exponentially large matrix $D : 4^k \times N(S)$ defined as $D_{rs} =$ `count(`r`, `s`)`, where r is a k-mer and s is a sequence of length S. Here $N(S) = (4^{S+1}-4)/(4-1)$, the number of sequences with length at most S.

Given the dictionary matrix, D, and vector of abundances of sequences in the microbiome, w, we can describe the observed k-mer profile, y, as a noisy version of the true k-mer profile

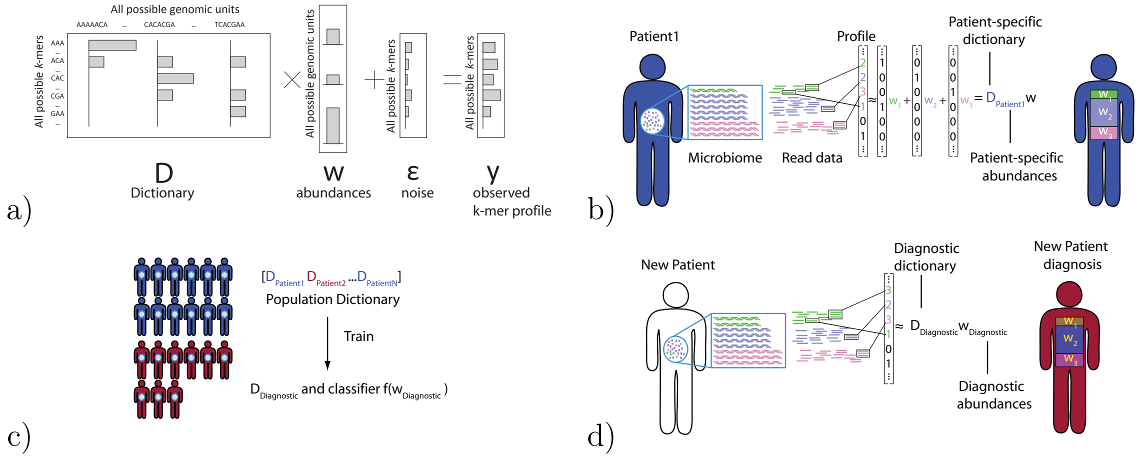

All possible genomic units
AAAAACA ... CACACGA ... TCACGAA
All possible k-mers
AAA
ACA
CAC
CGA
GAA
All possible genomic units
All possible k-mers
All possible k-mers
D
Dictionary
w
abundances
ε
noise
y
observed
k-mer profile
a)
Patient1
Profile
Patient-specific dictionary
Microbiome
Read data
Patient-specific abundances
w1 + w2 + w3 = D_Patient1 w
b)
[D_Patient1 D_Patient2 ... D_PatientN]
Population Dictionary
Train
D_Diagnostic and classifier f(w_Diagnostic)
c)
New Patient
Diagnostic dictionary
New Patient diagnosis
≈ D_Diagnostic w_Diagnostic
Diagnostic abundances
d)

Fig. 1. a) A sketch of a generative model of a read set profile of a metagenomic set using an exponentially sized **super-dictionary**. b) Learning of a **patient-specific dictionary** from a single patient's data discovers dictionary elements that represent the most abundant MGUs. c) The patient-specific dictionaries from both healthy and sick patients are aggregated into a **population dictionary** and a set of dictionary elements predictive of phenotype are selected yielding a **diagnostic dictionary**. d) Prediction of a new patient's phenotype from abundances of diagnostic MGUs.

by noting that $y = Dw + \epsilon$, see Fig. 1a.

In order to disambiguate from the later dictionaries, we call this exponentially sized dictionary the **super-dictionary**. In fact, any MGU dictionary will be a subset of the super-dictionary.

2. Methods

2.1. *Dictionary hierarchy*

We will be constructing dictionaries that are subsets of the **super-dictionary**. A **patient-specific dictionary** is a set of MGUs used to represent a particular patient's k-mer profile. A **population dictionary** is an aggregate of patient-specific dictionaries meant to represent k-mer profiles of multiple patients. We note that the MGUs found in one patient may be representative of other patients' k-mer profiles. Finally, a **diagnostic dictionary** is a subset of the population dictionary that is relevant for predicting a phenotype.

2.2. *Matching pursuit and greedy algorithms*

The matching pursuit algorithm[13] finds a representation of a signal by greedily selecting dictionary elements that best explain the signal's residual (see Algorithm 3.1). In our case, the signal corresponds to the k-mer profile and dictionary elements correspond to MGU sequences.

The most relevant observation about the matching pursuit algorithms is that *each* dictionary element is examined in order to find the one that best correlates with the residual. Given the exponential size of the super-dictionary, the matching pursuit search requirement is not feasible in polynomial time. Therefore, we turn to the area of weak greedy algorithms, in which asymptotic convergence is guaranteed even if the chosen dictionary element in each iteration does not correlate optimally with the residual.[14] This permits the use of randomized schemes that sample and accept dictionary elements if they explain a prespecified fraction of the residual signal.

The probabilistic matching pursuit (PMP) algorithm[15] leverages this intuition by probabilistically sampling dictionary elements. By avoiding an exhaustive search, PMP enables the use of large dictionaries; however, PMP methods do not explicitly optimize a likelihood, and sampling is restricted to conditioning on an element's correlation with the residual. We require a more flexible framework, and so present a generalized PMP (GPMP) algorithm (Algorithm 3.2). GPMP iteratively chooses dictionary elements that increase the likelihood of the data, $p(y|D, w)$, by sampling from a proposal distribution, $q(j|y, D, w)$.

3. Algorithm

Our patient-specific dictionary construction algorithm follows the GPMP framework. To specify the algorithm, we must select a likelihood $p(y|D, w)$ to optimize and proposal distribution $q(j|y, D, w)$ for sampling the dictionary.

3.1. *Likelihood*

The Gaussian distribution with fixed unit variance is a common choice of likelihood in matching pursuit applications,

$$\log p(y|D, w) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - D_{i,:}w)^2. \tag{1}$$

This likelihood corresponds to linear regression, and its primary benefit is the computational efficiency with which it can be optimized.

A second choice of likelihood, corresponding to Poisson regression, is

$$\log p(y|D, w) = \sum_{i=1}^{n} y_i(D_{i,:}w) - \exp\{D_{i,:}w\} - \log(y_i!). \tag{2}$$

In contrast to linear regression, which treats both positive and negative observations, Poisson regression is meant to model non-negative data, such as read counts. Notably, the expected value and variance of a Poisson random variable are equal.

A third choice of likelihood, corresponding to Negative Binomial regression, is

$$\log p(y|D, w) = \sum_{i=1}^{n} y_i(D_{i,:}w) + \frac{1}{\alpha}\log(1 - \exp\{D_{i,:}w\}) + \log\frac{\Gamma(y_i + (1/\alpha))}{\Gamma(y_i + 1)\Gamma(1/\alpha)}. \tag{3}$$

Like Poisson count models, Negative Binomial models have been used to model non-negative, integral data; however, with the additional dispersion parameter α, they can model count data with less constricted mean-variance relationships.[16]

3.2. *Dictionary element proposal*

We wish to propose a sequence j whose k-mer profile is likely to increase the objective $\log p(y|D_{I\cup j}, w_{I\cup j})$ compared to $\log p(y|D_I, w_I)$. Given this goal we can easily construct a forward sampling algorithm that will produce a reasonable candidate sequence. Specifically, we

Algorithm 3.1. *Matching Pursuit*

```
Input: $D, c$
Output: $w$, such that $\|y - Dw\|_2 \leq c$
initialize $w_j = 0, \forall j$
while($\|y - Dw\|_2 \leq c$)
  $R = y - Dw$
  $c_j = \left| \frac{\langle R, D_k \rangle}{\langle D_k, D_k \rangle} \right|, \forall j$
  $k = \text{argmax}_j \, c_j$
  $I = I \cup \{k\} \quad w_k = c_k$
```

Algorithm 3.2. *Generalized Probabilistic Matching Pursuit*

```
Input: $D, y, c$
Output: $I$ and $w$ such that $\log p(y|D, w) > c$
$I = \emptyset, w_j = 0, \forall j$
while($\log p(y|D, w) \leq c$)
  sample $\{j\}$ from $q(j|y, D, w)$
  $I = I \cup j$
  $w_I = \text{argmax}_v \log p(y|D_I, v)$
return $I, w$
```

Algorithm 3.3. *Dictionary element proposal*

```
Input: $D, y, w, m$, Orthogonal,
  set of all observed $k$-mers $K$
Output: a candidate dictionary element $s$
$s = $ EMPTY$, I = \{i | w_i \neq 0\}$
repeat
  $K_s = $ overlap$(s, K, m) \cup \{$EMPTY$\}$
  if(Orthogonal)
    $K_s = K_s - \cup_{i \in I}$ kmers$(i)$
  foreach($l \in K_s$)
    $s' = $ extend$(s, l)$
    $I' = I \cup \{s'\}$
    $w_{I'} = \text{argmax}_w \log p(w|y, D_{I'})$
    $\pi_l = p(y|D_{I'}, w_{I'})$
  sample $l^*$ from normalized $\pi$
  $s = $ extend$(s, l^*)$
until($l^* = $ EMPTY)
```

initialize a new dictionary element (contig) by sampling a k-mer based on the increase in likelihood if that k-mer, alone, were to enter the model as an element. When extending a contig, an overlapping k-mer is sampled according to the change in likelihood that would result if it were added to the growing element. If no k-mer sufficiently improves the likelihood, then the algorithm may sample the `EMPTY` k-mer (i.e. choose to terminate extension), and thus complete an iteration. The likelihood biases the algorithm toward sampling k-mers that occur with high frequency.

Another proposal distribution produces orthogonal dictionary elements, and is accordingly used by orthogonal matching pursuit algorithms. Since the entries in our dictionary are always nonnegative, two dictionary elements will be orthogonal, ($\sum_k D_{k,i} D_{k,j} = 0$), if and only if their corresponding MGUs do not share any k-mers. Algorithm 3.3 implements both of these choices specified by argument `Orthogonal`.

3.3. *The AMP algorithms*

We introduce four algorithms based on choices of likelihood and dictionary element proposal.

(1) Gaussian Assembly Matching Pursuit (GAMP) combines the likelihood from (1) and the non-orthogonal dictionary proposal.

(2) Poisson Assembly Matching Pursuit (PAMP) combines the likelihood from (2) and the non-orthogonal dictionary proposal.
(3) Negative Binomial Assembly Matching Pursuit (NAMP) combines the likelihood from (3) and the non-orthogonal dictionary proposal.
(4) Orthogonal Assembly Matching Pursuit (OAMP) uses the orthogonal dictionary proposal.

Because the orthogonal dictionary proposal strongly constrains dictionary construction, the choice of likelihood is irrelevant.

3.4. *Population dictionary construction and patient summarization*

Given a learned patient-specific dictionary we are tasked with constructing a dictionary that can be used universally across the whole patient population, the **population dictionary**. We can construct this dictionary by pooling all patient specific dictionaries, but here we face two challenges:

(1) How do we estimate abundances of the population dictionary elements in each patient?
(2) Which of the population dictionary elements are diagnostically relevant?

Abundance estimation A patient's k-mer profile may be regressed onto the population dictionary in order to estimate MGU abundances. We utilize Negative Binomial (NB) regression due to its flexibility in modeling potentially overdispersed data, such as read counts.[16] NB models have been fruitfully applied in RNA-Seq data analysis,[17] and we have found that abundances estimated by NB regression – regardless of dictionary origin – are more accurate than those estimated using other likelihoods (data not shown).

Using abundances as predictors MGU abundance estimates can be directly used as predictors of phenotype. In terms of interpretability, logistic regression is most appealing. In our experiments we use an efficient implementation of sparse logistic regression.[18] The sparsity inducing, ℓ_1-penalty selects only a small portion of the features to participate in phenotype prediction from an otherwise large population dictionary. Because the optimal scale, λ, of the ℓ_1-penalty is unknown, it must be estimated from the data. The data are therefore split into training, validation, and test sets – the validation set is used to determine the λ parameter. The sparsest model that classifies the validation set statistically as well as the best model is chosen. The classification accuracy of this logistic regression model is then evaluated on the test set.

The chosen set of MGUs that are predictive of phenotype correspond to parts of the population dictionary that can be diagnostically useful. Thus, they compose the **diagnostic dictionary**.

4. Implementation

We customized an implementation of a succinct suffix trie[19] to store suffix and prefix k-mer tries. The counts of each k-mer are also stored during trie construction. While the AMP algorithms' implementation is straightforward, here we draw attention to two issues relating to likelihood optimization and read storage and querying.

The AMP assemblers are string-based and rely on greedy extension. However, extension is stochastic and is guided by the likelihood of the observed k-mer profile (Algorithms 3.2 & 3.3). When updating the weight of a growing contig to a conditional maximum likelihood value (i.e. computing $w_j = \text{argmax}_{v_j} \log p(y|D, v_j))$, GAMP equates the first partial derivative of the likelihood (with respect to the contig's weight) to zero and solves for w. NAMP and PAMP, on the other hand, utilize Newton-Raphson updates to find a w that maximizes the likelihood (a closed-form solution for w_I does not exist when equating the gradient of the negative binomial or Poisson likelihoods to zero).

5. Results and Discussion

To assess the ability of our methods to produce discriminative diagnostic dictionaries, we turned to synthetic and real data experiments. We put particular focus on the efficiency with which our AMP methods could produce representations relevant for phenotype prediction.[b]

In all synthetic experiments we worked with k-mers of fixed read length. The real dataset consisted of a mix of 75bp and 44bp read datasets. Hence we used k-mer length of 44bp, using shorter reads directly as k-mers. From each longer 75bp read we constructed 3 44-mers with 16bp spaced starting offsets. In our dictionary element proposal algorithm we required that a k-mer achieve an overlap of 20bp with a growing contig to be considered a candidate for appending.[c] Finally, to estimate the classification accuracy we performed a 10-fold cross-validation with an inner cross-validation on the training and validation sets to select λ (the held out data in the outer fold were not used during training).

5.1. *Baselines*

For comparision, we chose to analyze the quality of dictionaries produced by SOAPdenovo[12] and a pure random projection method.[11]

For synthetic experiments, SOAPdenovo was run on each sample using a single thread and a minimum k-mer overlap (option `-K`) of 21 for extension purposes. For the real data from Ref. 1, we used the SOAPdenovo contigs already generated in their paper.[d] Because SOAPdenovo's assembly is not likelihood driven, the order in which contigs are produced is not interpretable. Thus, the longest SOAPdenovo contigs with high coverage were added in a random fashion when evaluating successively larger population dictionaries.

Random projections (RP) summarize a set of points by projecting them into a lower dimensional subspace defined by a intelligently chosen, but random basis. If done properly, the relative distances between points before and after projection will be, on average, approximately preserved. In the case of metagenomic samples, we treat each sample's k-mer profile as a K dimensional point. Application of RP produces a new, smaller set of features that are sums

[b] Efficiency refers to the size of the population dictionary required to produce a diagnostic dictionary capable of achieving a particular classification accuracy.

[c] This amounts to using `m=20` in Algorithm 3.3.

[d] They used `-K` 21 and `-K` 23 for 44bp and 75bp reads, respectively.

of randomly weighted k-mer profiles, each with dimension $C < K$. If we have N samples then, $w^{\mathrm{RP}} = PY$ where Y is the $K \times N$ matrix of k-mer profiles from all samples, P is the $C \times K$ random projection matrix, and w^{RP} is the resulting $C \times N$ projected form of Y. These new C dimensional features are roughly akin to abundances produced by the AMP methods and are treated as such during our classification step. Indeed, an implicitly constructed dictionary matrix can be defined as matrix D that satisfies $P = (D^T D)^{-1} D^T$. The matrix P is constructed using the method described in Ref. 11 and refer to the algorithm as ARP.

5.2. *Synthetic data generation*

A/T SNP A 10Kb sequence was randomly generated and duplicated. The 5000th base in one duplicated copy was changed to an 'A' and the 5000th base in the other copy was changed to a 'T'. We then generated 100 synthetic metagenomic samples, 50 of which were phenotypically 'sick' and 50 of which were phenotypically 'healthy'. For each of the 100 samples, 20000 75bp reads with 2% noise were simulated from the 10Kb templates. A 50/50 and 33/67 ratio of the two variants were maintained for 'healthy' and 'sick' sample, respectively.

Distinct species For this synthetic experiment 40 10Kb sequences were randomly generated. From this true dictionary, we generated 100 synthetic metagenomic samples, 50 of which were phenotypically 'sick' and 50 of which were phenotypically 'healthy'. For each of the 100 samples, 40000 75bp reads with 3% noise were simulated from the 10Kb templates with varying coverage. Average baseline mixing proportions of the templates followed an exponential decay; however for 'sick' samples, the relative abundances of the 7th, 13th, and 24th most abundant templates were altered by 1%, 0.67%, and 0.33%, respectively (see Supp. Info. Fig. 1 for exact abundances).

Synthetic community The Genome Institute at the Washington School of Medicine has produced many draft-quality genomes of various human gut microbes.[e] We selected 31 microbial species' genomes to represent actual genera and phyla found in the human gut. From this true dictionary, we generated 100 samples, 50 'healthy' and 50 'sick', each with 10 million, 75bp reads with 3% noise. Baseline mixing proportions of each microbe in all samples were set in accordance with relative abundances reported in Ref. 1 and Ref. 20 based on the genus and phylum they represent; however, the relative abundances of 3 microbes in 'sick' samples were altered by 1%, 2%, and 3% (see Supp. Info. Fig. 2 for exact abundances).

5.3. *Synthetic data results*

Fig. 2 shows the performance of the methods in classifying 'healthy' and 'sick' samples from the synthetic experiments.

In the A/T SNP experiment, the discriminative abundances are driven by the reads spanning the SNP position. Without noise, all methods converge quickly and classification is trivial for SOAPdenovo and OAMP as orthogonality requirements ensure that none of the shared

[e] Freely available from `http://genome.wustl.edu/pub/organism/Microbes/Human_Gut_Microbiome/`.

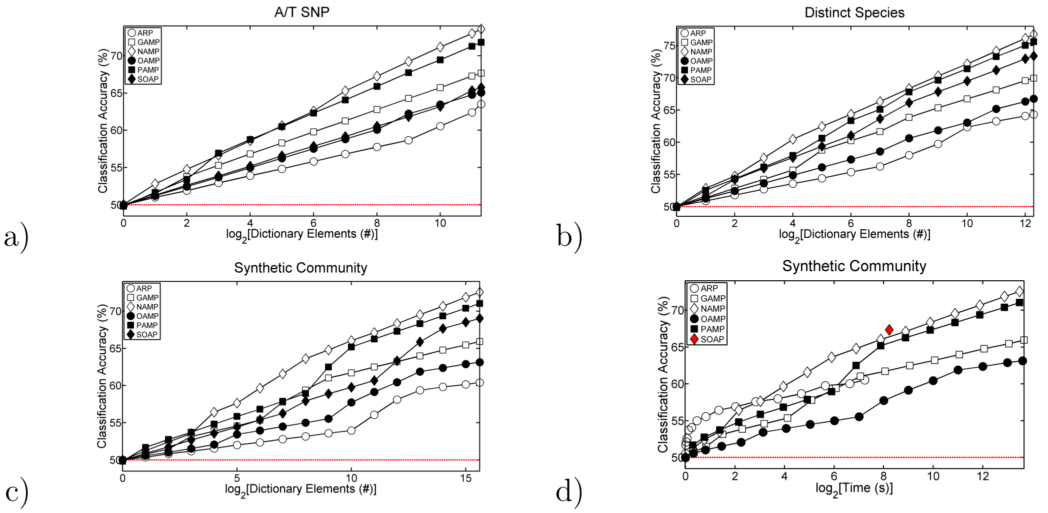


Fig. 2. a,b,c) Mean classification accuracies of the methods on synthetic datasets. The dashed red line corresponds the performance expected when always predicting 'healthy'. Plots with confidence bands around the mean can be seen in Supp. Info. Fig. 3. d) Performance comparison with respect to running time of each method on the synthetic community experiment. SOAPdenovo performance is marked with a single red diamond since the order in which it produces contigs is not interpretable.

reads between species are available for the second contig (data not shown). However, in a more realistic, noisy setting contig construction is more difficult. Nevertheless, NAMP, PAMP and GAMP begin to discover sequences containing the discriminative SNP within the first 8 contigs, before the other methods.

In the distinct species experiment, species' k-mer profiles are nearly orthogonal. Without noise, OAMP and SOAPdenovo reconstruct the true dictionary within the first 40 iterations (data not shown). With noise, OAMP spends more iterations constructing subcontigs of the true dictionary elements that are non-discriminitave. By exploring only well-supported edges in its De Bruijn graph construction, SOAPdenovo better handles the noise, constructs longer contigs, and thereby discovers significant features more quickly.

Interestingly, in all scenarios, including the synthetic community, there is a steady and consistent difference in performance between GAMP, PAMP, and NAMP. This illustrates clear ordering between the three choices of likelihood: NAMP $\succ$ PAMP $\succ$ GAMP. Additionally, NAMP and PAMP discover discriminative features sooner than the other methods,[f] and with the exception of the A/T SNP experiment, SOAPdenovo consistently outperforms GAMP. These results suggest that given the appropriate read count model, likelihood driven assembly can direct the early discovery of predictive features. Finally, the subpar performance of ARP on all experiments demonstrates the benefit of computing abundances of sensible contigs in a manner consistent with the nature of the data.

[f] This comparison is not directly applicable to SOAPdenovo as its assembly is not order dependent.

5.4. *Human Gut Metagenome Analysis*

In addition to synthetic experiments, we tested our method on data from Ref. 1. This data set contains 576 gigabases of sequence data obtained from the fecal samples of 124 Spanish and Danish individuals, 25 of whom have inflammatory bowel syndrome (IBD). Population dictionary pools for the AMP methods were constructed by aggregating the first 1000 dictionary contigs greater than 500 bp of each patient. For SOAPdenovo's pool we took the longest 124000 contigs of the roughly 6.6 million contigs greater than 500 bp produced by SOAPdenovo in Ref. 1. ARP's pool was constructed by producing 1000 random projections for each of the 124 patients, since ARP does not have a concept of a patient specific dictionary. From each of their respective pools, successively larger population dictionaries were constructed in order to evaluate each method's classification accuracy. Length distributions for the contigs used in each population dictionary can be found in Supp. Info. Fig. 4.

Fig. 3a) shows the method performances in classifying individuals based on their health status (IBD or healthy). We see that all methods discover relevant contigs, but at a different rate. The leading algorithm is NAMP, followed closely by PAMP, and thereafter SOAPdenovo. We see that GAMP and OAMP are relatively close in terms of performance but for different reasons. The OAMP is affected by the orthogonality requirement while Gaussian likelihood is overly greedy, driven by the quadratic cost on the residual.

From the final GAMP, SOAPdenovo, PAMP, and NAMP dictionaries, 11, 18, 18, and 19 metagenomic units, respectively, were found to have non-zero weight, suggesting their importance as potential biomarkers for IBD (Fig. 3b)). We obtained KO (KEGG orthologous groups) numbers for each of these features using KAAS, an annotation server that queries the KEGG database.[21] For discovered enzymes we additionally mined the KEGG BRITE database to obtain a functional annotation. Finally, as a measure of consistency between our method and an independent biological study, we noted any commonalities between our annotations and those of Ref. 22 (see Supp. Info., Fig. 5). Of our 48 features, 10 were found to be either enriched or depleted in the Ref. 22 analysis. In particular, 4 were related to the PTS, a system important for sugar transport into the cell and recently found to include biomarkers for IBD.[23] We additionally found nitrate reductase among our significant features. Nitrate reductase plays an important role in the conversion of nitrate to nitrite and nitric oxide, neither of which can be synthesized by human DNA. Unsurprisingly then, elevated levels of nitric oxide have been found to correlate with IBD.[24] Finally, we noted the presence of vanillate monooxygenase, an agent that may play a role in xenobiotic degredation of phenolic compounds, such as p-cresol, another correlate of IBD.[25]

5.5. *Time and memory*

Fig. 2a,b,c) and 3 describe the efficiency of the various methods in terms of accuracy gained per added dictionary element. To gauge computational efficiency, it is important to consider efficiency with respect to running time.

Fig. 2d) illustrates the performance of the various methods with respect to time in the synthetic community experiment and corresponds with the accuracies depicted in Fig. 2c). For the AMP methods, time required to achieve a particular classification accuracy was calculated

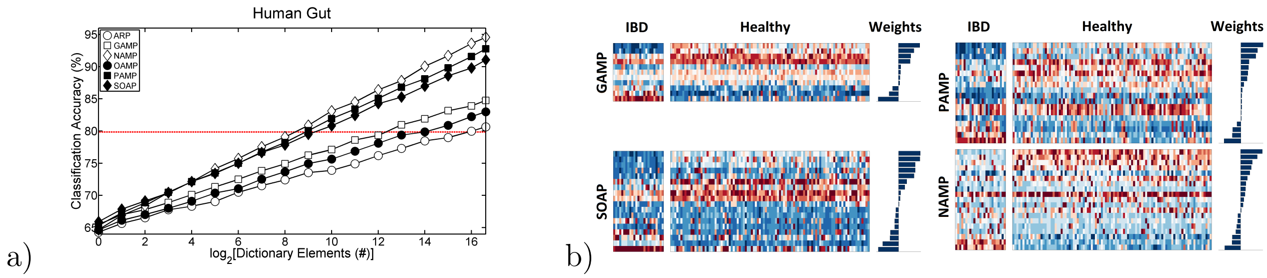


Fig. 3. a) Mean classification accuracy of each method on the real dataset. Dashed red line is the performance expected by always predicting 'healthy'. Plots with confidence bands around the mean can be seen in Supp. Info. Fig. 3. b)Most predictive dictionary element abundances for the healthy and sick patients stemming from the different methods (GAMP, SOAP, PAMP, NAMP) as well as weights of these abundances in a sparse logistic regression trained model.

as the sum total of the times required to generate each contig used in the corresponding population dictionary. For the ARP the performance curve is parameterized by the number of rows in the projection matrix. Each successive point on the AMP and ARP curves corresponds to a two-fold increase in the number of contigs over the previous point. For SOAPdenovo we measured the total running time required to assemble all synthetic samples (2.99×10^4 seconds) and noted how many contigs were produced (5.00×10^6). Thus, we extrapolated the time required to generate a final population dictionary of size 50000 to be $50000 \div (5.00 \times 10^6) \times (2.99 \times 10^4) = 299$ seconds. SOAPdenovo's performance is depicted as a single point since the contigs it produces are not necessarily order dependent.

SOAPdenovo reaches its final 67% accuracy before NAMP and PAMP, and handily outperforms GAMP and OAMP. The AMP methods are not as time efficient due to the expensive floating point arithmetic (e.g. computing exponents and logarithms) associated with the likelihood computations. However, NAMP and PAMP offset these inefficiencies by nearly reaching the same accuracy as SOAPdenovo in the same time and with a dictionary $1/25^{th}$ the size. Ultimately, with equally large dictionaries as SOAPdenovo, NAMP and PAMP provide superior performance by classifying 4-5% more accurately.

The AMP methods additionally require less memory than SOAPdenovo. On average, SOAPdenovo requires 2358 bytes per 75 bp read, whereas the AMP methods require 2037 bytes per 75 bp read (Supp. Info. Fig. 6). These reads were taken from the synthetic community experiment.

6. Conclusion

We introduced the Assembly Matching Pursuit family of methods for metagenomic dataset summarization and analysis. Our AMP methods follow a novel generalized matching pursuit paradigm, which guides dictionary construction using likelihood based principles. Within this framework, we explored the appropriateness of popular likelihood choices for modeling read counts and accordingly derived the GAMP, PAMP, and NAMP assemblers. In investigating an alternative proposal distribution, we derived the OAMP assembler, which enforces orthogonality among its contigs.

We also introduced a simple abundance estimation protocol that directly regresses k-mer profiles of any read sample on a set of dictionary sequences. Indeed, a dictionary does not have to be composed of contigs from our AMP methods. It may generated by SOAPdenovo, any another assembler, or in the future, set to be a large sequence database.

By coupling AMP assembly with a negative binomial based abundance estimator, we have put forth a simple method of aggregating sample dictionaries into a population dictionary from which learned abundances can be leveraged as predictors of phenotype. In both synthetic and real datasets we show that this new family of methods does significantly better in phenotype discrimination than random projections. Further, due to their simplicity, the methods easily handle large scale datasets, such as in Ref. 1, which spans 0.6 terabases. Finally, while we focused on medical applications as an illustration, the method is applicable to other metagenomic, and in principle, RNA-seq studies.

References

1. J. Qin, R. Li, J. Raes, M. Arumugam *et al.*, *Nature* **464**, 59 (March 2010).
2. P. Hugenholtz, B. M. Goebel and N. R. Pace, *J. Bacteriol.* **180**, 4765 (Sep 1998).
3. C. Burke, P. Steinberg, D. Rusch, S. Kjelleberg and T. Thomas, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 14288 (Aug 2011).
4. D. Knights, L. Parfrey, J. Zaneveld, C. Lozupone and R. Knight, *Cell Host & Microbe* **10**, 292 (October 2011).
5. J. H. Friedman and J. W. Tukey, *IEEE Trans. Comput.* **23**, 881 (1974).
6. S. Mallat, *A wavelet tour of signal processing*Wavelet Analysis and Its Applications Series, Wavelet Analysis and Its Applications Series (Academic Press, 1999).
7. D. L. Donoho, *IEEE Transactions on Information Theory* **41**, 613 (1995).
8. E. J. Candès and T. Tao, *IEEE Transactions on Information Theory* **52**, 5406 (2006).
9. J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang and S. Yan, *Proceedings of the IEEE* **98**, 1031 (June 2010).
10. A. Gionis, P. Indyk and R. Motwani, Similarity search in high dimensions via hashing1997.
11. D. Achlioptas, *Journal of Computer and System Sciences* **66**, 671 (June 2003).
12. R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang and J. Wang, *Genome Res.* **20**, 265 (Feb 2010).
13. S. G. Mallat and Z. Zhang, *IEEE Trans. Signal Process.* **41**, 3397 (1993).
14. V. Temlyakov, *Greedy Approximation* (Cambridge University Press, 2011).
15. S. E. Ferrando, E. J. Doolittle, A. J. Bernal and L. J. Bernal, *Signal Processing* **80**, 2099 (October 2000).
16. J. Hilbe, *Negative Binomial Regression*, 2nd edn. (Cambridge University Press, 2011).
17. S. Anders and W. Huber, *Genome Biol.* **11**, p. R106 (2010).
18. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin, *Journal of Machine Learning Research* **9**, 1871 (2008).
19. D. Okanohara, ux-trie `http://code.google.com/p/ux-trie`, (2012).
20. M. Arumugam, J. Raes, E. Pelletier *et al.*, *Nature* **473**, 174 (May 2011).
21. Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa and M. Kanehisa, *Access* **35**, 182 (2007).
22. S. Greenblum, P. J. Turnbaugh and E. Borenstein, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 594 (Jan 2012).
23. A. L. Francl, T. Thongaram and M. J. Miller, *BMC Microbiology* (2010).
24. G. Kolios, V. Valatas and S. G. Ward, *Immunology* (2004).
25. M. H. van Nuenen, K. Venema, J. van der Woude and E. J. Kuipers, *Digestive Diseases* **49**, 485 (2004).

CHARACTERIZATION OF THE METABOCHIP IN DIVERSE POPULATIONS FROM THE INTERNATIONAL HAPMAP PROJECT IN THE EPIDEMIOLOGIC ARCHITECTURE FOR GENES LINKED TO ENVIRONMENT (EAGLE) PROJECT

DANA C. CRAWFORD

Department of Molecular Physiology and Biophysics, Center for Human Genetics Research,
Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: crawford@chgr.mc.vanderbilt.edu

ROBERT GOODLOE

Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: robert.j.goodloe@vanderbilt.edu

KRISTIN BROWN-GENTRY

Center for Human Genetics Research, Vanderbilt University, 1207 17th Avenue, Suite 300
Nashville, TN 37232, USA
Email: kristin.brown@chgr.mc.vanderbilt.edu

SARAH WILSON

Center for Human Genetics Research, Vanderbilt University, 1207 17th Avenue, Suite 300
Nashville, TN 37232, USA
Email: sarah.wilson@chgr.mc.vanderbilt.edu

JAMIE ROBERSON

Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: jamie.l.roberson@vanderbilt.edu

NILOUFAR B. GILLANI

Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: nila.gillani@vanderbilt.edu

MARYLYN D. RITCHIE

Department of Biochemistry and Molecular Biology, Center for System Genomics, Pennsylvania State University, 512 Wartik Lab
University Park, PA 16802, USA
Email: marylyn.ritchie@psu.edu

HOLLI H. DILKS

Department of Molecular Physiology and Biophysics, Center for Human Genetics Research,
Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: holli.dilks@chgr.mc.vanderbilt.edu

WILLIAM S. BUSH

Department of Biomedical Informatics, Center for Human Genetics Research, Vanderbilt University,
2215 Garland Avenue, 519 Light Hall

Nashville, TN 37232, USA
Email: william.s.bush@vanderbilt.edu

Genome-wide association studies (GWAS) have identified hundreds of genomic regions associated with common human disease and quantitative traits. A major research avenue for mature genotype-phenotype associations is the identification of the true risk or functional variant for downstream molecular studies or personalized medicine applications. As part of the Population Architecture using Genomics and Epidemiology (PAGE) study, we as Epidemiologic Architecture for Genes Linked to Environment (EAGLE) are fine-mapping GWAS-identified genomic regions for common diseases and quantitative traits. We are currently genotyping the Metabochip, a custom content BeadChip designed for fine-mapping metabolic diseases and traits, in~15,000 DNA samples from patients of African, Hispanic, and Asian ancestry linked to de-identified electronic medical records from the Vanderbilt University biorepository (BioVU). As an initial study of quality control, we report here the genotyping data for 360 samples of European, African, Asian, and Mexican descent from the International HapMap Project. In addition to quality control metrics, we report the overall allele frequency distribution, overall population differentiation (as measured by F_{ST}), and linkage disequilibrium patterns for a select GWAS-identified region associated with low-density lipoprotein cholesterol levels to illustrate the utility of the Metabochip for fine-mapping studies in the diverse populations expected in EAGLE, the PAGE study, and other efforts underway designed to characterize the complex genetic architecture underlying common human disease and quantitative traits.

1. Introduction

In the last seven years, genome-wide association studies (GWAS) have been used extensively to identify common genetic variants associated with human diseases and quantitative traits. While there are many replicated and mature, known relationships between genomic regions and phenotypes, very few individual genetic variants have been identified as the risk variant for downstream molecular studies or personalized medicine applications. The lack of true functional variants revealed by GWAS stems from the fact that GWAS is based on linkage disequilibrium (LD), the non-random association of alleles at different variants along the chromosome. That is, GWAS fixed-content products mostly assay presumably neutral common genetic variants that are in LD or "tag" other genetic variants not directly assayed resulting in GWAS-identified regions that probably contain the true risk (unassayed) variant.

To identify the true risk variant, a major proposed activity in the "post-GWAS" era is *fine mapping*. In a fine-mapping experiment, the GWAS-identified region is densely interrogated via thousands of common and rare variants. Fine-mapping experiments can also take advantage of the known LD differences observed across populations. For example, populations of African-descent have lower levels of LD compared with populations of European-descent and therefore may be useful in identifying the risk variant masked by higher levels of LD in other populations. Fine mapping across populations is also useful for identifying population-specific variants associated with phenotypes.

In recognition for the need to fine-map mature GWAS-identified regions originally identified in European-descent populations, the National Human Genome Research Instituted established the Population Architecture using Genomics and Epidemiology (PAGE) study to genotype African American and Asian populations linked to phenotypes using the Illumina Metabochip, a custom iSelect BeadChip designed to fine-map GWAS-identified regions for metabolic diseases and traits. We as Epidemiologic Architecture for Genes Linked to Environment (EAGLE) are genotyping ~15,000 DNA samples linked to de-identified electronic medical records in the Vanderbilt University biorespository (BioVU) for fine mapping within the PAGE study. As the first step in quality control, EAGLE has genotyped 360 HapMap samples from European, African, Asians, and Mexican-descent populations. This short report describes the quality control, variant properties, and the potential for fine mapping of GWAS-identified regions in the anticipated populations within EAGLE and the PAGE study.

2. Methods

2.1. *Study populations*

DNA samples were obtained by the PAGE Coordinating Center from the Coriell Cell Repositories[1]. A total of 360 samples overlapping the International HapMap Project collection were obtained, including 30 trios of Northern and Western European ancestry from Utah from the Centre d'Etude du Polymorphisme Humain (CEPH) collection (CEU; catalog ID HAPMAPPT01), 90 unrelated individuals representing 45 individuals each from Tokyo, Japan and Beijing, China (ASN; catalog ID HAPMAPPT02), 30 trios from the Yoruba in Ibadan, Nigeria (YRI; catalog ID

HAPMAPPT03), and 30 trios from communities of Mexican origin in Los Angeles, California (MEX; catalog ID HAPMAPV13). Samples were chosen to reflect the overall genetic ancestry of epidemiologic and clinical-based samples available in the PAGE study[1].

2.2. *Genotyping*

Aliquots of HapMap DNA samples were distributed by the PAGE Coordinating Center to individual PAGE study sites. The Vanderbilt DNA Resources Core genotyped the Illumina Metabochip on the HapMap samples distributed by the PAGE Coordinating Center on the Illumina iScan (San Diego, California). The Metabochip is a custom BeadChip targeting 196,725 genetic variants. Common and less common genetic variants were chosen from among the first iteration of the 1000 Genomes Project and represent index GWAS-identified variants regardless of disease or phenotype as of 2009; regions targeted for fine-mapping for specific GWAS-identified regions associated with coronary artery disease, type 2 diabetes, QT-interval, body mass index/obesity, lipid traits, glycemic traits, and blood pressure; mitochondrial markers; HLA markers; sex chromosome markers; and ancestry informative markers[2, 3]. Illumina software GenomeStudio (v1.7.4) was used to determine the genotype calls for each variant for each sample, and manual re-clustering was performed on all mitochondrial and Y chromosome variants. Data were stored and accessed by the Vanderbilt Computational Genomics Core for quality control and downstream analyses using BC Platforms (Espoo, Finland).

2.3. *Statistical methods*

Standard quality control metrics were generated using PLINKv1.07[4] and PLATOv0.84[5]. F_{ST} calculations were based on the Weir and Cockerham algorithm[6] implemented in PLATO. Allele frequencies and F_{ST} were calculated for CEU, YRI, JPN and CHB combined (ASN), and MEX unrelated samples separately. Linkage disequilibrium (r^2) was calculated using independent samples stratified by race/ethnicity using Haploviewv4.2[7].

3. Results

We genotyped 360 DNA samples from the International HapMap collection including 90 CEU, 90 YRI, 90 ASN, and 90 MEX on the Illumina Metabochip. From the 360 samples, 358 (99%) samples were successfully genotyped. And, out of the targeted 196,725 genetic variants on the Metabochip, we obtained data for 185,788 genetic variants for an overall pre-quality control call rate of 94.44%. From this initial dataset, we then performed quality control as outlined by Buyske et al[2] (Table 1).

Table 1. Number of genetic variants removed from Metabochip dataset after quality control, by criteria and HapMap population. We performed quality control steps appropriate for a single dataset as outlined by Buyske et al.[2]. Lower genotyping call rates were observed for YRI compared with other HapMap populations consistent with our observations for targeted genotyping in EAGLE (data not shown).

Criteria	SNP Failure Determination	# SNPs removed				
		CEU	**YRI**	**CHB**	**JPN**	**MEX**
Call Rate	< 0.95	14515	73445	11851	13585	14871
Mendelian Errors	> 1 (out of 30 trios)	97	10	0	0	144
Replication Errors	> 2	0	0	0	0	0
Hardy-Weinberg Equilibrium *p*-value	$< 1 \times 10^{-6}$	11	1	11	10	19
Discordant calls versus HapMap database	> 3 (out of 90 samples)	329	178	285	292	301

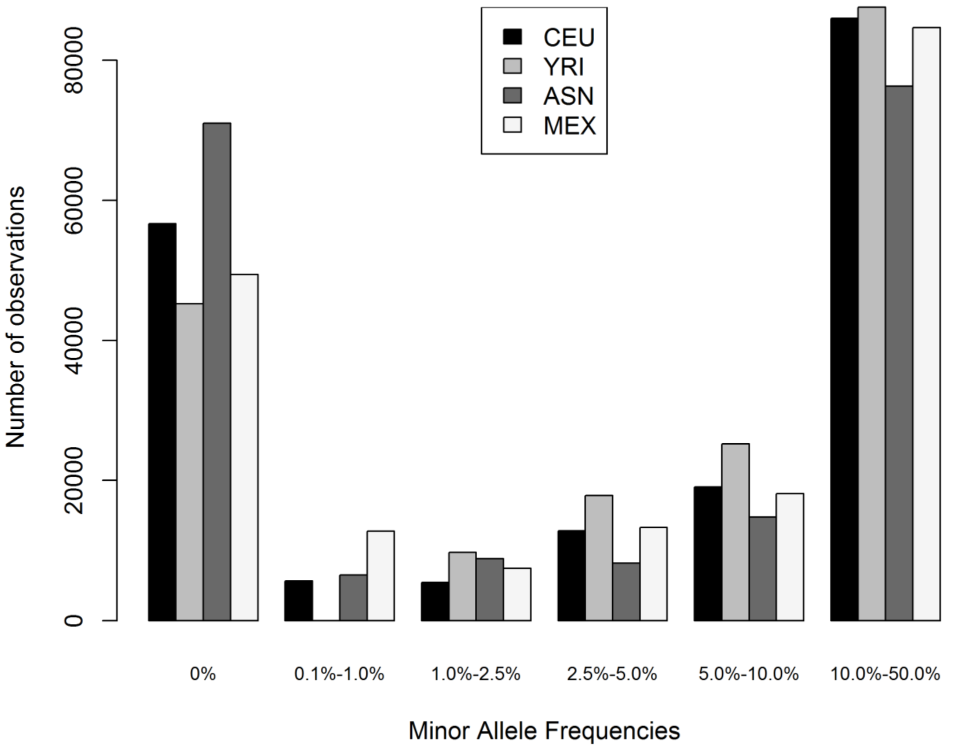


Figure 1. Distribution of minor allele frequencies of genetic variants assayed by the Metabochip, by HapMap population. Allele frequencies were determined in the founder (unrelated) samples of Northern and Western European ancestry (CEU; n=60), West African ancestry (YRI; n=60), Asian ancestry (ASN; n=90), and Mexican ancestry (MEX; n=60). On the x-axis, genetic variant frequencies were binned as monomorphic, rare (0.1%-1-%), less common (1-2.5%), and common (2.5-5%, 5-10%, and 10-50%) by population. Number of observations for each bin is given on the y-axis.

To examine potential population differences for genetic variants targeted by the Metabochip, we first determined minor allele frequencies for every variant by HapMap population. As shown in Figure 1, the majority of variants for this custom BeadChip are polymorphic. More than one half (58% for ASN) to up to three-quarters (75% for YRI) of the alleles assayed by the Metabochip occurred at greater than 1% frequency. Conversely, one quarter (24% for YRI) to more than one-third (38% for ASN) of the variants were monomorphic in this small sample set.

We also calculated a fixation index, $F_{ST,}$ for all pair-wise population comparisons. F_{ST} is an estimate of population differentiation ranging from 0 (no measureable genetic differentiation) to 1.0 (very great genetic differentiation), and its distribution for Metabochip-targeted variants in HapMap samples is given in Figure 2. The majority (76%) of F_{ST} values are less than 0.15 for all genetic variant pair-wise population comparisons. The most population differentiation was observed between YRI and ASN. Conversely, the least population differentiation was observed between CEU and MEX.

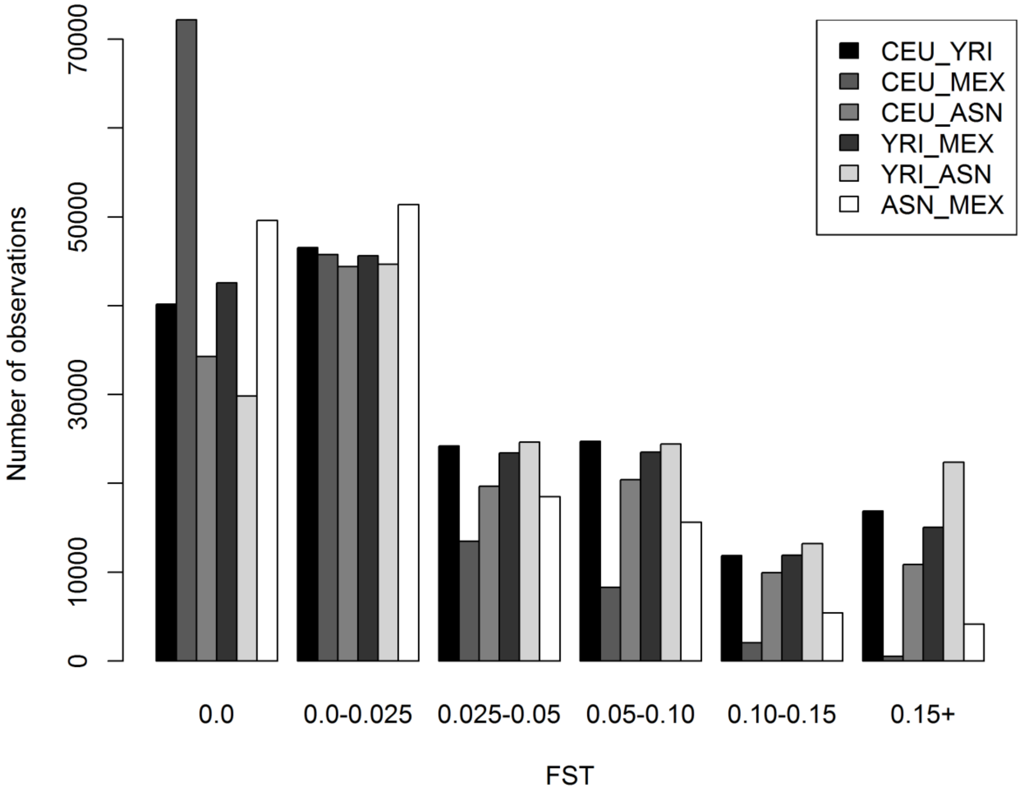


Figure 2. Distribution of genetic differentiation (F_{ST}) by HapMap population pairwise comparison. F_{ST}, a measure of population differentiation, was calculated per SNP in PLATO based on the Weir and Cockerham algorithm[6] for each HapMap population pair. Calculations were performed on unrelated samples of Northern and Western European ancestry (CEU; n=60), West African ancestry (YRI; n=60), Asian ancestry (ASN; n=88), and Mexican ancestry (MEX; n=60). On the x-axis, F_{ST} values were binned no difference (zero), >0.0-0.25, >0.025-0.05, >0.05-0.10, >0.10-0.15, >0.15 by pair-wise population comparison. Number of observations for each bin is given on the y-axis

We mapped the most highly differentiated SNPs ($F_{ST} > 0.15$) to dbSNP identifiers (143,750 successfully mapped to known SNPs), and examined the degree to which alleles altered the expression or function of genes using annotation resources from the Genome-Wide Annotation Repository (http://gwar.mc.vanderbilt.edu). We defined two categories of SNP annotation for this analysis: predicted changes to protein function via SIFT and PolyPhen2 algorithms [8, 9], and prior associations to expression levels of nearby genes [10, 11]. The total number of SNP and gene annotations is shown in tables 2 and 3.

Table 2. Number of differentiated SNPs showing functional effects

Population Comparison	SIFT (Deleterious)	PolyPhen2 (Possibly or Probably Damaging)	Significant eQTL	Total functional SNPs*	Total Differentiated SNPs
ASN/MEX	6	12	202	218	4059
YRI/ASN	23	50	786	844	21565
YRI/MEX	15	33	620	654	14716
CEU/ASN	10	24	445	474	10641
CEU/YRI	13	28	598	631	16405
CEU/MEX	0	1	15	16	510

*this total accounts for overlap between annotations

Table 3. Number of distinct genes affected by differentiated SNPs

Population Comparison	SIFT (Deleterious)	PolyPhen2 (Possibly or Probably Damaging)	Significant eQTL	Total Genes Affected*
ASN/MEX	5	12	127	141
YRI/ASN	24	49	610	663
YRI/MEX	17	31	444	481
CEU/ASN	9	24	260	285
CEU/YRI	15	26	455	489
CEU/MEX	0	1	15	16

*this total accounts for overlap between annotations

Using this collection of genes associated to differentiated SNPs through functional annotations, we performed gene enrichment analysis to identify specific biological mechanisms that likely have altered function between ethnic groups. This analysis revealed multiple pathways showing differences between CEU and MEX and CEU and ASN populations. KEGG pathways showing significant adjusted p-values ($p < 0.05$) are shown in Table 4.

Notably, the most significantly enriched pathways between CEU and MEX indicate a dramatic difference in the functional properties of glutathione and drug metabolism through cytochrome P450. Enrichment of these three pathways is the result of a single SNP – rs1010167 -- altering expression of three genes, *GSTM1*(p=3.88e-7), *GSTM2*(p=1.54e-7), and *GSTM4*(p=8.44e-7)[11]. This SNP falls within a region of chromatin that has been functionally categorized as an active promoter by the analysis of Ernst et al. in multiple cell types [12], and is confirmed to bind multiple proteins via ChIP-seq data as reported by the HaploREG database [13].

Table 4. Pathways with significant enrichment for highly differentiated functional alleles.

Population Comparison	KEGG Pathway	Reference Genes	Observed Genes	Expected Genes	P-value	P-value (adjusted for multiple testing)
CEU/MEX	Glutathione metabolism	24	3	0.04	1.02E-05	9.47E-05
CEU/MEX	Metabolism of xenobiotics by Cytochrome P450	30	3	0.06	2.03E-05	9.47E-05
CEU/MEX	Drug metabolism - Cytochrome P450	29	3	0.05	1.83E-05	9.47E-05
CEU/ASN	Allograft rejection	26	6	0.83	0.0001	0.0007
CEU/ASN	Graft-versus-host disease	22	6	0.7	4.70E-05	0.0007
CEU/ASN	Systemic lupus erythematosus	54	9	1.71	4.35E-05	0.0007
CEU/ASN	Arginine and proline metabolism	17	5	0.54	0.0001	0.0007
CEU/ASN	Autoimmune thyroid disease	26	6	0.83	0.0001	0.0007
CEU/ASN	Antigen processing and presentation	29	6	0.92	0.0002	0.0013
CEU/MEX	Asthma	17	2	0.03	0.0004	0.0014
CEU/ASN	Type I diabetes mellitus	30	6	0.95	0.0003	0.0016
CEU/MEX	Intestinal immune network for IgA production	24	2	0.04	0.0009	0.0018
CEU/MEX	Type I diabetes mellitus	30	2	0.06	0.0013	0.0018
CEU/MEX	Allograft rejection	26	2	0.05	0.001	0.0018
CEU/MEX	Graft-versus-host disease	22	2	0.04	0.0007	0.0018
CEU/MEX	Autoimmune thyroid disease	26	2	0.05	0.001	0.0018
CEU/MEX	Antigen processing and presentation	29	2	0.05	0.0013	0.0018
CEU/ASN	Intestinal immune network for IgA production	24	5	0.76	0.0008	0.0039
CEU/ASN	Riboflavin metabolism	8	3	0.25	0.0016	0.007

Remaining pathways showing high differentiation in the CEU/ASN and CEU/MEX comparisons are largely immune-related, and are driven mostly by functional changes to the Major Histocompatibility Complex (MHC) found on chromosome 6. Interestingly, there were no significant pathways found for differentiated functional SNPs involving YRI comparisons.

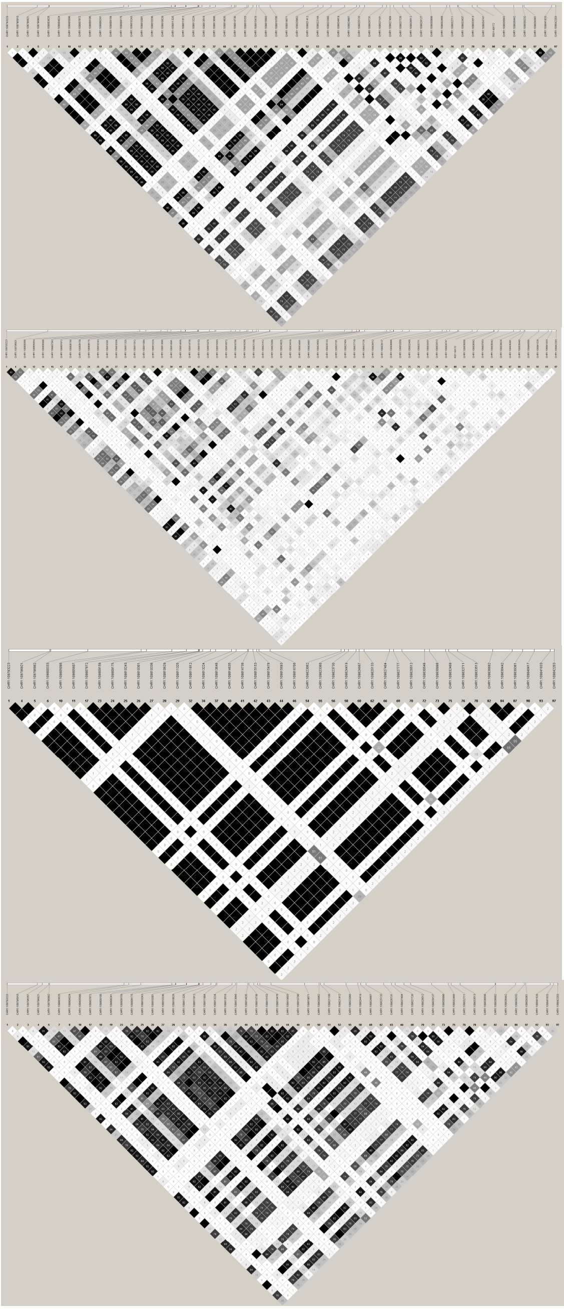


Figure 3. Extent of linkage disequilibrium (r^2) for 50kb region targeted by the Metabochip containing genome-wide association study (GWAS)-identified *CELSR2/PSCR1/SORT1* by HapMap population. Pair-wise linkage disequilibrium (LD) was calculated on unrelated samples using HaploView for European-descent [a)CEU; n=60], African [b)YRI; n=60], Asian [c)ASN; n=88], and Mexican [d)MEX; n=60] HapMap populations. For each LD plot, the genetic variants are labeled by chromosomal position at the top from 5′ to 3′. Each square represents a pair-wise LD statistic and they are coded on a gray scale where black is perfect LD (r^2=1) and white to gray is weak LD. The numbers in select squares represent the LD metric for that pair-wise comparison (for example, 1 is r^2=0.01).

To illustrate the fine-mapping potential of densely targeted regions on the Metabochip, we calculated linkage disequilibrium (r2) by HapMap population for the CELSR2/PSRC1/SORT1 locus known to be associated with low-density lipoprotein cholesterol levels from GWA studies in European-descent populations[14-16]. Consistent with the observations of Buyske et al[17] in samples from African American and Swedish participants, we observed less LD in YRI compared with CEU for this genomic region. To extend the observations made by Buyske et al, we examined LD for the same genomic region in HapMap samples of Asian and Mexican ancestry (Figure 3 c,d). As observed with minor allele frequency and FST, the CEU and MEX populations displayed similar levels of LD for this genomic region. In contrast, the ASN population had LD patterns that were distinct from CEU, YRI, and MEX LD patterns. For the ASN population, the CELSR2/PSRC1/SORT1 locus contained strong pair-wise LD statistics punctuated by weak LD.

4. Conclusions

We demonstrate here that the Metabochip custom BeadChip produces high-quality data for diverse populations from the International HapMap Project. We further show that the majority of variants observed in all populations considered were common and that a sizeable fraction of variants were monomorphic. Finally, we demonstrate population differences in both allelic diversity and LD patterns, both of which will impact the effectiveness of fine-mapping efforts that employ this BeadChip in the post-GWAS era.

Many of the observations reported here were expected based on population genetics theory and recent empirical genome-wide data from the International HapMap Project[18, 19] and 1000 Genomes Project[20]. That is, as expected, the greatest population differentiation (as measured by F_{ST}) was observed between African-descent and Asian-descent populations[21]. However, other observations such as the proportion of common and rare variants did not follow expectations given the bias in genetic variant selection for this custom BeadChip[22]. From our FST analysis, we also observe significant differentiation of functional alleles within drug metabolism and auto-immune associated pathways between CEU and ASN/MEX populations. These variants may explain some aspects of ethnic differences in HLA-based autoimmune disease susceptibility, and indicates that cytochrome P450 drug metabolism may be altered in individuals of Mexican ancestry.

A major limitation of this study is sample size. With only 60 to 90 independent samples per HapMap population, our ability to observe rare alleles targeted by the Metabochip was limited for any HapMap population. Indeed, although the shape of the allelic distribution was similar, proportionally more variants in our dataset were classified as common or monomorphic compared with Buyske et al reflecting our limited ability to observe rare variants. Larger sample sizes will be required to take advantage of the full range of the allelic spectrum targeted by the Metabochip for fine mapping.

A final observation made here that will impact fine-mapping efforts is the extent of LD for an LDL-C associated region across populations. As Buyske et al[2] noted, the breakdown of LD in African Americans for this region (and West Africans here) will be useful in identifying the true risk variant in a region with high LD in European populations. However, we note in ASN that the same genomic region has very high

LD and thus this custom BeadChip may not fine map equally well for all targeted GWAS-identified regions for all populations. Because this custom BeadChip was designed using early iterations of the 1000 Genomes Project data, additional iterations of chips designed for fine mapping will be required to capture the latest genetic diversity data now emerging in non-European descent populations from later releases of the 1000 Genomes Project.

Acknowledgements

This work was supported by NIH U01 HG004798 and its ARRA supplements. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work.

References

1. Matise TC, Ambite JL, Buyske S, Carlson CS, Cole SA, Crawford DC, Haiman CA, Heiss G, Kooperberg C, Marchand LL, Manolio TA, North KE, Peters U, Ritchie MD, Hindorff LA, and Haines JL (2011) The Next PAGE in Understanding Complex Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. American Journal of Epidemiology 174 (7):849-859
2. Buyske S, Wu Y, Carty CL, Cheng I, Assimes TL, Dumitrescu L, Hindorff LA et al (2012) Evaluation of the Metabochip Genotyping Array in African Americans and Implications for Fine Mapping of GWAS-Identified Loci: The PAGE Study. PLoS ONE 7 (4):e35651
3. Center for Statistical Genetics. MetaboChip SNP details. University of Michigan . 2012. 7-26-2012
4. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, and Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analysis. Am J Hum Genet 81 (3):559-575
5. Grady, B. J., Torstenson, E., Dudek, S. M., Giles, J., Sexton, D., and Ritchie, M. D. Finding unique filter sets in PLATO: a precursor to efficient interaction analysis in GWAS data. Pac Symp Biocomput , 315-326. 2010.
6. Weir, B. S. and Cockerham, C. C. Estimating F-statistics for the analysis of population structure. Evolution 38(1358), 1370. 1984.
7. Barrett JC, Fry B, Maller J, and Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21 (2):263-265
8. Ng PC and Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11 (5):863-874
9. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7 (4):248-249
10. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, and Pritchard JK (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4 (10):e1000214
11. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de GA, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, and Dermitzakis ET (2007) Relative impact of

nucleotide and copy number variation on gene expression phenotypes. Science 315 (5813):848-853

12. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, and Bernstein BE (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473 (7345):43-49
13. Ward LD and Kellis M (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 40 (Database issue):D930-D934
14. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC et al (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40 (2):161-169
15. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466 (7307):707-713
16. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, Cooper GM et al (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40 (2):189-197
17. Buyske S, Wu Y, Carty CL, Cheng I, Assimes TL, Dumitrescu L, Hindorff LA et al (2012) Evaluation of the Metabochip Genotyping Array in African Americans and Implications for Fine Mapping of GWAS-Identified Loci: The PAGE Study. PLoS ONE 7 (4):e35651
18. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437 (7063):1299-1320
19. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449 (7164):851-861
20. (2010) A map of human genome variation from population-scale sequencing. Nature 467 (7319):1061-1073
21. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467 (7311):52-58
22. Keinan A and Clark AG (2012) Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants. Science 336 (6082):740-743

INSIGHTS INTO DISEASES OF HUMAN TELOMERASE FROM DYNAMICAL MODELING

SAMUEL COULBOURN FLORES
Cell and Molecular Biology Department, Uppsala University, Biomedical Center, Box 596, 75124 Uppsala, Sweden
Email: samuel.flores@icm.uu.se

GEORGETA ZEMORA
Max F. Perutz Laboratories, University of Vienna, Dr. Bohrgasse 9/5, 1030 Vienna, Austria

CHRISTINA WALDSICH
Max F. Perutz Laboratories, University of Vienna, Dr. Bohrgasse 9/5, 1030 Vienna, Austria

Mutations in the telomerase complex disrupt either nucleic acid binding or catalysis, and are the cause of numerous human diseases. Despite its importance, the structure of the human telomerase complex has not been observed crystallographically, nor are its dynamics understood in detail. Fragments of this complex from Tetrahymena thermophila and Tribolium castaneum have been crystallized. Biochemical probes provide important insight into dynamics. In this work we summarize evidence that the T. castaneum structure is Telomerase Reverse Transcriptase. We use this structure to build a partial model of the human Telomerase complex. The model suggests an explanation for the structural role of several disease-associated mutations. We then generate a 3D kinematic trajectory of telomere elongation to illustrate a "typewriter" mechanism: the RNA template moves to keep the end of the growing telomeric primer in the active site, disengaging after every 6-residue extension to execute a "carriage return" and go back to its starting position. A hairpin can easily form in the primer, from DNA residues leaving the primer-template duplex. The trajectory is consistent with available experimental evidence. The methodology is extensible to many problems in structural biology in general and personalized medicine in particular.

1. Introduction

Telomere maintenance has broad implications for human health and longevity. Positive lifestyle changes including exercise and smoking cessation result in increased telomerase activity in the immune system.(1) Cancer cells also have vigorous telomerase activity (1) as required for immortality. (2) Despite its importance, the human Telomerase complex has not been solved crystallographically. However various components of the complex from other organisms are available. We use these to predict the structure of key components of human Telomerase by a novel homology modeling approach and illustrate its functional mechanism. We also give a possible explanation for the role of several disease associated mutations.

The primary enzymatic task of Telomerase is the extension of the leading telomeric DNA strand following the sequence of an RNA template. This is done one DNA residue at a time at the active site of the Reverse Transcriptase (RT). The RT domain shares motifs with RNA and DNA polymerases, suggesting similar function and mechanism.(3) The RNA Binding Domain (RBD) appears to help position the primer-template duplex for extension.(3) The carboxy-terminal extension (CTE) is implicated in DNA binding.(3)

In this work we focus on the RT, RBD, and the primer-template duplex. Key regions of RBD and most of the RT are conserved across organisms. In particular a published multiple sequence alignment (MSE)(4) shows considerable sequence identity between human and *T.castaneum* ("beetle") in a substantial portion of TERT. A crystallographic structure of beetle TERT with a putative primer-template duplex now exists, but considerable debate exists as to whether it is truly telomerase;(5) in this work we present the evidence that it is. We then use the beetle structure as the template for a partial homology model of human Telomerase. The model explains the probable role of disease associated residues and is consistent with biochemical probes of structure and dynamics.

Considerable insight into the dynamics of primer elongation exists from biochemical experiments. The primer-template duplex is always about 7 base pairs (bp) long as determined by dimethyl sulfate footprinting assays in yeast,(6) because bp's are denatured at the distal end as they form at the proximal end. In human, the template region is 6 bases long while an alignment region on the 3' side adds 5 bases for a total of 11; however it is believed that 11 base pairs would never form simultaneously during elongation as subsequently denaturing multiple base pairs would require too much energy. (7) Dissociation would also be difficult for an excessively long duplex. (7)

When the extending primer reaches the 5' end of the template, the template disengages and reattaches to the primer having shifted by six residues, ready for another six-residue extension to be be added (6). Meanwhile as DNA residues exit the primer-template duplex, they queue to join a hairpin or quadruplex which may help drive processivity.(8)

The N-Terminal domain (TEN) is a low-conservation region of TERT (3) known to be important for primer positioning and elongation. (9) *T.Castaneum* telomerase has no TEN domain, a point we will return to.
Additional insight comes from prior structural modeling. The TEN, RBD, RT and primer-template duplex domains were ambitiously predicted by homology modeling and docking by Steczkiewicz and collaborators. (10) However the primer-template duplex in that model is about 15 (10) rather than 7 bp long as it is in yeast (6) and is even longer than the 11 bp discussed above. Also, TEN residues 170-175 which have been experimentally implicated in the active site (9) are about 9 RNA (not DNA) nt away from the active site in the model. (10) That model also includes the CTE, which we did not model due to the low sequence identity between beetle and human in that domain. Lastly, the mechanism of processivity proposed in that work is based on a low-order normal mode expansion. The RBD and RT domains of that model, on the other hand, agree with those presented in this work.

In this work we thus present a knowledge-based structural and dynamical model of human telomerase, including much of TERT, the template, and a telomeric extension. We generate the structural model of TERT by homology modeling. The dynamical model incorporates additional biochemical information. We address the debate on the function of the beetle structure. (11) The results provide an explanation for the role of various disease associated mutations. Our model also supports a role in processivity for the primer extension hairpin. We show that MMB (formerly RNABuilder) (12) is a structural and dynamical modeling code with many potential applications in molecular biology. Its economy, versatility, and ease of use make it a good tool to use for examining the effect of individual genetic variation on disease phenotype.

2. Methods

2.1 *Validating human-beetle sequence alignment using secondary structure*

Generating the structure of the human TERT by threading to the existing beetle TERT structure (13) requires a multiple sequence alignment, which is available from the telomerase database.(4). Note that not all domains have sufficient sequence identity for alignment. As a first validation of the alignment, we predicted the secondary structure of the human TERT using the Jpred server, which is not biased by the use of the known *T.thermophila* or beetle TERT structures.(14)

2.2 *Aligning T. thermophila RNA Binding Domain to beetle TERT*

For further validation of the MSE and to support telomerase function of the beetle structure, we used the former as the basis for a rigid-body structural alignment (15,16) of the crystallographically observed *T.thermophila* RNA Binding Domain (RBD) (13) onto the beetle TERT. (11)

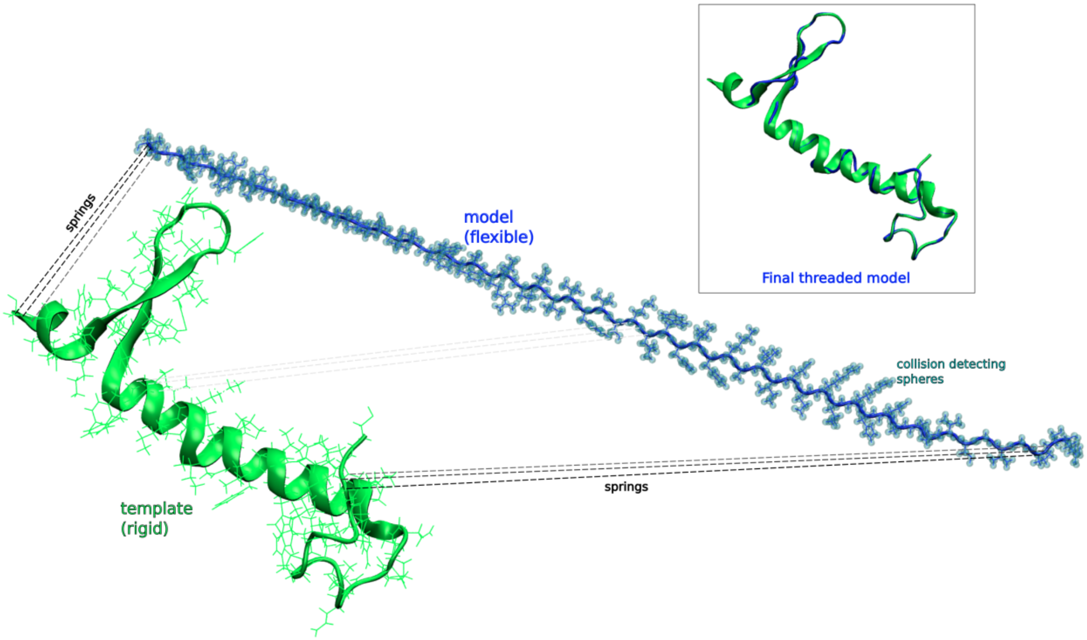


Figure 1. Illustration of internal coordinate threading procedure.
Threading is done as follows. The template (here a fragment of the beetle RBD is used, in green) is made rigid. In the model (here the corresponding human peptide fragment, in blue), bond lengths and angles are fixed, but torsion angles are free to vary (except for proline which has additional freedom for ring closure). Springs connect backbone atoms in the template with backbone atoms in the model which correspond according to sequence alignment (a representative set are shown as black dashed lines). Collision detecting spheres (transparent cyan) prevent steric clashes within the model. The system dynamics are allowed to proceed until the backbones are aligned. The final threaded model (inset) has a backbone RMSD of 1.002Å with respect to the template in this example.

2.3 *Threading human to beetle TERT*

In prior work we showed how MMB/RNABuilder can be used for RNA threading. (16) In this work we show that the package can also do protein threading.(17) The threading forces can be combined with other forces, constraints, and coordinate matching features available in this full-featured modeling package. We aligned a flexible human TERT protein backbone to the rigid beetle TERT(11) template by connecting corresponding backbone atoms with springs, while steric clashes were economically prevented by means of collision detecting spheres.(15) The approach of using internal coordinate dynamics(18) to align the backbone in this way is unique to this work;(17) most other protein threading algorithms work by rigid fragment assembly, segment matching, spatial restraint, and artificial evolution.(19) The approach is preferred for this work because it is economical, gives us full control over the alignment, conserves chemistry and sterics, allows protein, DNA, and RNA(16) to be threaded simultaneously, and permits dynamical (16) rather than only static modeling of the mechanism of primer extension as we will show. An illustrative example of this process is shown in Figure 1 above.

The basis for the correspondence for all threaded TERT fragments was the Telomerase Database MSE.(4) Note that much of TERT is highly diverged and therefore not all residues are aligned (Figure 2). (4).

```
525                  545 (RBD)         563    570   576
GVGCVPAAEHRLREEILAKF LHWLMSVYVVELLRSFFYVTETTFQKNRLFFYRKSVWSKLQSIGIRQHLKRVQLREL
------HHHHHHHHHHHHHH HHHHH--HHHHHH---EEEEEE----EEEEEEHHHHHHHHHHHHHHHHHH-------
YDAIPWLQNVEPNLRPKLLL HNLFLLDNIVKPIIAFYYKPIKTLNGHEIKFIRKEEYISFESKVFHKLKKMKYLVEV
---HHHHH-----HHHHHHH HHHHHHHHHHHHHHHHHEEEEE------EEEEHHHHHHHHHHHHHHHHHHH--EEE-

602   617                    648                     681                704
SEAEV LTSRLRFIPKPDGLRPIVNMDY EKRAERLTSRVKALFSVLNYERA LGLDDIHRAWRTFVLRVR PELYFVKVD
-HHHH ---EEEEEE----EEEEEE--- --HHHHHHHHHHHHHHHHHHHHH ---HHHHHHHHHHHHHHH --EEEEEE-
QDEVK PRGVLNIIPKQDNFRAIVSIFP DSARKPFFKLLTSKIYKVLEEKY KTSGSLYTCWSEFTQKTQ GQIYGIKVD
----- --EEEEEEE----EEEEEEE-- ---HHHHHHHHHHHHHHHHHHHH ------HHHHHHHHHHH- ---EEEEEE

713
VTGAYDTIPQDRLTEVIASIIKP
---EE----HHHHHHHHHH----
IRDAYGNVKIPVLCKLIQSIPTH
E-------HHHHHHHHHH-----

803     811                                                   865 869
SGLFDVFLRFMCHHAVRIRGKSYVQCQGIPQGSILSTLLCSLCYGDMENKLFAGIRRDGLLLRLVDDFLLVTPHLTH
HHHHHHHHHHHH---EEE---EEEEE--------HHHHHHHHHHHHHHHHHH-------EEEEEE--EEEEE--HHH
SEKKNFIVDHISNQFVAFRRKIYKWNHGLLQGDPLSGCLCELYMAFMDRLYFSNLDKDAFIHRTVDDYFFCSPHPHK
HHHHHHHHHHHHHEEEE----EEEE---------HHHHHHHHHHHHHHHHHH-------EEEE---EEEEEE-HHHH

880      889        898 902
AKTFLRTL RGV        VNLRKTVVNFPVEDEAL
HHHHHHHH HHH        E---EEEEE--------
VYDFELLI KGV        VNPTKTRTNLPTHRHPQ
HHHHHHHH HHH        EEEEEEEE---------
```

Key:

Black numbers: human residue number range of aligned fragment
Dark green: human sequence
Light green: predicted human secondary structure (Jpred)
Red: beetle sequence
Orange: observed beetle secondary structure

Figure 2. Alignment of human to beetle sequence and comparison of predicted human to observed beetle secondary structure. Note wide agreement between the two secondary structures, in particular in the RBD. Human residues of special interest include the active site (712,868,869, cyan highlight – note conservation), positively charged disease associated residues (570, 811, 865, 901, 902, green highlight – note only 865 and 902 are charged in beetle), and highest OPRA propensity from beetle (563, 576, yellow highlight).

As mentioned the beetle telomerase has no TEN. We found it difficult to dock the *Tetrahymena* TEN to the beetle TERT. The beetle TERT has a very crowded RT active site, with little room for an additional protein to be involved. We found no clear shape complementarity with the *T.Tetrahymena* TEN; significant structural rearrangement would be needed just to get the crucial TEN residues 170-175 near the 3' end of the primer. Our model for this reason does not include TEN.

2.4 *Predicting RNA binding interfaces on beetle TERT*

The correct position of the RNA component of Telomerase is unknown except for the location of the primer-template duplex (11). We therefore used OPRA (Optimal Protein RNA Area)(20) to predict points of high RNA-binding propensity on beetle TERT.

2.5 *Generating the primer elongation trajectory*

We used the available biochemical knowledge to generate a dynamical trajectory of elongation as follows.

In a preparatory stage, we threaded (16) the human Telomerase RNA (TR) template portion spanning residues 53 to 59 (18) onto the putative RNA component of the beetle primer-template duplex. (11) We attached a spring to pull the 5' end of the modeled template flanking region (residue 38) so as to pass near the predicted RNA binding hotspot in the T-domain. This was motivated by biochemical evidence suggesting the double-helical template boundary element (which includes residues 32-37) binds to the RBD. The T-domain is therefore one candidate for the location of the boundary element. We threaded seven residues of the modeled primer onto the DNA portion of the beetle duplex.

To model the addition of a single residue, we then shifted the primer by one residue in the 5' direction and attached a new residue at the 3' end. We shifted the template by one residue position in its 3' direction. We released the most distal base pair of the duplex. We repeated this process for a total of seven residues. Once sufficient primer residues exited the duplex, we enforced a stem-loop.

As a final step, following the addition of a nucleotide conjugate to the 5'-most template RNA residue, we released the template and reattached it 6 residue positions in the 5' direction, ready for another round of extension.

2.5 *Functional assay*

The plasmids pcDNA6hTERT and pBSU1hTR were used for transient transfections of HEK cells and were a kind gift from J. Lingner. The hTERT mutations K570N, R865C and K902N were generated in the pcDNA6 vector as well. Cell extracts and direct telomerase assays were done as described in (21) and (22), respectively.

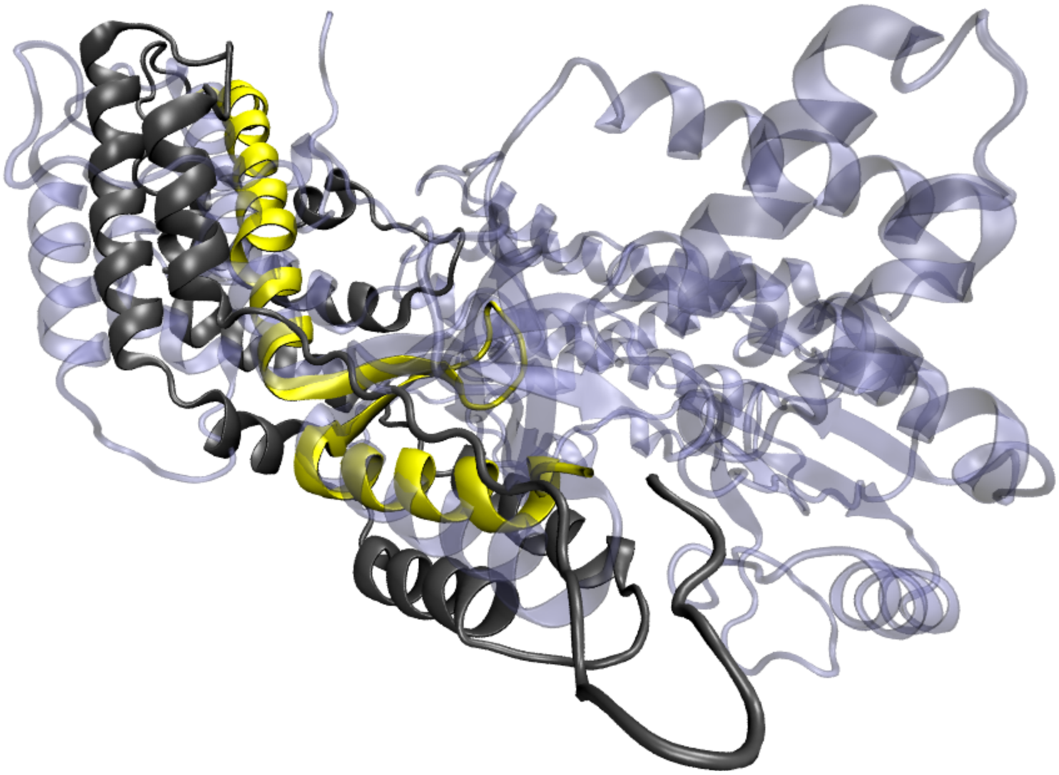
3. Results

3.1. *Validating human-beetle sequence alignment*

We compared to the predicted human secondary structure(14) to the observed beetle secondary structure(11) and found that they agreed in 232 of 283 aligned residues (Figure 2).

3.2. *Aligning Tetrahymena RBD to beetle*

We found that the *T.thermophila* RBD T-domain was aligned incorrectly in the telomerase database(4) as evidenced by a lack of secondary structure correspondence and the failure of the sequence alignment to guide correct 3D alignment between beetle and *tetrahymena* TERT. Once we corrected this the secondary structure aligned properly and we were able to align structurally (Figure 3)



```
CVPAAEHRLREEILAKF-LHWLMSVYVVELLRSFFYVTETTFQKNRLFFYRKSVWSKLQSIGIRQLKRVQLREL
---HHHHHHHHHHHHHH HHHHH--HHHHHH---EEEEEE----EEEEEEEHHHHHHHHHHHHHHHHH------
IPWLQNVEPNLRPKLLLKHNLFLLDNIVKPIIAFYYKPIKTLNGHEIKFIRKEEYISFESKVFHKKKMKYLVEVQDEVK
HHHHH-----HHHHHHHHHHHHHHHHHHHHHHHHEEEEEE------EEEEHHHHHHHHHHHHHHHHHH--EEE------
TQKRKYYISDKRKILGDLIVFIINKIVIPVLRYNFYITEKHKEGSQIFYYRKPIWKLVSKLTIVKLEE
-HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEEEEE-------EEEEHHHHHHHHHHHHHHHHHH

Dark green:  human sequence
Light green: human predicted secondary structure (Jpred)
Red: Beetle sequence
Orange: Beetle observed secondary structure
Yellow highlight: Tetrahymena sequence
Gray:  Tetrahymena observed secondary structure
```

Figure 3. Alignments of the T-motif

Upper panel: Structural alignment of Tetrahymena T-motif (yellow) and remainder of Tetrahymena TRBD (gray). Beetle TERT is shown in transparent gunmetal. Rigid structural alignment is based on the known beetle-Tetrahymena sequence alignment. The T-motif is a unique fold, and the close structural alignment is evidence that the beetle structure is TERT. Therefore, we should be able to predict human structure based on alignment to beetle, in regions of high sequence identity.

Lower panel: Sequence alignment of the RNA Binding Domain T-motif. Note that previously published alignments are incorrect for Tetrahymena in this region.

3.3. *RNA binding interfaces on beetle TERT*

OPRA predicted a strong binding propensity in the RBD, in particular at the base of the β-hairpin. It also predicted a second binding region in the C-terminal domain (Figure 4) recapitulating its known DNA binding function.(3)

3.4. *Basic disease associated residues near the primer-template duplex*

A novel feature of MMB called "Physics where you want it" allows flexibilizing a selection of residues, leaving the remainder of the system fixed and rigid. An MD (in this case Amber99) force field is turned on for another (typically larger and enclosing) selection of residues. We flexibilized residues 570, 902, and 865 and turned on physics for those residues and for the primer-template duplex. We found that residues 570 and 902 rapidly gravitated towards the duplex, settling within 2Å of the DNA strand at the closest point. 901 was in a similar position to 902. Residue 865 came within 6Å and 8Å, respectively, of the RNA and DNA strands (Figure 4).

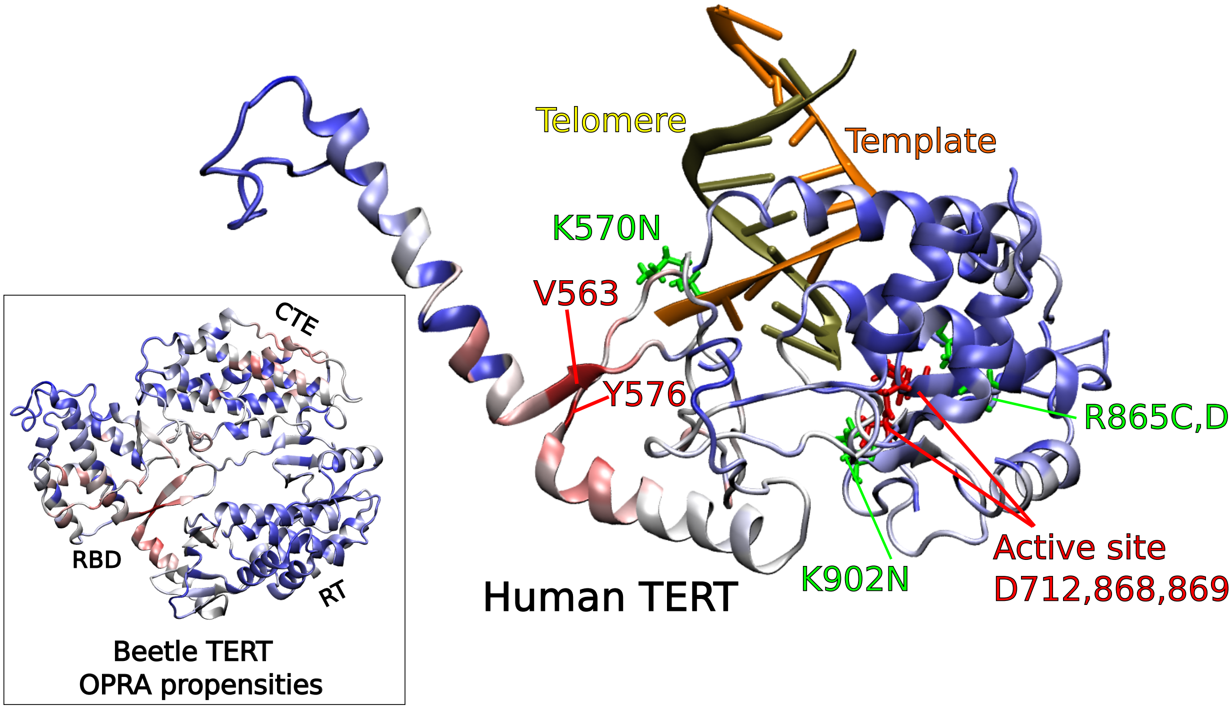


Figure 4. Human TERT charged disease-associated residues and OPRA RNA binding propensities in the vicinity of the duplex.

Labeled green residues are disease associated. Residue 570 lies very close to the primer backbone and is very likely to coordinate it. Residues 865 and 902 can come within a few Ångströms of the primer strand, and are also close to the active site residues (conserved ASP 712,868, and 869).

OPRA propensities were calculated in beetle TERT (inset) and then transferred to human. Highest propensities (red shading) were in the C-terminal extension (or Thumb) and RBD (see human residues 563 and 576). Lowest propensities are in blue shading.

3.5. *Basic beetle residues near the primer-template duplex*

For the beetle structure (PDB accession: 3KYL) we found that of 20 protein residues within 5Å of the 7 DNA residues in the 7-bp duplex, seven (144,194,406,416,418,437,477) were basic. Further, they were mostly in range to make contact. On the other hand, we found that 30 protein residues were within 5Å of the 7 RNA residues in the primer-template duplex. Of the 30, only two (206,210) were within the 5Å range, but at its very edge and apparently interacting with a portion of the DNA strand outside of the 7-bp duplex (Figure 5).

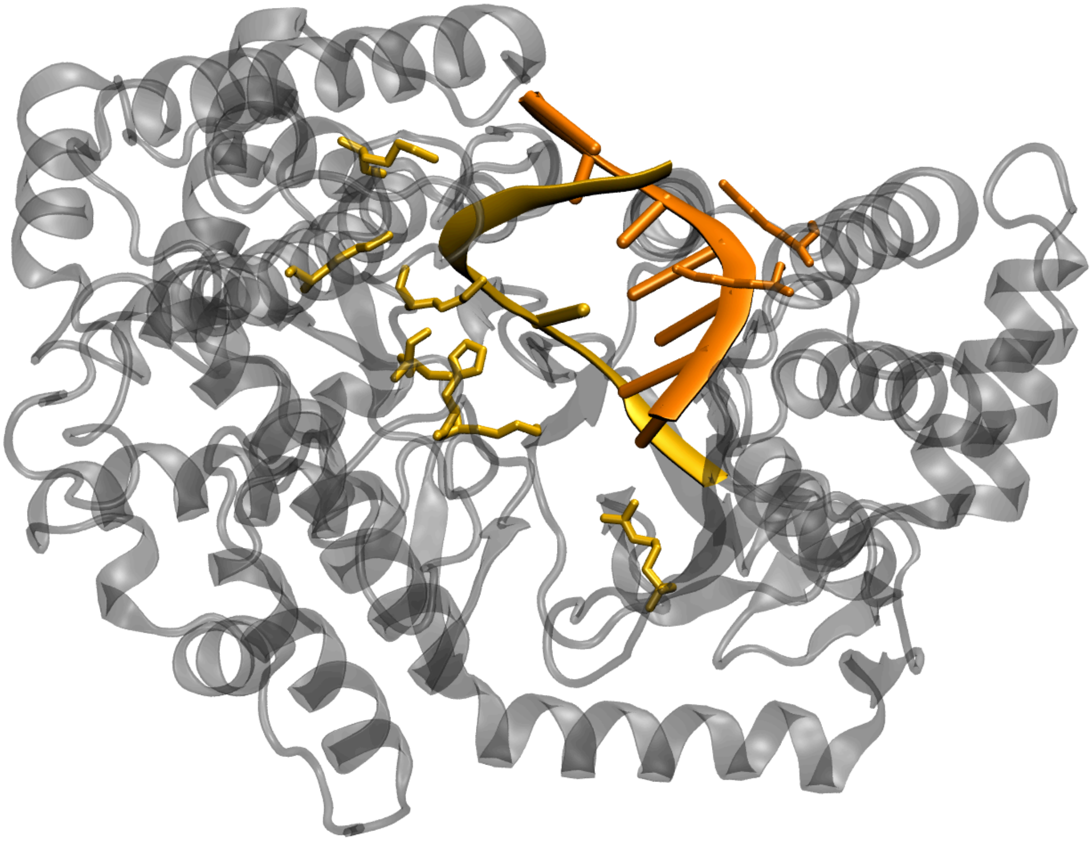


Figure 5 Basic residues within 5Å of the template-primer duplex in T. Castaneum.

Seven basic amino acid residues (gold) coordinate the seven primer residues (also gold) of the duplex. Two additional basic residues (orange) are within 5Å of the template portion of the duplex (also orange), but appear to be coordinating a portion of the primer outside the duplex. Thus all residues within 5Å of the duplex In beetle are binding the primer, not the template.

3.5. *Functional assay*

The hTERT variants carrying mutations K570N, R865C, K902N had been identified in patients with DC, AA or IPF (23-25). Measuring the activity of these mutant telomerase complexes revealed that the K570N and K902N variants showed no telomerase activity, while R865C showed a 20% telomerase activity (25). We now attempted to rescue this strong phenotype by increasing the telomeric primer concentration 10-fold in the extension assay; however this did not significantly improve the activity of mutant telomerases (data not shown).

3.6. *The dynamical trajectory of primer elongation*

We noted several possible routes for the 5' template flanking region to exit the primer-template duplex. The route used in our trajectory follows the β-hairpin, contacting the RNA binding hotspot at its base, and takes a direction tangent to its duplex portion (supplementary materials).

There is sufficient space to form a harpin in bursts as sufficient residues emerge from the primer-template duplex. We note that one or more hairpins may form, each with 8 base pairs in the stem. Alternatively, a single long hairpin may form, with its end loop migrating to 3' in six-residue increments as the primer elongates, or some combination of the two may occur.

4. Discussion

4.1. *Validating human-beetle sequence alignment*

Sequence identity alone does not necessarily imply structural homology; structural divergence is possible. In particular controversy surrounds the beetle structure used, with some authors pointing out that it may not be TERT(5). We present evidence that the beetle-human sequence alignment is correct, that the beetle structure is in fact TERT, and that there is a basis for homology modeling.

First, the beetle – *T.thermophila* sequence alignment was used as the basis for aligning the less controversial *T.thermophila* RBD to its putative beetle counterpart. The key conserved T-motif, a β-hairpin with flanking α-helices aligns well (Figure 4). Since this motif has not been observed outside of TERT to our knowledge, this observation indicates that the beetle structure is likely to be TERT, and that the sequence alignment is correct in this region.

Second, we ran the Jpred secondary structure prediction algorithm on the human sequence and found that the predicted human secondary structure matches the observed beetle secondary structure for 232 of 283 residues (Figure 2). Jpred(14) is not biased by the use of the beetle or *T.thermophila* TERT structures (C. Cole, personal communication). Therefore there is probably considerable structural homology in the aligned regions. The human sequence and predicted secondary structure also aligns to that of *T.thermophila* in the RBD (Figure 3), further bolstering the structural and functional correspondence (non-telomerase reverse transcriptases do not have this RBD).

As mentioned the TEN domain has been directly implicated in the active site (9) and yet is absent in beetle. We speculate that the role of the human TEN residues involved in elongation may be played in beetle by residues in the RT domain, leading to a crowded active site with little room for an additional protein subunit. This may explain why it is difficult to dock the *Tetrahymena* TEN to the beetle TERT.

4.2. *RNA binding interfaces on beetle TERT*

The results of the OPRA RNA binding interface prediction strengthen the case that the beetle structure is TERT. The binding hotspot in the CTE is consistent with its role in DNA binding.(3) The strong signal at the base of the β-hairpin confirms that this is the RNA Binding Domain.

4.3. *Basic beetle residues near the primer-template duplex*

The evidence of DNA binding being more important than RNA binding near the duplex prompted us to examine the beetle structure (3KIY) more closely. The finding that 7 of the 20 residues near the DNA part of the 7-bp duplex were basic, while only 2 of 30 the residues near the RNA part were basic, is strong indication that DNA binding is more crucial in this region (Figure 4). Experimentalists may find it profitable to mutate beetle TERT residues 144,194,406,416,418,437, and/or 477 to try to produce reduced DNA binding in vitro or, more interestingly, a beetle with a telomerase deficiency disease.

4.4. *Charged disease associated residues near the primer-template duplex*

Disease associated residue 570 is very likely in contact with the primer strand, a few residues from the terminus. Its charge, location in a highly conserved RNA binding region, and clear proximity to the DNA backbone make this prediction strong.

Residue 902 (and by extension 901) is also positioned to contact DNA. We verified this with a simple equilibration using MMB.

Residue 865 is spatially near the primer strand. It is also close in sequence to the active site residues, further suggesting a DNA binding role. However an equilibration using MMB, shows that 865 could contact either DNA or RNA without significant backbone motion.

4.5. *Insight into disease from structure*

Our rather conservative model leaves out diverged regions of TERT such as the CTE, and for that reason contains only five basic disease associated residues: 570, 811, 865, 901, and 902. Of these, all except 811 are within 9Å of DNA. A very limited "Physics where you want it" simulation showed that 865 is positioned such that it could also be contacting RNA, but that 570, 901, and 902 are very likely binding DNA only. 811 is about 19Å from the duplex. Our model therefore predicts a likely DNA-binding role for disease-associated residues 570, 901, 902, and perhaps 865. All basic residues within 9Å of the DNA in our model (with the exception of some at the periphery of that cutoff) are documented disease-associated residues, again suggesting that basic residues near the duplex have an essential DNA-binding function. If this model is correct, TERT with these mutations should have reduced affinity for DNA. A functional assay should show reduced activity, which should be recovered with a saturating concentration of DNA.

Inspection of the Steczkiewicz model also leads to interesting findings. Due to differences in the modeling approach the findings are overlapping but not identical. In particular, since their model included highly diverged regions, it includes many residues which ours did not. First, of the seven basic residues known to be disease-associated (486, 570, 811, 865, 901, 902, and 979), three (570, 865, and 979) are within 9Å of the DNA strand. Two of the latter (570 and 979) were also within 9Å of the RNA strand. Within 9Å of the DNA, there were 22 positively charged residues in total. In contrast, there were only 13 within the same radius of the RNA, again suggesting the greater importance of binding the stretch of DNA.

One interpretation of our computational model is that basic residues which coordinate the DNA strand are more likely to be essential, while those coordinating the RNA may be dispensable. This could be explained by the fact that while TERT has many points of contact with RNA, it has very few with DNA. In accordance with this idea, other basic residues within about 9Å of the DNA strand, when mutated, could be as-yet undiscovered causes of telomerase diseases. However in that case saturating concentration of DNA would be expected to recover function in mutants K507N, R965C and K902N, and it did not. This may mean that the residues contribute to the stability of TERT, and that a loss of charge leads to misfolding. Alternatively, it may be that the contribution of the residues to DNA binding is very strong, and we did not reach saturating conditions of DNA.

The Steczkiewicz model by our interpretation predicts that residues 499, 500, 570, 626, 631, 643, 646, 647, 649, 650, 710, 865, 955, 962, 968, 971, 972, 973, 979, 981, 983, and 1011 are within 9Å of the DNA part of the 7 bp duplex; accordingly clinicians should be on the lookout for mutations in these residues. Note that these include known disease-associated residues 570, 865, and 979. For costlier experimental assays, we suggest starting with residues within 5Å: 643, 649, 962, 972, 973, 970, 979, and 981). Note that the latter list includes known disease-associated residue 979. However we remind experimentalists that the Steczkiewicz model aligns highly diverged domains.

4.6. *Primer elongation movie*

We created a movie and figure (supplementary material) which shows a single cycle of primer elongation. One residue is added at a time, with the rest of the primer-template duplex shifting one residue position towards the distal end with each addition. Simultaneously, the distal base pair denatures.(6) Once sufficient residues emerge from the duplex, a hairpin forms.(8) In our model there is room for a series of such hairpins, or for a single long hairpin with and end loop that migrates to 3' in six-residue bursts, or for some combination of the two. We look forward to future workers elucidating how this is denatured to extend the lagging strand of the primer. At the end of one cycle of elongation, the template shifts six residues in the proximal direction, like an old-style typewriter carriage preparing to write another line.(6,26) The trajectory shows that this mechanism is sterically and geometrically feasible and consistent with existing structural and biochemical data. It further provides a structural basis for designing focused experiments to test specific steps of this process. We encourage other workers to modify the MMB/RNABuilder command file we provide to extend or modify our simulation or add more Telomerase subunits.

5. Conclusion

Considerable progress has been made on Telomerase structure and function, but this had not been turned into a 3D dynamical model. In this work we first presented the evidence that a recent *T.Castaneum* structure (3) is in fact TERT, addressing a topic of current debate. (11) We then built a threaded model of part of the RT and RBD domains and the primer-template duplex of human telomerase. We find that all four basic residues within 9Å of the primer-template duplex are disease associated, and further that at least three of them appear to be DNA binding. Similarly, we find that in the published beetle structure, basic residues cluster near the DNA and not the RNA strand of the duplex. This may suggest that several positively-charged, disease-associated residues are involved in coordinating DNA in human telomerase. We propose that such DNA-binding residues are more likely to be essential, whereas RNA-binding residues may be dispensable, since TERT has fewer points of contact with DNA than RNA. Our functional assay does not rule out the possibility that these residues also contribute to the stability of TERT. We used biochemical and biophysical results as constraints to generate a 3D kinematic model of primer extension as an illustration of this important process.

6. Distribution and supplementary materials

The telomere extension movie is distributed as a .mpg and a figure file at https://simtk.org/home/telomerase . The entire trajectory in .pdb format, the command file (in RNABuilder 2.2 syntax), and initial structure file are also available on request. RNABuilder 2.2 and more recent MMB distributions for Windows, OSX, and Linux are available for download from https://simtk.org/home/rnatoolbox .

7. Acknowledgements

We thank Juan Fernández-Recio and Laura Perez-Cano for running OPRA on the beetle TERT structure for us. S. Flores acknowledges funding from eSSENCE (essenceofscience.se). Joachim Lingner is acknowledged for his generous gift of plasmids. This work was supported by the Austrian Science Foundation (FWF) grant Y401 to C. Waldsich.

8. References

1. Sekaran, V.G., Soares, J. and Jarstfer, M.B. (2010) Structures of telomerase subunits provide functional insights. *Biochim Biophys Acta*, 1804, 1190-1201.
2. Kim, N.W., Piatyszek, M.A., Prowse, K.R., Harley, C.B., West, M.D., Ho, P.L., Coviello, G.M., Wright, W.E., Weinrich, S.L. and Shay, J.W. (1994) Specific association of human telomerase activity with immortal cells and cancer. *Science (New York, N.Y*, 266, 2011-2015.
3. Gillis, A.J., Schuller, A.P. and Skordalakes, E. (2008) Structure of the Tribolium castaneum telomerase catalytic subunit TERT. *Nature*, 455, 633-637.
4. Podlevsky, J.D., Bley, C.J., Omana, R.V., Qi, X. and Chen, J.J. (2008) The telomerase database. *Nucleic acids research*, 36, D339-343.
5. Wyatt, H.D., West, S.C. and Beattie, T.L. (2010) InTERTpreting telomerase structure and function. *Nucleic acids research*, 38, 5609-5622.
6. Forstemann, K. and Lingner, J. (2005) Telomerase limits the extent of base pairing between template RNA and telomeric DNA. *EMBO Rep*, 6, 361-366.
7. Hammond, P.W. and Cech, T.R. (1998) Euplotes telomerase: evidence for limited base-pairing during primer elongation and dGTP as an effector of translocation. *Biochemistry*, 37, 5162-5172.
8. Lue, N.F. (2004) Adding to the ends: what makes telomerase processive and how important is it? *Bioessays*, 26, 955-962.

9. Jurczyluk, J., Nouwens, A.S., Holien, J.K., Adams, T.E., Lovrecz, G.O., Parker, M.W., Cohen, S.B. and Bryan, T.M. (2011) Direct involvement of the TEN domain at the active site of human telomerase. *Nucleic acids research*, 39, 1774-1788.

10. Steczkiewicz, K., Zimmermann, M.T., Kurcinski, M., Lewis, B.A., Dobbs, D., Kloczkowski, A., Jernigan, R.L., Kolinski, A. and Ginalski, K. (2011) Human telomerase model shows the role of the TEN domain in advancing the double helix for the next polymerization step. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 9443-9448.

11. Mitchell, M., Gillis, A., Futahashi, M., Fujiwara, H. and Skordalakes, E. (2010) Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat Struct Mol Biol*, 17, 513-518.

12. Flores, S. and Altman, R. (2010) Turning limited experimental information into 3D models of RNA *RNA (New York, N.Y*, 16, 1769-1778.

13. Rouda, S. and Skordalakes, E. (2007) Structure of the RNA-binding domain of telomerase: implications for RNA recognition and binding. *Structure*, 15, 1403-1412.

14. Cole, C., Barber, J.D. and Barton, G.J. (2008) The Jpred 3 secondary structure prediction server. *Nucleic acids research*, 36, W197-201.

15. Flores, S.C. (2011) RNABuilder 2.2 Tutorial.

16. Flores, S., Wan, Y., Russell, R. and Altman, R. (2010) Predicting RNA structure by multiple template homology modeling. *Proceedings of the Pacific Symposium on Biocomputing*, 216-227.

17. Wallner, B. and Elofsson, A. (2005) All are not equal: a benchmark of different homology modeling programs. *Protein Sci*, 14, 1315-1327.

18. Flores, S., Sherman, M., Bruns, C., Eastman, P. and Altman, R. (2010) Fast flexible modeling of macromolecular structure using internal coordinates. *IEEE Transactions in Computational Biology and Bioinformatics*, 8, 1247-1257.

19. Xiang, Z. (2006) Advances in homology protein structure modeling. *Curr Protein Pept Sci*, 7, 217-227.

20. Perez-Cano, L. and Fernandez-Recio, J. (2010) Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. *Proteins*, 78, 25-35.

21. Cristofari, G., Adolf, E., Reichenbach, P., Sikora, K., Terns, R.M., Terns, M.P. and Lingner, J. (2007) Human telomerase RNA accumulation in Cajal bodies facilitates telomerase recruitment to telomeres and telomere elongation. *Molecular cell*, 27, 882-889.

22. Cristofari, G. and Lingner, J. (2006) Telomere length homeostasis requires that telomerase levels are limiting. *The EMBO journal*, 25, 565-574.

23. Armanios, M., Chen, J.L., Chang, Y.P., Brodsky, R.A., Hawkins, A., Griffin, C.A., Eshleman, J.R., Cohen, A.R., Chakravarti, A., Hamosh, A. *et al.* (2005) Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 15960-15964.

24. Xin, Z.T., Beauchamp, A.D., Calado, R.T., Bradford, J.W., Regal, J.A., Shenoy, A., Liang, Y., Lansdorp, P.M., Young, N.S. and Ly, H. (2007) Functional characterization of natural telomerase mutations found in patients with hematologic disorders. *Blood*, 109, 524-532.

25. Tsakiri, K.D., Cronkhite, J.T., Kuan, P.J., Xing, C., Raghu, G., Weissler, J.C., Rosenblatt, R.L., Shay, J.W. and Garcia, C.K. (2007) Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 7552-7557.

26. Zaug, A.J., Podell, E.R. and Cech, T.R. (2008) Mutation in TERT separates processivity from anchor-site function. *Nat Struct Mol Biol*, 15, 870-872.

SPECTRAL CLUSTERING STRATEGIES FOR HETEROGENEOUS DISEASE EXPRESSION DATA[†]

GRACE T. HUANG[1,2,3], KATHRYN I. CUNNINGHAM[4], PANAYIOTIS V. BENOS[1,3],
AND CHAKRA S. CHENNUBHOTLA[1]

[1]Department of Computational and Systems Biology
[2]Joint CMU-Pitt PhD Program in Computational Biology
[3]Clinical and Translational Science Institute
University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[4]Department of Computer Science,
University of Arizona, Tucson, Arizona, USA

Clustering of gene expression data simplifies subsequent data analyses and forms the basis of numerous approaches for biomarker identification, prediction of clinical outcome, and personalized therapeutic strategies. The most popular clustering methods such as *K*-means and hierarchical clustering are intuitive and easy to use, but they require arbitrary choices on their various parameters (number of clusters for *K*-means, and a threshold to cut the tree for hierarchical clustering). Human disease gene expression data are in general more difficult to cluster efficiently due to background (genotype) heterogeneity, disease stage and progression differences and disease subtyping; all of which cause gene expression datasets to be more heterogeneous. Spectral clustering has been recently introduced in many fields as a promising alternative to standard clustering methods. The idea is that pairwise comparisons can help reveal global features through the eigen techniques. In this paper, we developed a new recursive *K*-means spectral clustering method (ReKS) for disease gene expression data. We benchmarked ReKS on three large-scale cancer datasets and we compared it to different clustering methods with respect to execution time, background models and external biological knowledge. We found ReKS to be superior to the hierarchical methods and equally good to *K*-means, but much faster than them and without the requirement for *a priori* knowledge of *K*. Overall, ReKS offers an attractive alternative for efficient clustering of human disease data.

[†] This work is supported by NIH grants R01LM009657, R01LM007994 (PVB), GM086238. GTH was supported by grant UL1RR024153 to the Center for Translational Science (CTSI), University of Pittsburgh School of Medicine. KIC was supported by the NSF REU award DBI-1004555.

1. Introduction

The explosion of gene expression and other data collection from thousands of patients of several diseases has created novel questions about their meaningful organization and analysis. The Cancer Genome Atlas (TCGA)[1] initiative for example provides large heterogeneous datasets from patients with different types of cancers including breast, ovarian and glioblastoma. However, unlike data from model organisms and cell lines that have uniform genetic background, and where experiments are conducted under controlled conditions, disease samples are typically much more heterogeneous. Differences in the genetic background of the subjects, disease stage, progression, and severity as well as the presence of disease subtypes contribute to the overall heterogeneity. Discovering genes or features that are most relevant to the disease in question and identifying disease subtypes from such heterogeneous data remains an open problem.

Clustering, the unsupervised grouping of data vectors into classes with similar properties is a powerful technique that can help solve this problem by reducing the number of features one has to analyze and by extracting important information directly from data when prior knowledge is not available. As such, it has formed the basis of many feature selection and classification methods[2,3]. Hierarchical and data partitioning algorithms (like *K*-means) have been used widely in many domains[4] including biology[5,6]. They have become very popular due to their intuitiveness, ease of use, and availability of software. Their biggest drawbacks come from the usually arbitrary selection of parameters, such as the optimal number of clusters (for *K*-means) or an appropriate threshold for cutting the tree (for hierarchical clustering).

When applied to datasets from model organisms and cell lines, these clustering approaches have been quite successful in identifying biologically informative sets of genes[5,6]. However, the heterogeneity of the disease samples hinders their efficiency in them. Figure 1 shows an example of such a dataset; a dendrogram produced from the breast cancer TCGA data, in comparison to dendrogram generated from the less heterogeneous yeast expression data. It is obvious that the structure of the data makes it difficult to find a threshold to prune the tree to produce a satisfactory number of clusters, since every newly formed cluster is joined with a singleton node each time. Thus, despite its popularity, classical hierarchical clustering frequently performs poorly in discovering a satisfactory group structure within gene expression data. Tight clustering[7] and fuzzy clustering[8] attempt to build more biologically informative clusters either by focusing only on closely related genes while ignoring the rest, or by allowing overlap in cluster memberships. However, both methods suffer from long execution times. Similarly, Affinity Progation[9] has been applied on gene clustering successfully but at a significant cost in execution time. .

More recently, spectral clustering approaches have been used for data classification, regression and dimensionality reduction in a wide variety of domains, and has also been applied to gene expression data[10]. The spectral clustering formulation requires building a network of genes, encoding their pairwise interactions as edge weights, and analyzing the eigenvectors and eigenvalues of a matrix derived from such a network. To our knowledge, no systematic attempt has been made to-date to test and compare the performance of existing clustering methods in large-scale disease gene expression data, perhaps due to unavailability of suitable size datasets. In this paper, we evaluate the standard *K*-means and hierarchical clustering methods on three large

TCGA datasets. The evaluation is performed using intrinsic measures and external information. We introduce ReKS (Recursive *K*-means Spectral clustering), and compare it to the two aforementioned methods on the TCGA data. ReKS leverages the global similarity structure that spectral clustering provides, while saving on computing time by performing recursion. At each recursion step, we exploit the distribution of eigenvalues to select the optimal number of partitions, thus eliminating the need for pre-specifying *K*. We show that ReKS is very useful in deriving important biological information from patient gene expression data. Furthermore, we show how to add prior information from KEGG pathway to refine the cluster boundaries.

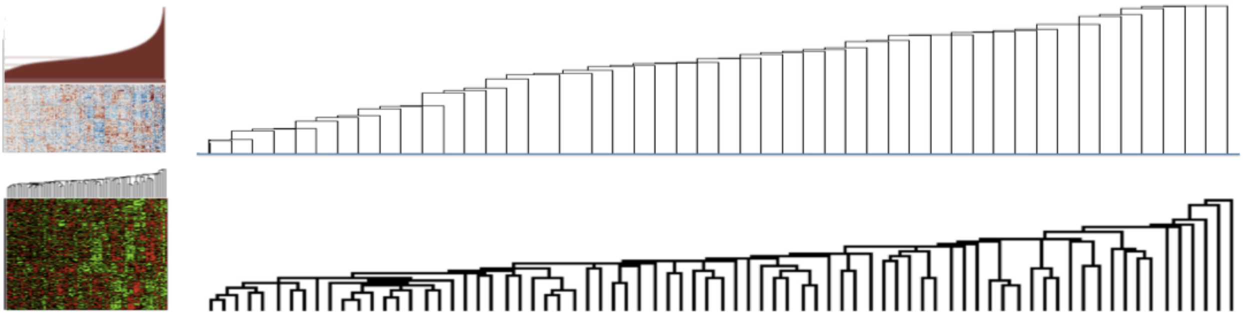


Fig. 1. Clustering patient data is more difficult than cell-based data. Partial views of dendrograms constructed from hierarchical clustering of the TCGA Breast Cancer expression data (top) and the yeast expression data (from Spellman *et al.*[11]). The dendograms suggest that it is easier to select a threshold to prune the tree and generate potentially meaningful clusters for the yeast data but not so for the breast cancer data.

2. Method

2.1. *Spectral Clustering*

The spectral clustering formulation requires building a network of genes, encoding their pairwise interactions as edge weights, and analyzing the vectors and eigenvalues of a matrix derived from such a network. This procedure is well established in the literature[12] so here we limit our discussion to the main points of the algorithm and use a Markov chain perspective to help us reason further about the idiosyncrasies of the algorithm when applied to cancer expression data.

A convenient framework for understanding the spectral method is to consider the partitioning of an undirected graph $G = (V, E)$ into a set of distinct clusters. Here the genes are represented as vertices v_i for $i = 1 \dots N$ where N is the total number of genes and network edges have weights w_{ij} that are non-negative symmetric ($w_{ij} = w_{ji}$) to encode the strength of interaction between a given pair of genes. Affinities denote how likely it is for a pair of genes to belong to the same group. Here we used as affinities a modified form of the correlation coefficient ρ_{ij}, calculated on the gene expression vectors:

$$w_{ij} = \exp\left(-\left(\sin\frac{arccos(\rho_{ij})}{2}\right)^2\right) \tag{1}$$

This is distance measure previously found to give empirical success in the clustering of gene expression data[9]. Note that high affinities correspond to pairs of genes that are likely to belong in the same group (e.g., participate in a pathway). In this paper, we ensured that the network is connected so that there is a path between any two nodes of the network. Our goal is to group genes into distinct clusters so that genes within each group are highly connected to each other, while genes in distinct clusters are dissimilar.

Spectral methods use local (pairwise) similarity (affinity) measurements between the nodes to reveal global properties of the dataset. The global properties that emerge are best understood in terms of a random walk formulation on the network[13–15] .The random walk is initiated by constructing a Markov transition matrix over the edge weights. Representing the matrix of affinities w_{ij} by W and defining the degree of a node by $d_j = \sum_i w_{ij}$, a Markov transition matrix M can be defined over the edge weights by

$$M = WD^{-1} \tag{2}$$

where D is a diagonal matrix stacked with degree values d_j. The transition matrix M can be used to set up a diffusion process over the network. In particular, a starting distribution p^0 of the Markov chain evolves to $p = M^{\beta}p^0$ after β iterations. As β approaches infinity, the Markov chain can be shown to approach a stationary distribution: $M^{\infty} = \pi\ 1^T$ is an outer product of 1 (a column vector of N 1s) and π (column vector of length N). It is easy to show that π is uniquely given by: $\pi_i = d_i/\sum_j d_j$ and is the leading eigenvector of M: $M\pi = \pi$ with eigenvalue 1.

We can analyze the diffusion process analytically by using the eigenvectors and eigenvalues of M. From an eigen perspective the diffusion process can be seen as[14]:

$$p^{\beta} = \pi + \textstyle\sum_2^n \lambda_j{}^{\beta} D^{0.5} u_j u_j{}^T D^{-0.5} p^0 \tag{3}$$

where the eigenvalue $\lambda_1 = 1$ is associated with stationary distribution π. The eigenvectors are arranged in decreasing order of their eigenvalues, so the second eigenvector u_2 perturbs the stationary distribution the most as $\lambda_2 \geq \lambda_k$ for $k > 2$. The matrix $u_2 u_2{}^T$ has elements $u_{2,i} \times u_{2,j}$, which means the genes that share the same sign in u_2 will have their transition probability increased, while transitions across points with different signs are decreased. A straightforward strategy for partitioning the network is to use the sign of the elements in u_2 to cluster the genes into two distinct groups.

Ng *et al*[16] showed how this property translates to a condition of piecewise constancy on the form of leading eigenvectors, i.e. elements of the eigenvector have approximately the same value with-in each putative cluster. Specifically, it was shown that for *K* weakly coupled clusters, the leading *K* eigenvectors of the transition matrix M will be roughly piecewise constant. The *K*-means spectral clustering method is a particular manner of employing the standard *K*-means algorithm on the elements of the leading *K* eigenvectors to extract *K* clusters simultaneously. We follow the recipe in Ng et al where instead of using a potentially non-symmetric matrix M, a symmetric normalized graph Laplacian $L = D^{-0.5}WD^{-0.5}$, whose eigenvalues and eigenvectors are similarly related to M, is used for partitioning the graph.

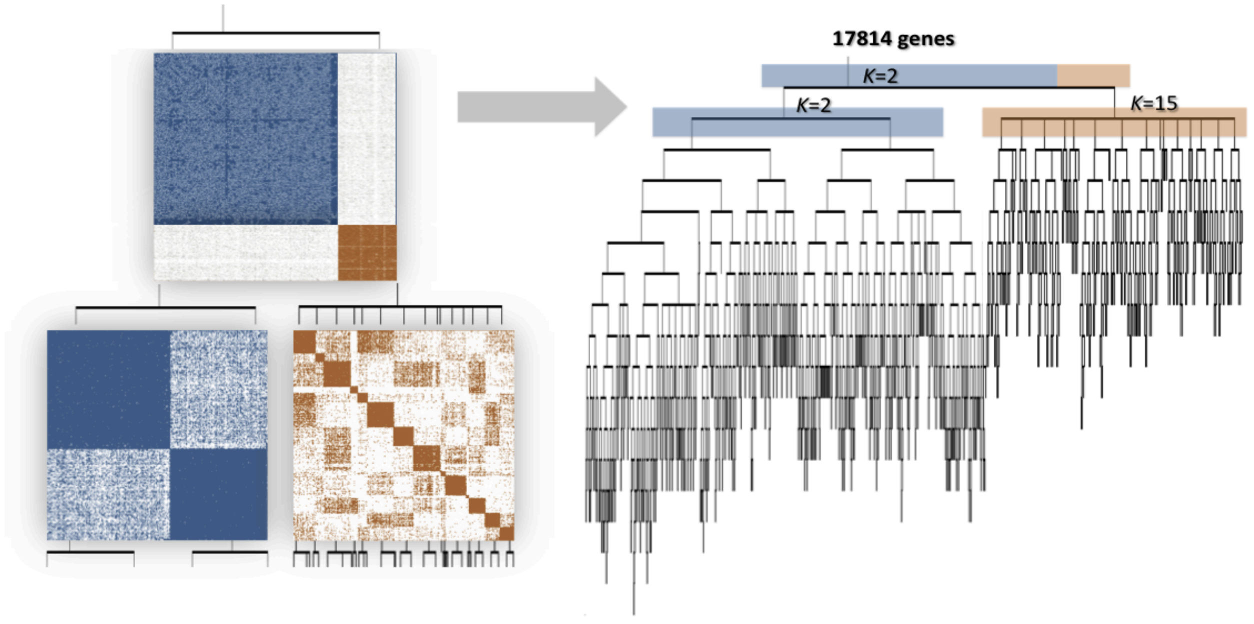
Spectral approaches have also some drawbacks. Their basic assumption of piecewise constancy in the form of leading eigenvectors need not hold on real data. Much work has been done to make this step robust, including the introduction of optimal cut ratios[17] and relaxations[18,19] and highlighting the conditions under which these methods can be expected to perform well[14].

Spectral methods can be slow as they involve eigen decomposition of potentially large matrices ($O(n^3)$). Recent attempts at addressing this issue include implementing the algorithm in parallel[20], speeding eigen decomposition with Nystrom approximations[21], building hierarchical transition matrices[22] and embedding distortion measures for faster analysis of large-scale datasets[23].

2.2. *Recursive K-means Spectral clustering algorithm (ReKS)*

In this paper, we will pursue a recursive form of *K*-means spectral clustering (ReKS), apply it on cancer expression data from patients and understand the intrinsic structure of the data by establishing a baseline clustering result. ReKS first defines an affinity matrix of all pairwise similarities between genes. We reduce the computational burden with sparse matrices, such that each gene is connected to a small number of its neighbors (default: 15) with varying affinities, and extract only a small subspace of eigenpairs (default: 20). In each recursion step, we determine the most appropriate subspace in which to run *K*-means using the eigengap heuristic, which is to compute the ratio of successive eigenvalues and pick *K* that satisfies: max{i: λ_i / λ_{i+1}, for i = 1 to 20}. We apply the eigengap heuristic at each recursion level to determine the optimal number of partitions at that level. In addition, to improve the convergence of the *K*-means algorithm we initiate the algorithm with orthogonal seed points. For each newly formed cluster, we extract the corresponding affinity sub-matrix and repeat the procedure.

In Figure 2(a) we illustrate the top two levels of ReKS recursion on the GBM dataset. At level-1 an obvious partition exists for the original affinity matrix. The genes are split into two clusters at this node, and for each cluster, a new affinity matrix is computed.



(a) (b)

Fig. 2. (a) Demonstration of the ReKS method on the GBM dataset at the first two iterations of *K*-means spectral decomposition recursions: two clusters are visible in the affinity map constructed from the entire dataset at the first level. From each, a new affinity matrix is constructed and spectral clustering repeated on the sub-affinity matrix. (b) Complete tree obtained by ReKS iterations. Each leaf node corresponds to a gene cluster in the final partition.

ReKS performs this procedure iteratively stopping when further split would cause all clusters to be 35 or smaller in size. The stopping threshold corresponds to the average number of genes that participate in a KEGG[24] pathway. In the end, we arrive at a tree where each leaf node represents a gene cluster. Note that with this procedure clusters of smaller than 35 genes could be obtained, for example due to an early split off the tree, as long as there is a cluster that is large in size. Figure 2(b) presents the full tree generated by ReKS on the GBM dataset.

3. ReKS evaluation on cancer patient data

3.1. *Data*

We applied ReKS on the three most complete TCGA gene expression datasets to date: Glioblastoma multiform (GBM) with a total of 575 tumor samples, Ovarian serous cystadenocarcinoma (OV) with a total of 590 tumor samples, and Breast invasive carcinoma (BRCA) with a total of 799 tumor samples. The level 3, normalized and gene-collapsed data obtained from the TCGA portal were downloaded and no further normalization was performed.

3.2. *Comparison of ReKS and other clustering strategies on TCGA data*

We compare our method against four other partition solutions: (1) average linkage hierarchical clustering, (2) average linkage hierarchical clustering on the spectral space, (3) K-means and (4) K-means on the spectral space. These algorithms are chosen to cover a range of common clustering techniques and clustering assumptions.

Agglomerative clustering methods build a hierarchy of clusters from bottom up. It is perhaps the most popular on gene expression data analysis[25], due to its ease of use and readily available implementations. We performed hierarchical agglomerative clustering using Euclidean distance and average linkage. A maximum number of clusters is specified to be comparable to the number of clusters K obtained when running ReKS. Since this choice might be considered favorable to ReKS, we also performed hierarchical clustering on the top three eigenvectors in the spectrum, using cosine distances to measure the distance on the resulted unit sphere. Note that hierarchical clustering is done from bottom up, using local similarities, and does not embed the global structure in its tree.

Similarly, standard K-means and K-means performed on the spectral space are included for benchmarking purposes. Given a number of clusters, K, the algorithm iteratively assigns members to centroids and re-adjusts the centroids of the clusters. K-means tends to perform well as it directly optimizes the intra-cluster distances, but tends to be slow especially as K increases. Here we used the default implementation of the K-means clustering algorithm in Matlab, with Euclidean distance, again using the K obtained from ReKS. We also ran K-means on the spectral space, effectively performing ReKS only once without choosing an optimal number of eigenvectors to use, but instead using K top eigenvectors.

Shown in Figure 3 are the distributions of the cluster sizes when applying the five methods to the three TCGA datasets. Hierarchical clustering, whether in the original or the eigenspace, produces a very skewed distribution of cluster sizes that is possibly an artifact of focusing on only local similarities. The *K*-means methods and ReKS produce cluster sizes that span roughly the same range. However, the *K*-means methods produce distributions that are artificially Gaussian, with relatively little clusters that contain small number of genes.

.

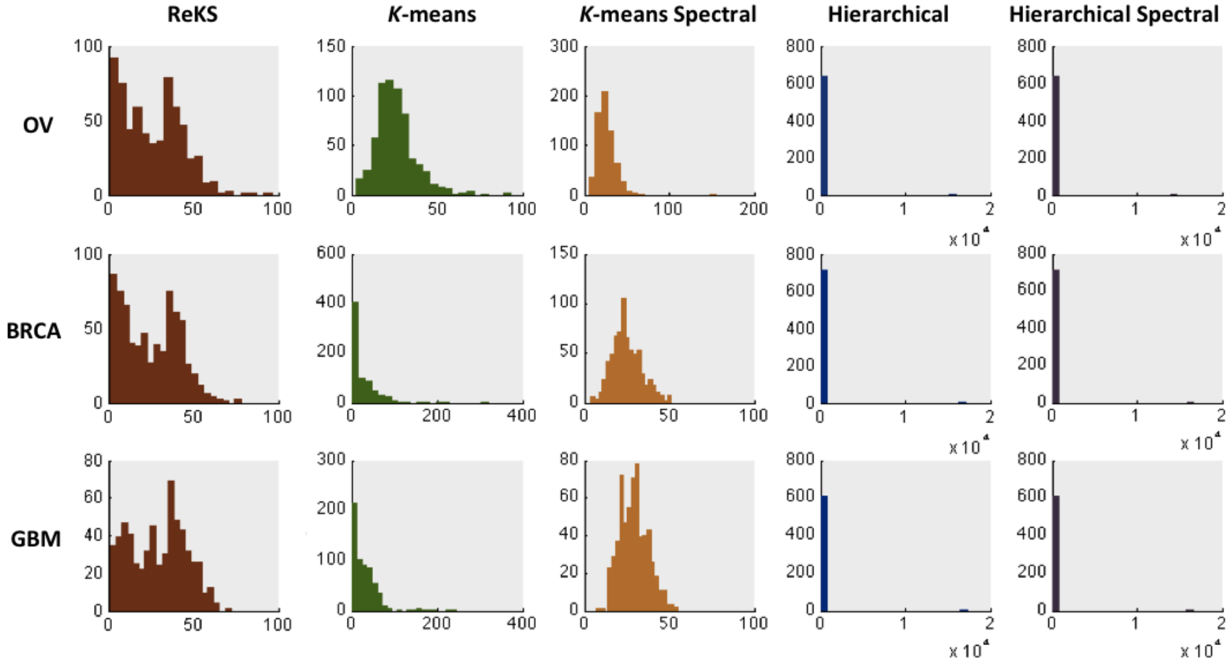


Fig. 3. Distribution of cluster sizes of ReKS and of other methods

3.3. *Cluster quality evaluation*

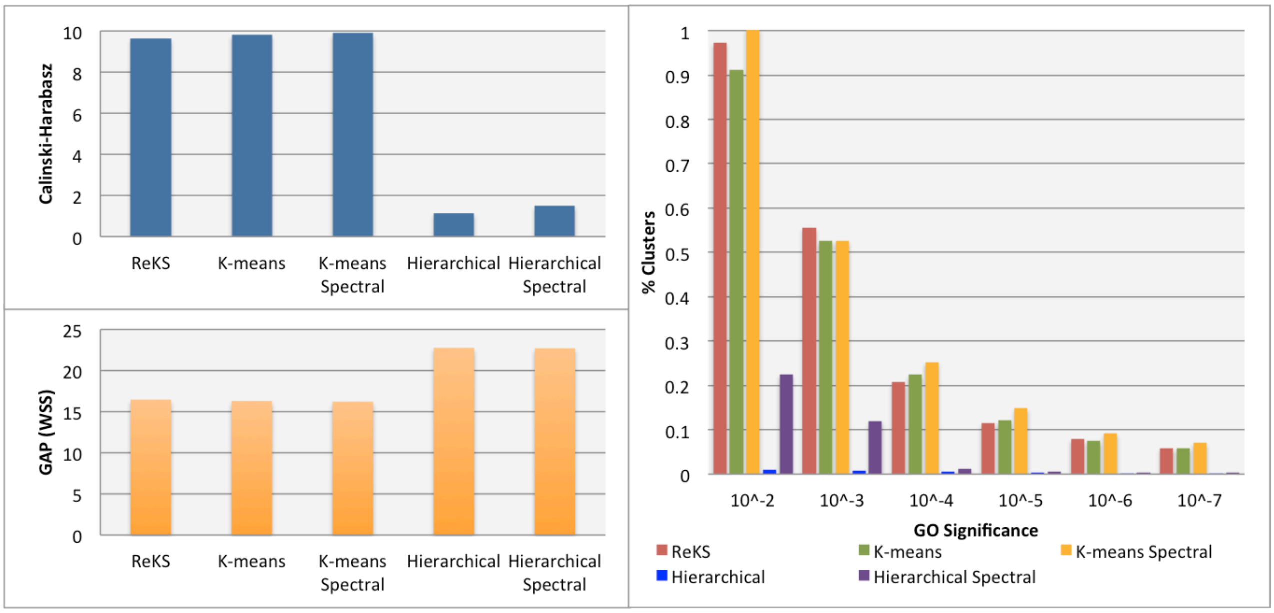
We evaluate the quality of the clusters obtained from each of the five methods (ReKS, *K*-means, *K*-means spectral, Hierarchical, Hierarchical spectral) using both intrinsic, statistical measures as well as external biological evidence, as detailed in the sections below.

3.5.1. Calinski-Harabasz

To evaluate the quality of the clusters, we used the Calinski-Harabasz measure[26], defined by:

$$CH = \frac{traceB/(K-1)}{traceW/(N-K)} \tag{4}$$

where $traceB$ denotes the error sum of squares between different clusters, $traceW$ is the intra-cluster square differences, N is the number of objects, and K is number of clusters. This statistic is effectively an adjusted measure of the ratio of between- vs. within- group dispersion matrices. A larger value denotes a higher compactness of the cluster compared to the inter-cluster distances. Figure 4(a) shows the performance of ReKS compared across other methods. Not surprisingly, ReKS outperforms hierarchical clustering in both the original data space as well as the spectral space, as hierarchical clustering produces some very large clusters with no apparent internal cohesion. The *K*-means based methods and ReKS are comparable in terms of cluster separation across the datasets.



(a) (b)

Fig. 4. (a) Cluster validity comparison with other methods using the Calinski-Harabasz and the GAP statistics (b) Gene Ontology(GO) enrichment across different range of p-values

3.5.2. GAP Statistic

The Gap statistic was proposed as a way to determine optimal cluster size[27]. In short, it is the log ratio of a reference within-cluster sum of square errors over the observed within-cluster sum of squares errors. The reference is usually built from a permutated set of genes that form *K* random clusters. Since we are comparing the (five) methods across the same dataset with the same *K*, it is fair to compare the performance of the observed within sum-of squares error only. With this direct proxy, ReKS performs at the same level as *K*-means based methods (shown in Figure 4(a)), and achieved a significantly lower sum-of-square distances than the hierarchical methods.

3.5.3. Gene Ontology Enrichment

Since no ground truth exists for gene cluster partition, we examine the overall quality of the clusters in terms of the amount of enrichment for Gene Ontology (GO) annotations. For each cluster, we test for GO enrichment using a variant of the Fisher's exact test, as described in the *weight01* algorithm of the topGO[28] package in R. The significance level of the test indicates the degree a particular GO annotation is over-represented in a given cluster. For a partition, we calculate the proportion of clusters annotated with a GO term at a p-value threshold. If a cluster has less than five members, the test is not performed. As shown in Figure 4(b), compared to hierarchical clustering, we observe that ReKS contains higher percentage of clusters that are significant at the specified levels, and especially so with more stringent p-value thresholds, and performs roughly the same as *K*-means methods. Finally, we observe that the spectral methods tend to perform better than their non-spectral counter-parts.

3.5.4. Execution Time

Table 1 shows the execution time of the five methods on a 3.4 GHz Intel Core i7 CPU. ReKS is slower than hierarchical clustering but compares favorably to K-means methods.

Table 1: Average execution time of the five methods

Methods	ReKS	K-means	K-means Spectral	Hierarchical	Hierarchical Spectral
Execution time	373s	6000s	1774s	90s	22s

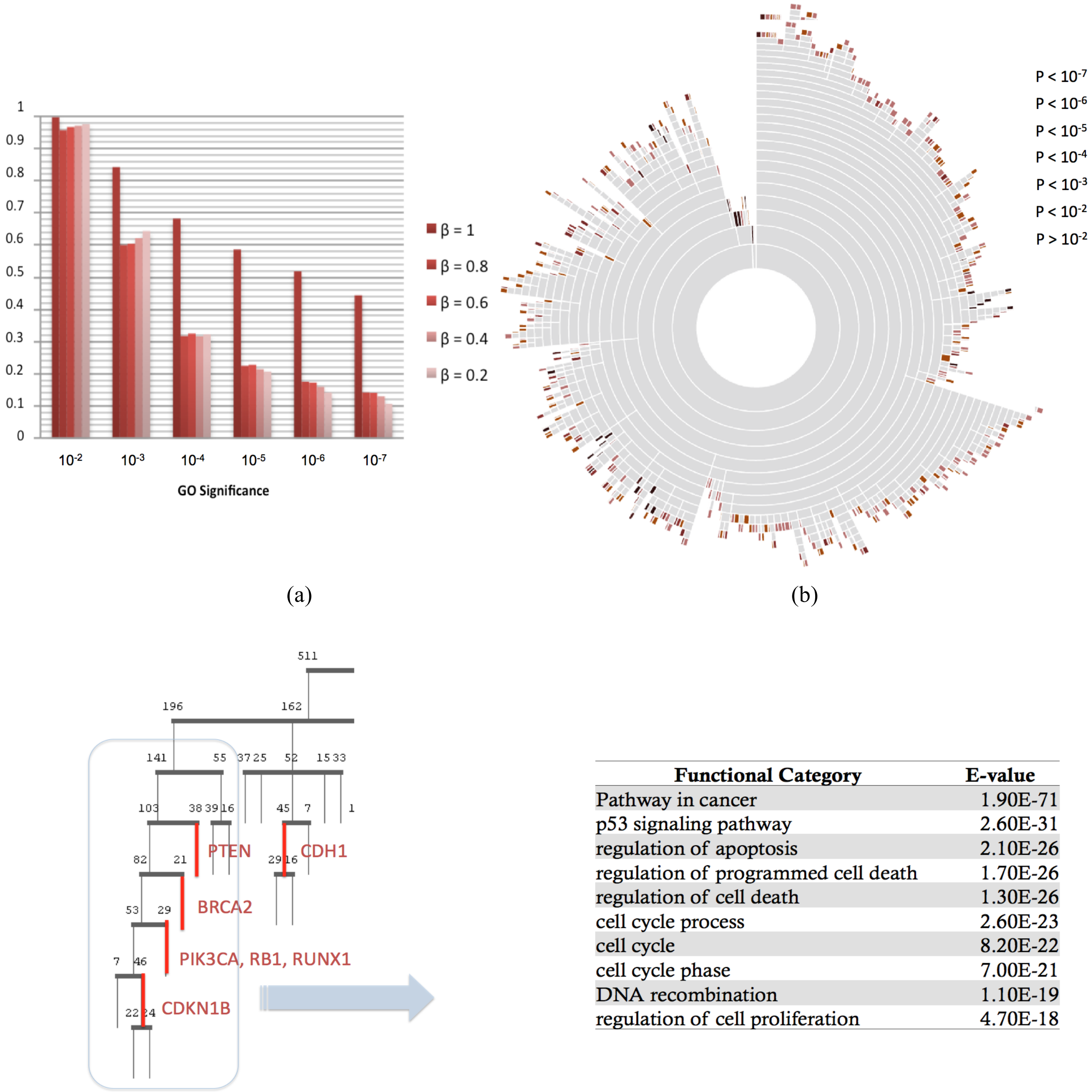
3.4. *Incorporation of Prior Information*

We use existing expert knowledge as prior information (from KEGG pathway[24]) to guide our clustering method, aiming to generate partitions that are even more biologically meaningful. The KEGG database includes a collection of manually curated pathways constructed from knowledge accrued from the literature. For the purposes of ReKS, we assume that the genes in a KEGG pathway are fully connected to each other (i.e., should belong in the same cluster). We code this prior knowledge in a constraint matrix U in which each column Uc is a pathway, and u_{ic}= 1, u_{jc}= -1 if a pair of genes i, j participate in the same KEGG pathway c. Similar to what was detailed in Ji *et al.*[29], where they supplied a prior for document clustering using K-means spectral decomposition, we apply a penalty term to the normalized graph Lapacian as follows:

$$L' = D^{-0.5}(W + \beta U^T U)D^{-0.5} \quad (5)$$

where $\beta \geq 0$ controls the degree of enforcement of the KEGG prior knowledge. As shown in Ji *et al.*, the eigenvectors of the K smallest eigenvalues of L' form the eigen-space represents a transformation of the affinity space embedded with prior information. We then proceeded to apply the K-means algorithm within the eigenspace, and iterate recursively as we did with ReKS. As shown in Figure 5(a), when we use a large amount of prior, not surprisingly the GO significance becomes very large. We observe the significance of the clusters do not drop very fast as β decreases. Therefore, small amount of prior at roughly β = 0.2 may be enough to enhance the biological significance of the ReKS clustering results.

We applied ReKS on the TCGA datasets at $\beta = 0.2$. A total of 715, 639, and 610 clusters are obtained for BR, OV, and GBM respectively. As shown in Figure 5(a), we observe that there exists a slight anti-correlation between how early a cluster splits off the tree and how significant the cluster is (ρ = -0.2112, p <10^{-7}). As a preliminary observation, how early a cluster is formed seems to imply the "tightness" of the cluster, this result seems to suggest that there is a slightly higher chance the clusters that form early to be more biologically significant. For example, in Figure 5(b) there is a tight histone H1 cluster that splits off the BRCA tree at the third level on the top. It has been shown that EB1089 treatment of breast cancer cell lines (MCF-7, BT20, T47D, and ZR75) is correlated with a reduction in CdK2 kinase activity towards phosphorylation of histone H1 and a decrease in DNA synthesis[30]. This cluster was not found in K-means spectral, K-means, and spectral hierarchical clustering results, and only exists in a mega-cluster in hierarchical clustering partition. Additionally, upon examining the resulted tree closely, we found that a few genes that have been implicated for breast cancer[31] cluster together or close to each other on the tree, as shown in Figure 5(c). When considering a few of these sub-clusters together, the top



Functional Category	E-value
Pathway in cancer	1.90E-71
p53 signaling pathway	2.60E-31
regulation of apoptosis	2.10E-26
regulation of programmed cell death	1.70E-26
regulation of cell death	1.30E-26
cell cycle process	2.60E-23
cell cycle	8.20E-22
cell cycle phase	7.00E-21
DNA recombination	1.10E-19
regulation of cell proliferation	4.70E-18

(a)

(b)

(c)

Fig. 5. (a) Effect of incorporation of prior information on the GO significance of the obtained clusters. β controls the degree of enforcement of the KEGG prior knowledge (b) A sunburst diagram for the BRCA dataset. In this alternative representation of the ReKS clustering results, each concentric circle represents a level of the tree. Each ring is sub-divided into clusters. The color of a leaf node denotes the GO significance of the cluster. There exists a small anti-correlation (ρ = -0.2112, p < 10^{-7}) between the level from which a cluster splits off, and its GO significance (c) A part of the tree enriched with genes implicated for breast cancer (level 2 and down), and the GO significance and categories of the 169 gene super-cluster (grey box).

functional categories that emerged are indeed caner and p53 pathways. We found several of these examples throughout the tree, all within 12 levels up to which the composition of the clusters remains stable when splitting the data into training and test sets. We note that PIK3CA, RB1, and RUNX1 do not cluster together in any of the other methods we compared to, nor does the rest of the genes we examined. This example suggests that the tree structure could be useful for inferring additional previously unknown biomarkers.

4. Discussion

In this study, we demonstrate the utility of a new recursive spectral clustering method we proposed as an alternative to traditional methods for clustering large-scale, human disease expression data. Consistent with previous findings[25], hierarchical methods are faster but perform relatively poorly. *K*-means methods can be accurate when the number of groups *K* is known. However, in the case of gene clustering of disease samples we are rather agnostic as to the number of the clusters we should expect. ReKS does not require the number of clusters to be known *a priori*, and is an order of magnitude faster than the original *K*-means algorithm. Also, compared to *K*-means spectral, ReKS enjoy a considerable speed gain by performing the decomposition and clustering iteratively, while maintaining a comparable performance even without directly minimizing the overall inter- and intra- cluster distances(sec 3.4).

By incorporating prior pathway information in the algorithm, ReKS additionally guides the clustering process toward a more biologically meaningful partition. We showed that the clusters obtained are biologically relevant in their enrichment in GO terms, and the size of the clusters has a more natural distribution than that of *K*-means or hierarchical clustering partitions. The clusters, being rather compact and constrained in size, could then be used in subsequent studies, where clusters of genes could potentially be used as predictors for disease classification. Not only does ReKS provide a partition of the gene space, the resulting tree structure provides a hint to the relative tightness of the clusters and potential targets. In the future, we wish to investigate the relationship between the relative position of the cluster in the tree and their potential strengths in classifying disease labels and other clinical variables. Also, it is possible to automatically calculate the optimal number of neighbors to be considered in each recursion level. For example, we can use an approach similar to eigengap, where the distribution of similarities for each node will be compared to the global distribution to identify the optimal number of informative neighbors. The above results indicate that, when applied to large clinical datasets, recursive spectral clustering offers an attractive alternative to conventional clustering methods.

5. Acknowledgements

The authors wish to thank Nermin Pinar Karabulut for help in the study of perturbation-based spectral clustering.

References

1. The Cancer Genome Atlas - Data Portal. at <https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp>
2. Butterworth, R., Piatetsky-Shapiro, G. & Simovici, D. A. On Feature Selection through Clustering., *IEEE International Conference on Data Mining.* **0**, 581–584 (2005).
3. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10869–10874 (2001).

4. Jain, A. K. & Dubes, R. C. *Algorithms for clustering data*. Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, (1988).
5. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863 – 14868 (1998).
6. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
7. Tseng, G. C. & Wong, W. H. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61**, 10–16 (2005).
8. Bezdek, J. C., & Pal, S. K. Fuzzy models for pattern recognition: Methods that search for structures in data. *IEEE Press, New York, NY (1992).*
9. Frey, B. J, and Dueck, D. “Clustering by Passing Messages Between Data Points.” *Science* **5814**: 972–976 (2007)
10. Braun, R., Leibon, G., Pauls, S. & Rockmore, D. Partition decoupling for multi-gene analysis of gene expression profiling data. *BMC Bioinformatics* **12**, 497 (2011).
11. Spellman, P. T. *et al.* Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
12. Luxburg, U. V., Belkin, M., Bousquet, O. & Pertinence A tutorial on spectral clustering. *Statistics and Computing 17(4)* (2007).
13. Meila, M. & Shi, J. A Random Walks View of Spectral Segmentation. *8th International Workshop on Artificial Intelligence and Statistics (AISTATS)*(2001).
14. Chennubhotla, C. & Jepson, A. D. Half-lives of eigenflows for spectral clustering. *Advances in Neural Information Processing Systems* 689–696 (2002).
15. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci.* **102**, 7426–7431 (2005).
16. Ng, A. Y., Jordan, M. I. & Weiss, Y. On Spectral Clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems* 849–856 (2001).
17. Shi, J. & Malik, J. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 888–905 (2000).
18. Bach, F. R. & Jordan, M. I. Learning Spectral Clustering. *Advances in Neural Information Processing Systems 16* (2003).
19. Tolliver, D. A. Graph partitioning by spectral rounding: Applications in image segmentation and clustering. *Computer Vision and Pattern Recognition* 1053–1060 (2006).
20. Song, Y., Chen, W., Bai, H., Lin, C. & Chang, E. Y. Parallel Spectral Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33,3,568-586 (2011)*
21. Drineas, P. & Mahoney, M. W. On the Nystrom Method for Approximating a Gram Matrix for Improved Kernel-Based Learning. *J. Mach. Learn. Res.* **6**, 2153–2175 (2005).
22. Chennubhotla, C. & Jepson, A. D. Hierarchical eigensolver for transition matrices in spectral methods. *Advances in Neural Information Processing Systems* 273–280 (2005).
23. Yan, D., Huang, L. & Jordan, M. I. Fast approximate spectral clustering. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* 907–916 (2009).
24. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
25. Souto, M. C. de, Costa, I. G., Araujo, D. S. de, Ludermir, T. B. & Schliep, A. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* **9**, 497 (2008).
26. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods* **3**, 1–27 (1974).
27. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistical Society* **63**, 411–423 (2000).
28. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).
29. Ji, X. & Xu, W. Document clustering with prior knowledge. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* 405–412 (2006).doi:10.1145/1148170.1148241
30. Wu, G., Fan, R. S., Li, W., Ko, T. C. & Brattain, M. G. Modulation of cell cycle control by vitamin D3 and its analogue, EB1089, in human breast cancer cells. *Oncogene* **15**, 1555–1563 (1997).
31. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. *Nature* (2012).

SYSTEMATIC IDENTIFICATION OF RISK FACTORS FOR ALZHEIMER'S DISEASE THROUGH SHARED GENETIC ARCHITECTURE AND ELECTRONIC MEDICAL RECORDS

LI LI*
Div. of Systems Medicine, Dept. of Pediatrics, Stanford University School of Medicine
Stanford, CA 94305, USA
Email: l3li@stanford.edu

DAVID RUAU*
Div. of Systems Medicine, Dept. of Pediatrics, Stanford University School of Medicine
Stanford, CA 94305, USA
Email: druau@stanford.edu

RONG CHEN
Personalis Inc., 1350 Willow Rd.,
Menlo Park, CA 94025, USA
Email: rong.chen@personalis.com

SUSAN WEBER
Stanford Center for Clinical Informatics, Stanford University School of Medicine
Stanford, CA 94305, USA
Email: scweber@stanford.edu

ATUL J. BUTTE
Div. of Systems Medicine, Dept. of Pediatrics, Stanford University School of Medicine
Stanford, CA 94305, USA
Email: abutte@stanford.edu

**: authors are equally contributed to this work*

Alzheimer’s disease (AD) is one of the leading causes of death for older people in US with rapidly increasing incidence. AD irreversibly and progressively damages the brain, but there are treatments in clinical trials to potentially slow the development of AD. We hypothesize that the presence of clinical traits, sharing common genetic variants with AD, could be used as a non-invasive means to predict AD or trigger for administration of preventative therapeutics. We developed a method to compare the genetic architecture between AD and traits from prior GWAS studies. Six clinical traits were significantly associated with AD, capturing 5 known risk factors and 1 novel association: erythrocyte sedimentation rate (ESR). The association of ESR with AD was then validated using Electronic Medical Records (EMR) collected from Stanford Hospital and Clinics. We found that female patients and with abnormally elevated ESR were significantly associated with higher risk of AD diagnosis (OR: 1.85 [1.32-2.61], p=0.003), within 1 year prior to AD diagnosis (OR: 2.31 [1.06-5.01], p=0.032), and within 1 year after AD diagnosis (OR: 3.49 [1.93-6.31], p<0.0001). Additionally, significantly higher ESR values persist for all time courses analyzed. Our results suggest that ESR should be tested in a specific longitudinal study for association with AD diagnosis, and if positive, could be used as a prognostic marker.

1. Introduction

Alzheimer's disease (AD) is the fifth-leading cause of death in older people and is the most common cause of dementia (up to 75%), with an approximately 26 million affected individuals worldwide estimated to reach 115 million by 2050 (*1-3*). Of those with Alzheimer's disease, an estimated 4% are under age 65, 6 % are 65 to 74, 44 % are 75 to 84, and 46% are 85 or older (*2*). Compared with men, women have a 1.54 fold increased risk for AD (95% CI, 1.21 to 1.96) (*4*).

About 25% of all AD cases have familial history (i.e., with 2 or more persons in a family having AD). Nevertheless, the main cause remains unknown, which may due to genetic and environment factors (*5*). AD is an irreversible and progressive brain disease, which can be diagnosed using behavioral observations and the gold standard for confirmation rely on neuropathologic findings of beta-amyloid plaques and intraneuronal neurofibrillary tangles upon autopsy examination (*6*). Therefore, identifying clinical manifestations of risk factors related with AD are critically needed for early diagnosis, prognostics and preventive care of AD. Currently, the known risk factors of AD are advancing age, family history, gender, *APOE* ε4 allelic variant, cardiovascular factors, mild cognitive impairment, life style, and head trauma, which were investigated through large scale epidemiological studies (*7-13*). However, these factors have relatively weak predictive effects. It is still necessary to find more potential risk factors which may contribute to AD development (*3*).

Over the past decade, Genome-Wide Association Study (GWAS) and candidate gene studies have identified genetic variants for thousands of diseases and traits (*14-16*). A previous study has shown the "human disease network" where two diseases were connected to each other if they shared at least one gene from Online Mendelian Inheritance in Man (OMIM), however, they did not integrate GWAS studies (*17*). We hypothesize that traits from GWAS studies might serve as additional risk factors for disease, here specifically looking at AD. We theorize that if a prior GWAS for a trait has identified a list of genes with variants that significantly match the list of genes with variants associated with AD, then that trait might serve as a predictive factor for AD.

In this study, we used those variants and develop a method to systemically identify associations between clinical traits and AD in a fast and efficient way. We searched for traits sharing common genetic variants with AD that could serve as a means to prognose AD, and possibly provide opportunities for life-style interventions and preventive drug treatment. We validated our novel finding using Electronic Medical Records (EMR) through an independent large patient cohort with more than 15,000 patients from Stanford Hospital and Clinics (SHC) (*18*).

2. Methods

2.1 Utilizing VARiant Informing MEDicine (VARIMED)

The overall experiment design is shown in Figure 1. GWAS have enabled the elucidation of the genetic architecture of hundreds of diseases, many of which are polygenic complex disorders. We have manually curated a unique database called VARiant Informing MEDicine (VARIMED) (*19*), holding manually curated, quantitative human disease-SNP associations extracted from the full text, figures, tables, and supplemental materials of human genetic related publications.

VARIMED is a comprehensive genetic association database with over 100 features stored including diseases (e.g. diabetes, lung cancer), clinical traits (e.g. blood pressure, creatinine levels), gene symbol, dbSNP, odds ratio, and published p-value of association from literature (*19*-*22*). Diseases are categorized and currently mapped to Concept Unique Identifiers (CUI) from the Unified Medical Language System. All the genetic variants (SNPs) were systematically annotated to the genes with the most recent NCBI Entrez gene identifiers using Entrez dbSNP by AILUN (*23*). At the time of this writing, VARIMED covers 8,962 human genetics papers from GWAS and candidate gene studies, including 87,553 SNPs annotated to 8,913 genes for 1,119 diseases and 1,257 clinical traits.

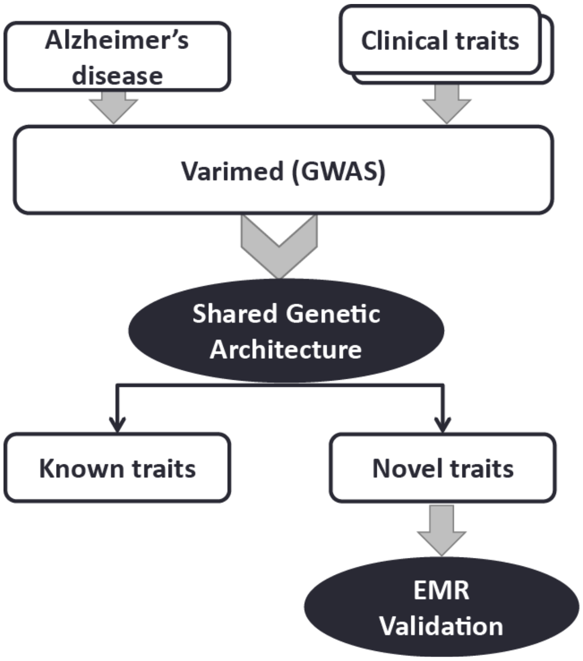


Figure 1: Work flow for entire experiment design

2.2 Assessing shared genetic architecture for Alzheimer's disease (AD) and clinical traits

We compared the shared genetic architecture for all available clinical traits in VARIMED with against Alzheimer's disease (AD) by first collecting all genetic variants related with AD and 1,257 traits. We selected only those variants associated at the gene level with AD and traits with p ≤ 1E-8 as a highly stringent threshold to reduce the chance of false positive results.

As some genes could be shared solely between a few traits, and other genes shared across thousands, we needed an approach to capture the specificity and relevance of the genetic association. We used a Term Frequency–Inverse Document Frequency (TF-IDF) weighing method (*24*) to take into account the popularity of the genes. The detailed calculation procedure is as follows. First, we calculated a term frequency (TF) using:

$$tf(i,j) = \frac{frequency\ of\ the\ gene\ i\ in\ phenotype\ j}{number\ of\ all\ genes\ in\ phenotype\ j} \tag{1}$$

where phenotype refers to a trait or the disease AD.

The *tf* score indicates the occurrence frequency level of gene *i* in phenotype *j*, similar to a precision measure. Then, we calculated the inverse document frequency (IDF) using:

$$idf(i) = \log_{10}(\frac{total\ number\ of\ phenotypes}{number\ of\ phentoypes\ containing\ gene\ i}) \qquad (2)$$

A larger *idf* score implies a lower popularity of gene *i* among the phenotypes (akin to a higher accuracy), which gives more weight to the gene as it might only be shared between these two phenotypes. Last, the TF-IDF score was calculated using:

$$tf\text{-}idf_{(i,j)} = tf_{(i,j)} \times idf_i \qquad (3)$$

A high weight in *tf-idf* is reached by a high gene frequency (in the given phenotype) and a low phenotype frequency of the gene across all phenotypes studied.

Thus, for every AD-trait pairs a TF-IDF score for every shared gene was computed. The similarity between AD and all traits was then estimated by the cosine distance based on *tf-idf* scores.

To evaluate the statistical significance of the distance scores obtained, we computed the False Discovery Rate (FDR) by random shuffling (1,000 times) the genes across all the traits and re-computing the AD-trait distance. The q-value was calculated as the ratio of the expected number of false positive over the total number of hypothesis tested (*25*). Q-value ≤ 0.01 was selected as threshold of significant association between AD and trait pairs.

2.3 Validation of novel finding from the independent electronic medical records

To assess the clinical relevance of our novel finding, we used electronic medical records (EMR) data extracted from Stanford Translational Research Integrated Database Environment (STRIDE). STRIDE is a research and development project at Stanford University to create a standards-based informatics platform supporting clinical and translational research (*18*). STRIDE contains a clinical data warehouse which is comprised of comprehensive clinical information such as ICD9 diagnoses codes, CPT procedure codes, and lab results on over 1.7 million pediatric and adult patients cared for at Stanford Hospital and Clinic. STRIDE has been implemented at SHC since 2005. We used patient data in STRIDE as an independent cohort specifically recruited for this study to validate the hypothetical associations observed between AD and traits at the genetic level. Patients with AD were retrieved using the ICD9 code = 331.0, the rest of the hospital population being considered as control.

Chi-square test and Mann–Whitney U test were used to investigate the effect of the traits and AD. All statistics and graphs were carried out by SAS 9.2 (SAS institute Inc., Cary, SC) and R 2.15.0 (*26*).

2.4 Ethical statement

Data collected from STRIDE did not contain any protected health information and thus the study was considered non-human subjects' research, as determined and approved by the Institutional Review Board at Stanford.

3 Result

3.1 Discovering genetic architecture related with Alzheimer's disease

From 8,962 GWAS and candidate genes studies implemented in VARIMED, we queried the number of unique SNPs, genes, and genetic studies associated with Alzheimer's disease (AD). We used a stringent and well-accepted p-value threshold ≤ 1E-8 as genome wide significant, and identified 89 SNPs within 28 genes published across 44 genetic studies associated (Table 1).

Table 1: Genes and number of genetic studies associated with Alzheimer's disease

Gene	SNP Count	P-value	Study Count
APOD	1	0	1
SORCS1	1	0	2
APOC1	1	1.00E-300	8
TOMM40	9	1.28E-299	10
PVRL2	18	5.65E-74	7
APOE	2	1.83E-67	8
BCL3	2	1.93E-21	3
ABCA7	1	5.00E-21	3
LRRC68	4	2.16E-20	2
BCAM	1	5.54E-19	1
CLU	2	1.10E-16	6
MS4A6A	6	1.20E-16	2
PCK1	1	2.00E-16	4
ZNF224	1	2.00E-16	4
CR1	7	3.70E-14	5
PVR	1	6.17E-12	2
NKPD1	1	1.04E-11	1
MS4A4A	2	4.71E-11	1
GAB2	3	9.66E-11	5
MTHFD1L	1	1.90E-10	2
CALHM1	1	2.00E-10	3
CLPTM1	1	2.00E-10	1
CEACAM16	1	7.68E-10	2
PICALM	13	9.57E-10	1
CD33	1	1.60E-09	2
MS4A4E	5	1.98E-09	1
MS4A2	1	2.94E-09	2
CD2AP	1	8.60E-09	2

3.2 Systematically identifying the significant traits with genetic architecture shared with AD

We identified 249 traits where at least one gene was genetically associated. In our study, a trait was defined as a human-related physical or cognitive measurement, which was not explicitly a predisposition to another disease. To evaluate the significance of the shared variants in AD and all possible trait pairings, we attributed to each gene a measure based on their popularity using TF-IDF weight adjustment, and tested for significance using random permutation (see Methods section 2.2). We identified 6 significant traits that paired with AD with q-value ≤ 0.01 (Table 2) based on the method we described above. All 6 traits originated from different published GWAS studies, suggesting that integrating different GWAS studies to discover underlying shared genetic architecture between diseases and traits can yield novel risk factors for the disease.

Among the 6 traits, 5 were related lipid tests and all shared variants in *APOC1*, *PVRL2*, and *TOMM40* genes in their genetics. *APOE* was shared in the lipid panel however was absent in Lipoprotein-Associated phospholipase a2 activity (Lp-PLA2) (Table 2). Erythrocyte sedimentation rate (ESR), a common immunology test to measure non-specific inflammation showed significant genetic association with AD through only one gene: complement component (3b/4b) receptor 1 (*CR1*). *CR1* was associated with ESR and AD solely and not with other phenotypes in VARIMED. *CR1* is a receptor and binds to *C3* and *C4* complement genes, which have been shown an increase in chronic inflammation (*27*), in risk of developing a myocardial infarction (*28*), and in deceased donor who progressed poor graft function due to cold ischemic injury with potential inflammation after kidney transplantation (*29*).

Among the 6 traits associated with AD, 5 associations were already known to be either risk factors or comorbidities of AD in the published literature (Table 2). Lipoprotein-Associated phospholipase a2 (Lp-PLA2) is a risk factor associated with the risk of dementia in the Rotterdam study, independently of cardiovascular and inflammatory factors (*30*). C-reactive protein (CRP) level is a risk factor where elevated CRP continues to predict increased dementia severity suggesting a possible proinflammatory endophenotype in AD (*31*). In addition, lipid level has been seen to increase in patients who have already developed AD. Apolipoprotein b (ApoB) level is increased in AD patients, suggesting that ApoE may not be the single factor in lipid metabolism to play a role in AD pathogenesis (*32*). Higher total cholesterol and LDL levels were significantly related to pathologically defined AD, which in turn suggests serum lipids have a role in the pathogenesis of AD and interventions may modify the progression of disease (*33,34*). Furthermore, the shared genes also explain the genetic cause between AD and these 5 traits.

Table 2: Clinical traits significant associated with Alzheimer's disease

Clinical Trait	**Gene Count**	**Common Genes**	**Gene Shared**	**Q-value**	**Reference**
Lipoprotein-Associated phospholipase a2 activity	12	3	*APOC1;PVRL2; TOMM40*	< 0.001	*30*
Apolipoprotein b levels	12	4	*APOC1; APOE; PVRL2; TOMM40*	< 0.001	*32*
C reactive protein levels	17	3	*APOC1; APOE; TOMM40*	0.002	*31*
LDL cholesterol levels	44	4	*APOC1; APOE; PVRL2; TOMM40*	0.002	*34*
Erythrocyte sedimentation rate	5	1	*CR1*	0.004	Novel
Cholesterol levels	50	4	*APOC1; APOE; PVRL2; TOMM40*	0.004	*33*

3.3 Clinical validation for novel trait ESR association with AD in an independent cohort

We identified ESR as a novel trait significantly sharing genes with genetic variants with AD. Since ESR is a well-known clinical measurement and non-specific marker of inflammation, and not known to be associated with AD, we evaluated the hypothesis that ESR might be abnormal before the diagnosis of AD. We obtained all ESR lab results from Stanford Hospital and Clinics from 2005 until July 15, 2012 for patients with and without an AD diagnosis. Our case cohort was constituted of 212 patients who were ever measured for ESR and had at least one diagnosis code of AD (mean age 81±8; range [48-96]) with 135 females and 78 males. We considered patients older than age 50, having at least one measurement of ESR, and never having a diagnosis code of AD as the control group, resulting in 15,040 unique patients. Reference ranges for Erythrocyte sedimentation rate (ESR) lab tests were defined 0-20 mm/hr for female <50, 0-30 mm/hr for female ≥ 50 years, and 0-20 mm/hr for male ≥ 50 years based on MedlinePlus (http://www.nlm.nih.gov/medlineplus/).

As AD is known to exhibit a sex difference in prevalence (*4*), we evaluated each gender separately. First, we compare the abnormal high ESR percentage for AD and control patients across all available time points (ESR measurement irrespective of the AD diagnosis code(s)) to test the overall association. Then, we compared the abnormal high ESR percentage within 1 year prior to our first diagnosis code of AD in AD patients, and first diagnosis code of anything other than AD in control patients, to investigate whether changes in ESR could be a risk factor to predict the AD incidence. Finally, we compared the ESR within 1 year after our first diagnosis code of AD in AD patients, and first diagnosis of anything other than AD in control patients, to evaluate whether ESR changes could be a consequence of the AD diagnosis.

In female, patients with abnormally high ESR (45%) (> 30 mm/hr) were significantly associated with having a diagnosis code of AD irrespective of lab and diagnosis timing (OR: 1.85 [1.32-2.61], p=0.0003). The effect was strengthened when looking at ESR measurements within 1 year prior to our first AD diagnosis for patients (OR: 2.31 [1.06-5.01], p=0.032), and within 1 year after our first AD diagnosis on patients (Table 3). Furthermore, ESR values were significantly higher across all time points (p<0.0001), within 1 year prior to diagnosis (p=0.0025), and within 1 year after diagnosis (p<0.0001) in AD versus controls by Mann–Whitney U test (Figure 1A).

Table 3: Clinical validation through electronic medical record from STRIDE by Chi-square test

Time Frame	**Gender**	**OR (95%CI)**	**% in each cohort having an abnormal high ESR (%, AD vs. Control)**	**P (Chi-square)**	**# of AD**	**# of Control**	**Total #**
All time points, irrespective of diagnosis timing	F	1.85 (1.32-2.61)	53.33% vs. 38.15%	0.0003	135	8769	8904
	M	1.42 (0.91-2.23)	56.41% vs. 47.60%	0.1216	78	6271	6349
ESR testing 1 year prior our first diagnosis	F	2.31 (1.06-5.01)	44.74% vs. 25.96%	0.032	38	104	142
	M	2.41 (0.94-6.18)	54.17% vs. 32.88%	0.0625	24	73	97
ESR testing 1 year after our first diagnosis	F	3.49 (1.93-6.31)	69.23% vs. 39.20%	<.0001	52	3194	3246
	M	1.79 (0.83-3.84)	52.79% vs. 66.67%	0.1302	30	2487	2517

In males, patient with high ESR (54%) (> 20 mm/hr) show a trend towards association with AD within 1 year prior to our patients' first AD diagnosis code (OR: 2.41 [0.94-6.18], p=0.0625) (Table 3). ESR values were overall significantly higher compared with control (p=0.0198) by Mann–Whitney U test (Figure 1B).

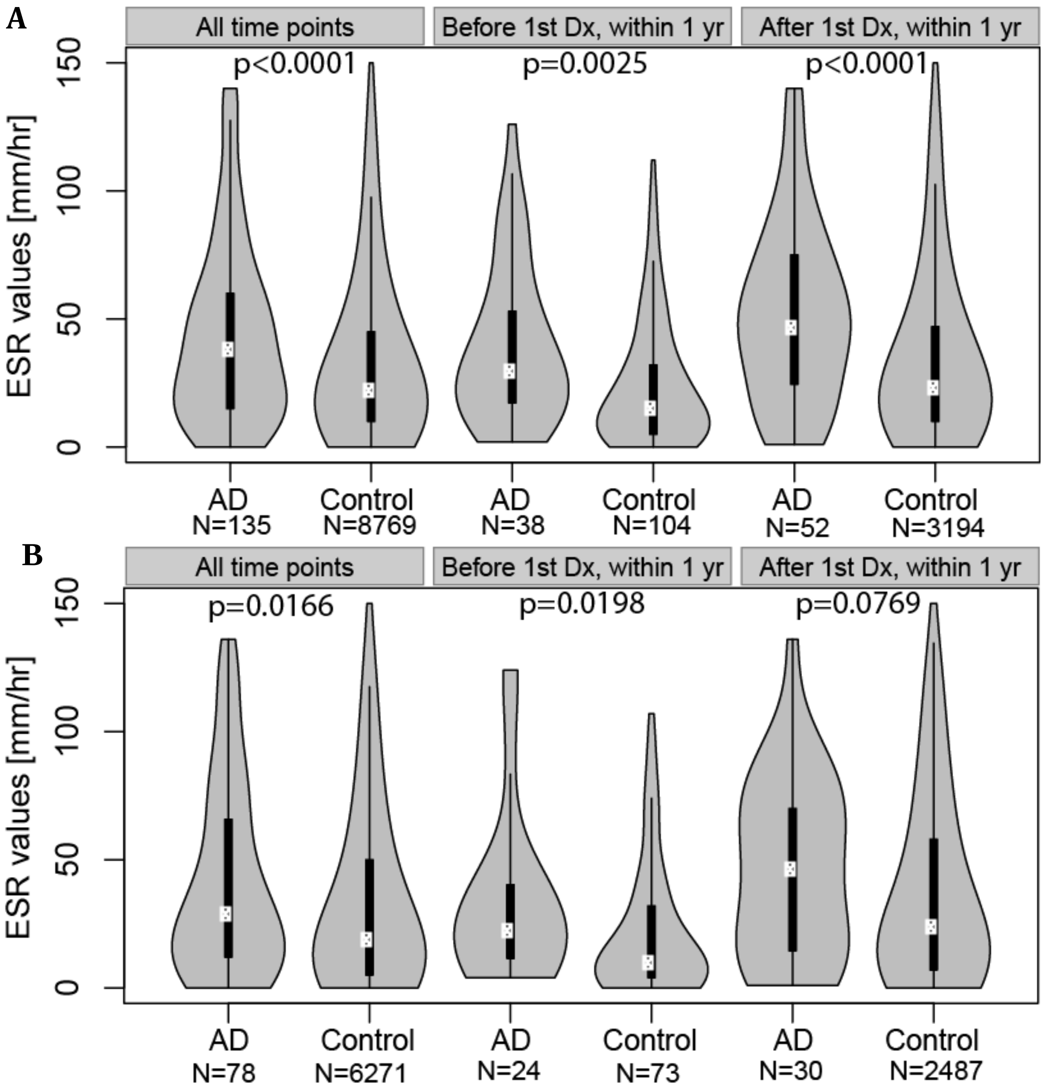
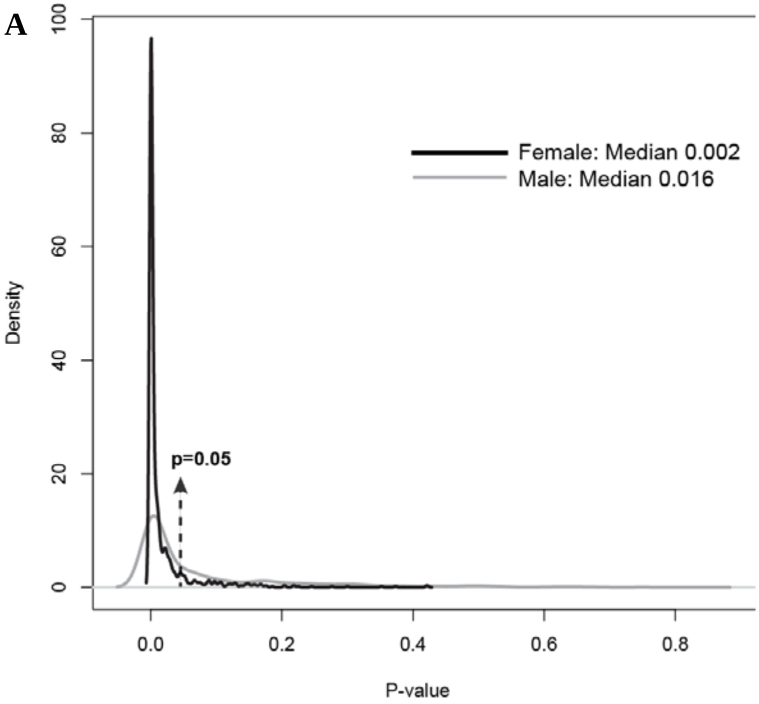


Figure 1: Violin plots (combination a boxplot and a kernel density plot) for ESR associated with AD overall time points, within 1 year lab tested prior to the 1st diagnosis, within 1 year lab tested after the 1st diagnosis for female (1A) and male (1B). In the black box plots, the bold black line boundaries indicate the 25th, 75th percentiles of ESR values, and white center squares indicate the median value of ESR. The outside grey shapes indicate density of the number of samples. P-values are reported by Mann–Whitney U test.

To match the ages of control patients to AD patients, we also performed a random sampling method to randomly select the same number of patients from controls whose ages fit the same distribution to the ages of the AD patients. As ESR is known to have values ranging from zero to higher, and with zero known to be the most frequently resulted normal value, we calculated a one-side p-value from T test by evaluating whether the mean of the lab value is higher in the AD patients, and repeated the process 1,000 times to generate the p-value distribution. We again tested for ESR values measured within 1 year prior to the 1st diagnosis and after the 1st diagnosis using the random sampling method to match the ages between control and AD cohorts. For instance, we randomly selected 38 female control patients matching the ages in our female AD cohort, where both cohorts had a measurement of ESR within 1 year prior to the 1st diagnosis. For within 1 year lab tested prior to the 1st diagnosis, the median p-values are 0.002 for female and 0.016 for male. For within 1 year lab tested after the 1st diagnosis, the median p-values are 0.025 for female and 0.161 for male. The distributions of p-values for prior to the 1st diagnosis and after the 1st diagnosis were shown in Figure 2A and Figure 2B. With the ESR being higher in AD cohorts compared to selected age-matched controls, this suggests that ESR might not be significantly confounded by age in our study.



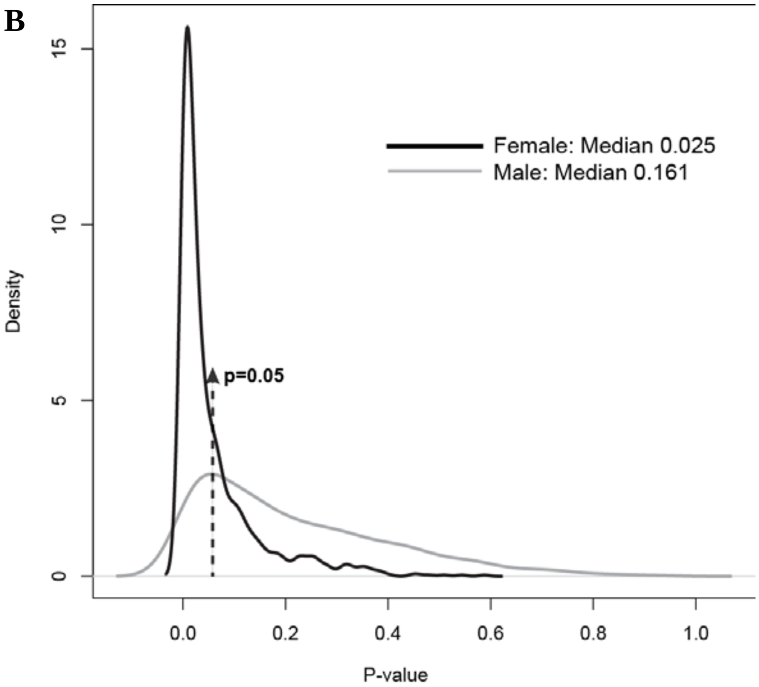


Figure 2: P-value distribution comparing age-matched AD and control groups for female (black curve) and male (grey curve) with 1,000 random samplings. ESR lab values within 1 year prior to our 1st diagnosis (2A), and ESR lab values within 1 year after our 1st diagnosis (2B). Dash line with arrow indicates p-value = 0.05.

4. Discussion

We developed a systematic approach to identify genetic associations between traits and diseases susceptibility based on common genetic architecture, aiming at identifying potential novel prognostic or risk markers for disease. In this study we focused on traits associated with Alzheimer's disease (AD) as a proof of concept, and we identified 6 clinical traits associated with AD. Five of these traits were known but one was a novel finding. We retrospectively validated our novel finding using EMR data from more than 15,000 patients at SHC.

We observed a significant association between ESR and AD, especially in female patients above 50 years old. Female patients who had abnormally elevated ESR levels had 2.31 higher chance of developing subsequent AD within a year of that lab test, compared to control patients, indicating ESR is a risk factor to AD that could be tested in a prospective trial for AD prognosis. A previous study has also reported the increased trend for ESR in AD female, although it did not reach significance due to a very small sample size (*35*). Moreover, we found that ESR persists in its elevation in female patients diagnosed with AD, suggesting that inflammation may play a role in the pathophysiology of AD (*36*), but we cannot rule out its elevation as secondary due to therapy of AD. A possible mechanism involve the complement gene inflammatory pathway including *C3*, *C4* and *C1Q* (*27-29*) as *CR1* was in common with ESR, currently used as a non-specific inflammation marker. If ESR proves to be a useful marker in specific prospective trials, we would also suggest that patients diagnosed with AD could be closely monitored for ESR as a trigger for intervention modification, such as adjusting non-steroidal anti-inflammatory medications (*36*). We would suggest that a robust prediction model could be developed combining ESR and other current risk factors including age, lipid panel, and environmental factors and validated using multi-center EMR data, then further validated in a prospective study.

Presently, the small sample size for the case cohort represents the limiting factor for a broader implication. We acknowledge that AD patients are relatively older than the control and the control are not exactly matched, as we used a retrospective study design based on our EMR, and not a randomized prospective trial.

Though we showed 6 significant traits with q-value ≤ 0.01, we acknowledge that threshold parameters could be altered. For example, the seventh trait on our list associated with AD would be high-density lipoprotein cholesterol (HDL-C) level, with q = 0.011. A recent study has shown that higher levels of HDL-C were indeed associated with a decreased risk of both probable and possible AD compared with lower HDL-C levels (*37*).We could increase our significance cutoff for more novel findings. However, in this study, we used a well-accepted stringent q-value cutoff from random shuffling to avoid identifying false positive.

We do acknowledge that our discovered association and validation cannot fully distinguish the causal direction of the association or if a single associated mutation in a shared gene systematically influences both phenotypes. Regardless, we do suggest that the strategy we adopted here captures and exploits relevant genetic association between disease and traits. The approach described here could in theory be applied to any disease in order to refine their risk factors model. Investigating clinical traits that share genetic architecture with a disease, and validating these traits through EMR data is a powerful and efficient way to identify risk factors, prognostics, and diagnostic markers for complex disease.

5 Acknowledgments

This work was supported by Lucile Packard Foundation for Children's Health. We thank Chirag Patel for discussion and suggestions to this work. We thank Alex Skrenchuk for computer cluster IT support.

References

1. A.M. Minino *et al.*, *Natl Vital Stat Rep* **59**, 1 (Dec 7, 2011).
2. L.E. Hebert *et al.*, *Arch Neurol* **60**, 1119 (Aug, 2003).
3. J. Povova *et al.*, *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub* **156**, 108 (Jun, 2012).
4. K. Andersen *et al.*, *Neurology* **53**, 1992 (Dec 10, 1999).
5. K. Blennow *et al.*, *Lancet* **368**, 387 (Jul 29, 2006).
6. T.D. Bird. in *GeneReviews* (eds. Pagon, R.A. *et al.*) (Seattle (WA), 1993).
7. L. Fratiglioni *et al.*, *Ann Neurol* **33**, 258 (Mar, 1993).
8. S.T. Pendlebury *et al.*, *Lancet Neurol* **8**, 1006 (Nov, 2009).
9. R.A. Whitmer *et al.*, *Neurology* **71**, 1057 (Sep 30, 2008).
10. A. Solomon *et al.*, *Dement Geriatr Cogn Disord* **28**, 75 2009).
11. H.C. Hendrie *et al.*, *Alzheimers Dement* **2**, 12 (Jan, 2006).
12. M. Kivipelto *et al.*, *Arch Neurol* **62**, 1556 (Oct, 2005).
13. T.M. Sivanandam *et al.*, *Neurosci Biobehav Rev* **36**, 1376 (May, 2012).
14. , *Nature* **447**, 661 (Jun 7, 2007).
15. A.D. Johnson *et al.*, *BMC Med Genet* **10**, 6 2009).
16. U.P. Steinbrecher *et al.*, *Arterioscler Thromb* **12**, 608 (May, 1992).
17. K.I. Goh *et al.*, *Proc Natl Acad Sci U S A* **104**, 8685 (May 22, 2007).
18. H.J. Lowe *et al.*, *AMIA Annu Symp Proc* **2009**, 391 2009).
19. R. Chen *et al.*, *PLoS One* **5**, e13574 2010).
20. R. Chen *et al.*, *PLoS Genet* **8**, e1002621 (Apr, 2012).
21. C.J. Patel *et al.*, *Bioinformatics* **28**, i121 (Jun 15, 2012).
22. S. Suthram *et al.*, *PLoS Comput Biol* **6**, e1000662 (Feb, 2010).
23. R. Chen *et al.*, *Nat Methods* **4**, 879 (Nov, 2007).
24. H.C. Wu *et al.*, *Acm Transactions on Information Systems* **26**, 2008).
25. J.D. Storey *et al.*, *Proc Natl Acad Sci U S A* **100**, 9440 (Aug 5, 2003).
26. R. Ihaka *et al.*, *Journal of computational and graphical statistics* **5**, 299 1996).
27. S.K. Nadar *et al.*, *J Hum Hypertens* **21**, 261 (Apr, 2007).
28. A. Muscari *et al.*, *Am J Med* **98**, 357 (Apr, 1995).
29. M. Naesens *et al.*, *J Am Soc Nephrol* **20**, 1839 (Aug, 2009).
30. M. van Oijen *et al.*, *Ann Neurol* **59**, 139 (Jan, 2006).
31. S.E. O'Bryant *et al.*, *J Geriatr Psychiatry Neurol* **23**, 49 (Mar, 2010).
32. P. Caramelli *et al.*, *Acta Neurol Scand* **100**, 61 (Jul, 1999).
33. T. Matsuzaki *et al.*, *Neurology* **77**, 1068 (Sep 13, 2011).
34. G.T. Lesser *et al.*, *Dement Geriatr Cogn Disord* **27**, 42 2009).
35. D. Robinson *et al.*, *Journal of the American Geriatrics Society* **43**, 1177 1995).
36. E.E. Tuppo *et al.*, *Int J Biochem Cell Biol* **37**, 289 (Feb, 2005).
37. C. Reitz *et al.*, *Arch Neurol* **67**, 1491 (Dec, 2010).

A CORRELATED META-ANALYSIS STRATEGY FOR DATA MINING "OMIC" SCANS[*]

MICHAEL A PROVINCE

Division of Statistical Genomics, Washington University School of Medicine, Box 8506, 4444 Forest Park Blvd
St. Louis, MO, 63105, USA
Email: mprovince@wustl.edu

INGRID B BORECKI

Division of Statistical Genomics, Washington University School of Medicine, Box 8506, 4444 Forest Park Blvd
St. Louis, MO, 63105, USA
Email: iborecki@wustl.edu

Meta-analysis is becoming an increasingly popular and powerful tool to integrate findings across studies and OMIC dimensions. But there is the danger that hidden dependencies between putatively "independent" studies can cause inflation of type I error, due to reinforcement of the evidence from false-positive findings. We present here a simple method for conducting meta-analyses that automatically estimates the degree of any such non-independence between OMIC scans and corrects the inference for it, retaining the proper type I error structure. The method does not require the original data from the source studies, but operates only on summary analysis results from these in OMIC scans. The method is applicable in a wide variety of situations including combining GWAS and or sequencing scan results across studies with dependencies due to overlapping subjects, as well as to scans of correlated traits, in a meta-analysis scan for pleiotropic genetic effects. The method correctly detects which scans are actually independent in which case it yields the traditional meta-analysis, so it may safely be used in all cases, when there is even a suspicion of correlation amongst scans.

[*] This work is partially supported by grants AG023746, HL088215, DK075681 and DK089256 from the USA National Institutes of Health

1. Introduction

Meta-analysis is becoming a common tool for integrating findings across multiple OMIC scans (e.g. Hsu et al., 2010; Moutselos et al., 2010). The advantages are most obvious when investigators do not have access to all of the source data, but only to summary results from each study. But such a meta analysis strategy is sometimes analytically preferred even when all of the individual level data are available, in situations where there is enough potential for study heterogeneity that a combined supermodel, mega-analysis would require estimation of many cross-study-by-covariate interaction terms (e.g. Ioannidis et al., 2002).

However, one potential problem with such meta-analyses is the danger of hidden non-independencies between elements of the scans that can occur when data are generated with overlapping subjects, related subjects, or other information. For example, there are overlapping subjects in several large scale NIH sponsored genetic epidemiology studies, such as the Framingham Heart Study, the NHLBI Family Heart, the HyperGen study, ARIC study, etc. Even if subjects are distinct across studies, if there are closely related subjects (e.g. siblings) across studies, this can cause non-independence of the observations and potentially inflate type I error in meta-analyses. The reason is that such non-independence violates the basic i.i.d. (independent and identically distributed) random variables assumptions of most statistical tests and models, including traditional meta-analysis ones, so that if a type I error (false positive) occurs in one study, and there is overlapping information to another study, then the other study is more likely to reflect this same false-positive trend in its corresponding result. A meta-analysis which ignores this fact, will take the reinforcement of "signals" between the two studies as a sign of independent replication, and overstate the significance of the meta-findings.

Conneely and Boehnke (2010) provide a method of conducting meta-analysis of correlated SNPs within an LD region, on multiple correlated traits, but they assume that the multiple studies that are being meta-analyzed are strongly independent, and they do not consider the possibility of overlapping subjects or correlated information between the OMIC scans. Riley et al. (2007) discuss the properties a bivariate random effects meta analysis model, in which they estimate what they call "between study correlation," but it is clear from their hierarchical model that they are in fact making the assumption that studies are strongly independent of one another. Their "between study correlation" is actually the correlation between the true parameter values within a study, as distinct from what they call the "within study correlation" which is the correlation between the pair of estimates of the parameters for each study. So they are not modeling the kinds of across scans correlations that would arise from overlapping subjects or any of the other reasons that we consider in our correlated meta-analysis model. Lin and Sullivan (2009) provide an efficient method for analyzing overlapping subjects in multiple GWAS to avoid inflation of type I error, but their approach is only applicable to case/control data, not quantitative traits, and it requires either having access to all individual level data or at least having a complete census accounting of the exact numbers of overlapping cases and controls. Sometimes such information is not easily shared amongst studies, due to IRB concerns, and sometimes, such overlaps may not even be known to the investigators, as subjects may not always volunteer that they are participating in multiple studies. Turchin and

Hirschhorn (2012) have provided a clever way to forensically detect overlapping subjects using cryptographic hashes on GWAS data that preserves confidentiality of subjects. But as stated above, overlapping subjects is not the only cause for non-independence.

Hartung (1999) proposed the first true correlated meta-analysis test, using an approach similar to ours based upon the inverse normal MVN. But his approach estimates a single dependency correlation amongst all scans, assuming they are all equally correlated, which can be problematic when some pairs of OMIC scans are more related than others. Additionally, his method estimates this single correlation for each hypothesis (or "OMIC unit", below) separately. This works well under the null hypothesis OMIC units, avoiding accumulation of evidence from correlated false-positives which results in inflation of type I error. But this approach can overcorrect for those OMIC unit tests under either the complete alternative hypothesis or partial alternative (incomplete null), resulting in potential loss of power. There, we want correlated true positive evidence to accumulate—we do not want to correct it out. Our approach, first proposed by Province in 2005 for combining linkage scans, estimates the complete MxM correlation matrix for M OMIC scans, and thus allows for different scans to be correlated at different levels. We also estimate the average study dependency correlation matrix across all OMIC units in a set of scans, exploiting the biological fact that most of the OMIC units will be under the null. Thus, our method is more likely to only be correcting for dependencies under the null, retaining power under the alternative (or partial alternative) OMIC units.

2. Methods

2.1. *OMIC unit of inference*

Definition: An "OMIC unit of inference" is the basic unit for which statistical testing is performed for a particular OMIC scan.

For example, in a micro-array experiment, the OMIC unit of inference would be the gene, since the scan would consist of one statistical test for each of the 20,000 genes on the array. In a proteomic scan, the OMIC unit would be "proteins", since we have one statistical test for each measured protein. In a linkage scan, the OMIC unit of inference might be the linkage markers themselves, or it might be centimorgan locations equally spaced throughout the genome, at which Identity-By-Descent (IBD) estimates have been made for each relative pair. There, we would have one multipoint LOD-score for each cM location. In a Whole Exome Sequencing (WES) experiment where the goal is to find rare variants, the OMIC unit of inference could be the variants themselves if power is sufficient to support individual testing of rare variants. But often there is not enough power to detect rare variants at an individual variant level (unless the variants are unusually penetrant). More commonly a statistical burden test is applied, so that the OMIC unit inference would be the gene, not the variant. Even though the smallest unit of measurement in the WES is the variant, the statistical tests are conducted at the gene level not the variant level, by collapsing/weighting all exonic variants within the same gene into a single composite predictor for that gene. So the gene is considered the OMIC unit of inference in this case. In a GWAS, the most natural OMIC unit of inference would be SNPs genotyped on the GWAS chip. But this might be reduced to a gene-level OMIC unit of inference, by taking, say, only the most significant SNP for each gene. Or it might be expanded to include all SNPs catalogued in a standard reference panel, such as HapMap or 1000 Genomes, by first performing genetic imputation (estimating the unmeasured SNPs via haplotype matching) and then conducting statistical tests on each imputed as well as measured SNP.

2.2. *Harmonization of OMIC units of inference and missing data patterns*

In order to conduct any meta-analyses of multiple OMIC scans we must first put them all on a common OMIC unit of inference scale, so that we can see if the combined evidence reinforces or destroys the overall signal at that particular OMIC unit. Exactly how this is done depends very much on the scientific goals, the types of OMIC scans one wishes to meta-analyze, the granularity and extent of the available data, and many other factors which we will not address in this paper. It is important to note that the methods we propose here do not depend upon the details of the process by which harmonization of OMIC units of inference across scans is accomplished. Nor is it necessary that this is accomplished comprehensively with identical OMIC units of inference across all OMIC scans. We can in fact have quite complex patterns of missing data within and across the OMIC scans, and our method will still apply. We do not make the strong assumption that data are missing "at random", but instead, we make the slightly weaker assumption that the missing data patterns are "ignorable." For example, we may wish to meta-analyze "I" different expression array experiments along with "K" different GWAS scans in combination with "J" linkage scans, and "M" WESs. We can reduce each of these I+J+K+M scans at the gene OMIC unit of inference, by taking the "most significant" SNP/gene in the GWASes, the highest LOD score over each gene in the linkage scans, and use burden tests for each gene in the exome scans. But we may also be interested in going beyond the genes into the intergenic regions, leaving out the expression arrays, and meta-analyzing the J+K+M GWAS+linkage+exome scans. We might define intergenic OMIC units as contiguous regions of open chromatin defined in functional experiments, contiguous regions of high species conservation. Or we might include some of the genes from the "I" expression arrays for those regions that are "known" to play regulatory role for the genes (e.g. eQTL regions for the gene).

2.3. *Correlated Meta-Analysis of p-values*

Suppose we have conducted N different OMIC scans on M common OMIC units of inference. Let $\underline{P}_{NxM}=(p_{ij})$ for i=1,…,N and j=1,…,M be the NxM matrix of p-values for the N scans of across the M OMICs units For each i, j let Z = Φ-1(1-P) denote the element-wise monotonic "complement probit" transformation of p-values to z-scores, where Φ(z) denotes the cumulative distribution function of the unit normal, N(0,1), i.e. $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2} dx$

For each row "j," corresponding to a particular OMIC unit of inference, we look at the Nx1submatrix formed by taking only the jth -row of $\underline{P}_{NxM}$, and denote this vector by $\underline{P^{(j)}}_{Nx1}$. It is this set of p-values that we wish to meta-analyze to test the combined effect of the jth OMIC unit across all N scans.

We apply a theorem from multivariate normal (MVN) distribution theory, whose proof is found in Anderson (2003):

Theorem 1: If $\underline{Z}_{kx1}$ is a MVN random variable, $\underline{Z}_{kx1} \sim N[\underline{\mu}_{kx1},\Sigma_{kxk}]$, and $\underline{D}_{1xk}$ is any vector of constants, then the linear combination $\underline{D}_{1xk}\ \underline{Z}_{kx1} \sim N[\underline{D}_{1xk}\ \underline{\mu}_{kx1},\ \underline{D}_{1xk}\ \Sigma_{kxk}\ \underline{D'}_{kx1}]$. In particular, when k=N, $\underline{D}_{1xN} = \underline{1}_{1xN}$ (i.e. the 1xN vector of all "1"s) and $\underline{\mu}_{Nx1} = \underline{0}_{Nx1}$ (i.e. the Nx1 vector of all "0"s), then out meta-analysis Z-value is given by

$$Z_{meta} = \sum_{i=1}^{N} Z_i = SUM(\underline{Z}_{Nx1}) = \underline{D}_{1xN}\ \underline{Z}_{Nx1} \sim N[0, \underline{D}_{1xN}\ \Sigma_{NxN}\ \underline{D'}_{Nx1}] = N[0, SUM(\Sigma_{NxN})] \qquad (1)$$

We then convert back from the z-score scale back to the p-value scale using the monotonic inverse transformation to the one above, to get the meta-analysis p-value: $P_{meta} = 1 - \Phi(Z_{meta})$. Note that if the jth OMIC unit does not have a significance test result for one or more of the N scans, then the corresponding entries in $\underline{P^{(j)}}_{Nx1}$ will be missing (this will happen if data are low quality for that OMIC unit in one or more scans or if OMIC unit harmonization is not complete across scans for whatever reason). In such cases, we may use the basic property of the MVN distribution that every sub-dimensional space is also MVN distributed. Specifically, if $N_j \leq N$ is the number of non-missing p-values for the jth OMIC unit, and we denote by $\underline{P^{(j)*}}_{Njx1}$ the N_jx1 submatrix of $\underline{P^{(j)}}_{Nx1}$ with all missing p-value rows deleted, then Theorem (1) still applies to $k=N_j$, the corresponding sub-dimensional components of $\underline{D^{(j)*}}_{1xNj}$, $\underline{\mu^{(j)*}}_{Njx1}$ and $\Sigma^{(j)*}_{NjxNj}$ being the non-missing submatricies of $\underline{D}_{1xN}$, $\underline{\mu}_{Nx1}$ and Σ_{NxN}, respectively.

We illustrate the application of this theorem to the meta-analysis of OMIC scans with two extreme mathematical examples.

Example 1: k independent OMIC scans. For each common OMIC unit, j=1, 2, ..., M we denote the "k" p-values at that jth OMIC unit by the kx1 vector $\underline{p^{(j)}}_{kx1} = (p^{(j)}_1, p^{(j)}_2. ..., p^{(j)}_k)'$. We transform these elementwise to z-scores via the complement probit transformation given in Equation (1), above, so that $\underline{Z^{(j)}}_{kx1} = (Z^{(j)}_1, Z^{(j)}_2. ..., Z^{(j)}_k)'$. Then $\forall$ j=1, 2, ..., M we have $\underline{Z^{(j)}}_{kx1} \sim N(\underline{0}_{kx1}, ,\Sigma_{kxk} = I_{kxk}]$, where I_{kxk} is the kxk identity matrix, so that $Z^{(j)}_{meta} = \sum_{i=1}^{k} Z^{(j)}_i \sim N[0,k]$

Example 2: **k completely correlated (equal) scans.** $\underline{Z}_{kx1} = (Z_1, Z_2. ..., Z_k)'$. Then

$\underline{Z}_{kx1} \sim N(0, ,\Sigma_{kxk} = \mathbb{1}_{kxk}]$, where $\mathbb{1}_{kxk}$ is the kxk matrix, with all elements equal to "1.". Because each $Z_i \sim N[0,1]$, then for each i, $Z_i = Z_1$ (i.e., they are actually all equal), so that

$$VAR\left[\sum_{i=1}^{k} Z_i\right] = VAR[kZ_1] = k^2 VAR[Z_1] = k^2 \times 1 = k^2$$

which is the same result we get from applying Theorem 1. $VAR\left[\sum_{i=1}^{k} Z_i\right] = sum(\Sigma_{kxk}) = sum(1_{kxk}) = k \times k = k^2$

These two examples represent the extreme cases and make sense. In Example 1, when all scans are really independent, the Σ_{kxk} matrix is the identity, and our correlated meta-analysis method is the same as the traditional meta-method. In the other extreme, in Example 2, when all scans are completely correlated, the Σ_{kxk} matrix becomes the 1_{kxk} matrix, and there is really just one scan, so the meta-analysis should recognize this and just return the original scan, which our correlated meta-method does, with no inflation of type I error.

2.4. *Estimating Σ_{kxk} from the tetrachoric correlations amongst the OMIC scans themselves*

Unlike the two simple examples above, in practice with real data, we will not know the values of the Σ_{kxk} matrix. However, if entire OMIC scans are available to us, we can exploit this fact to estimate Σ_{kxk} , by making using of the following assumption:

Assumption 1: In any OMIC scan, by far (in fact, by several orders of magnitude) most of the statistical tests will be under the NULL hypothesis --.e. the OMIC units of inference are actually statistically independent of the phenotype being scanned.

With this assumption, we can use the <u>observed correlations</u> between OMIC scans to obtain our estimate of Σ_{kxk}., and then apply Theorem 1 to conduct our correlated meta-analysis.

Note that this is a <u>biologically</u> motivated assumption, which stems from our understanding of the OMIC architecture of phenotypes and traits, i.e. that for any fixed trait, most of the genome, exome, proteome, etc. is neutral. There may be (hopefully is) some "contamination" of the alternative hypotheses somewhere within the OMIC scan (otherwise, our scans are fruitless). But OMIC analysis experience tells us (as does population genetic theory), that the number of true OMIC signals should be far outnumbered by the number of noise OMIC units of inference. Nonetheless, if we simply estimate Σ_{kxk} across scans between corresponding OMIC units there are two problems. The first is that the p-value scale is uniformly distributed under the null hypothesis and we are assuming that we are dealing with a MVN distribution in Theorem 1. But this can be easily handled by making use of the complement probit transformation discussed above. The second, bigger problem is that the contamination of the OMIC units that are under the alternative should not so easily be dismissed as trivial. Yes, Assumption 1 tells us that they are small in frequency, but it says nothing about their impact on the estimate of Σ_{kxk}. Correlated, highly significant true signal results may be small in number but highly influential on the estimate of Σ_{kxk}. Worse, we do not want to over-estimate Σ_{kxk}. because we are downweighting the results of our meta-analysis by its magnitude. Σ_{kxk} is supposed to be estimating the correlation between OMIC scans <u>only for those OMIC units under the H0</u>, because we want to avoid accumulating evidence across highly correlated scans that are only due to correlations of type I errors (due to overlapping subjects, relatedness, etc.). But we do NOT want to down weight evidence at OMIC units that are under the alternative hypothesis. In fact, we want the meta-p-values to be as significant as possible here. But we do not know which OMIC units are under the null and which are under the H1 (or we would not be doing the meta-analysis in the first place). We can minimize the impact of this contamination of the alternative hypothesis by using the tetrachoric correlation instead of the Pearson correlation in Σ_{kxk}. To do this, we first truncate all of the individual scan Z scores into two categories ($Z \leq 0$) vs ($Z > 0$). Then, at each common OMIC unit, if there are M scans, we form the M dimensional 2x2x…x2 table of scores across all scans, from which we estimate the tetachoric correlation matrix Σ_{kxk}. The tetrachoric correlation is less sensitive to contamination from the alternative hypothesis, because it lumps them with all moderately significant and even non-significant findings at the $P<0.5$ level. Thus, it provides some protection for over correction of the correlation amongst OMIC scans.

3. Results

3.1. *Simulations*

To validate the correlated meta-analysis method, we performed a series of simulation experiments. We generated 100 simulation replications of 3 OMIC scans on 10,000 OMIC units each, with a known correlational structure. For each replication, we conducted both the traditional meta-analysis (which assumes the 3 scans are independent) as well as our correlated meta-analysis procedure, which estimates the tetrachoric correlations between scans of the truncated p-values across the OMIC units of inference, and then uses that correlation matrix to correct the meta-analysis inference, as described above. The results of our simulations are shown in Table 1, where we catalog the distribution of estimates across the 100 simulation experiments, and Figure 1, where we show the Quantile-Quantile (Q-Q) plot for a typical (the first) replication's meta-analysis of the 3 scans.

Table 1: Distributions (Mean, Min, Max) across 100 simultation replications of parameter estimates from Traditional vs. Correlated meta-analyses of 3 OMIC scans

Parameter	**Expected Value**	**Parameter Estimates using Traditional Meta**			**Parameter Estimates using Correlated Meta**		
		Mean	**Min**	**Max**	**Mean**	**Min**	**Max**
3 Independent OMIC Scans							
$\rho(Z_1,Z_2)$	0	[0]	[0]	[0]	-0.00082	-0.04044	0.04886
$\rho(Z_1,Z_3)$	0	[0]	[0]	[0]	0.00121	-0.04380	0.03524
$\rho(Z_2,Z_3)$	0	[0]	[0]	[0]	0.00066	-0.04143	0.03101
$\mu(Z_{Meta})$	0	0.00016	-0.02125	0.02762	0.00160	-0.02127	0.02761
$\sigma(Z_{Meta})$	1	1.00001	0.98363	1.01877	0.99977	0.97254	1.02056
3 Correlated OMIC Scans							
$\rho(Z_1,Z_2)$	0.5	[0]	[0]	[0]	0.50153	0.47154	0.52387
$\rho(Z_1,Z_3)$	0.2	[0]	[0]	[0]	0.20110	0.16289	0.23495
$\rho(Z_2,Z_3)$	0.9	[0]	[0]	[0]	0.89994	0.88774	0.91027
$\mu(Z_{Meta})$	0	0.00015	-0.02820	0.04023	0.00011	-0.01961	0.02798
$\sigma(Z_{Meta})$	1	1.43792	1.40103	1.46045	0.99983	0.98170	1.01774

As can be seen in Table 1, when the 3 scans are actually independent (top half of the table), our correlated meta-method correctly senses this and accurately estimates that the 3 tetrachoric correlations between the scans (ρ_{12}, ρ_{13} and ρ_{23}) are all nearly zero for all 100 replications (as evidenced by the fact that the mean, min and max are all roughly equal to one another and to the expected value of zero). Not shown in the table, we also compared the tetrachoric estimates of the correlations between scans, to the Pearson correlations on the continuous (non-truncated) transformed p-values, which should be more correct under the null. The average pairwise difference between the tetrachoric and pearson correlation estimates between the 3 scans in Table 1 across all simulation replications were -0.001, 0.0004, and -0.0003 for $\rho(Z_1,Z_2)$, $\rho(Z_1,Z_3)$, $\rho(Z_2,Z_3)$, respectively in the top half of the table (independent scans), and 0.001, -0.001, -0.002 for the bottom half of the table (dependent scans). More importantly, no tetrachoric correlation estimate differed from its corresponding Pearson correlation estimate by more than 0.037 in any simulation replication under any condition. Thus, in all cases (not just on average), the tetrachoric correlation provides accurate estimates of the correlation between scans under the null, but unlike the Pearson correlations, it reduces the impact of of correlations between true positives (for which we do NOT want to "correct away," instead we want such evidence to accumulate in favor of the alternative).

Figure 1: Q-Q plots under the H0 for replication 1 in a simulation of a meta-analysis of 3 OMIC scans

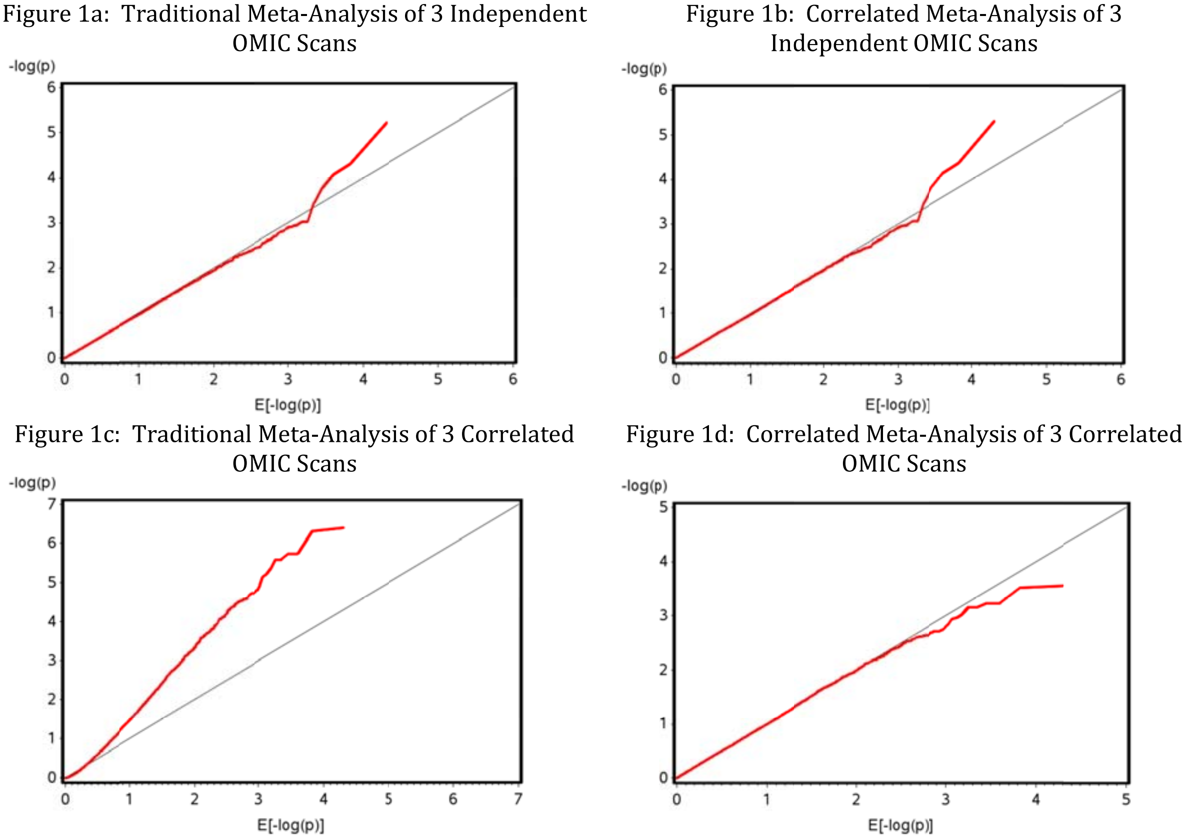


Figure 1a: Traditional Meta-Analysis of 3 Independent OMIC Scans

Figure 1b: Correlated Meta-Analysis of 3 Independent OMIC Scans

Figure 1c: Traditional Meta-Analysis of 3 Correlated OMIC Scans

Figure 1d: Correlated Meta-Analysis of 3 Correlated OMIC Scans

This results in little bias as well as little loss in power in our correlated meta-test as compared to the traditional meta-analysis (which fixes the 3 between study correlations to zero). Thus, both the tradtional and the correlated meta analysis produce meta-Z scores that are nearly normal and have mean nearly zero and variance nearly one, as expected. The Q-Q plots in Figures 1a and 1b (for the 3 independent scans), verify that there is no inflation of type I error in this case, but more importantly from the correlated meta-analysis view, there is no over conservatism. The p-values are just as true estimating the 3 correlations to be nearly zero as they are assuming them to be zero, so there is no harm in applying the correlated meta-analysis procedure even when the 3 scans really are independent. The correlated meta will tell us what it thinks are the correlations, and will correct for them, regardless of their magnitude.

For the 3 correlated OMIC scans (bottom half of Table 1), we generated the scans pairwise correlated at 0.5, 0.2 and 0.9, respectively, and then applied traditional as well as our correlated meta-analysis method. As can be seen, our correlated meta-method accruately estimated the 3 correlations using the tetrachoric approach, across all 100 replications. More importantly, our correlated meta-analysis method correctly produces Z-scores with the proper 0,1 first and second moments. Whereas the traditional meta-analysis, yields a meta-Z-score with a badly inflated variance (Standard Error is approximately 1.4). This results in the traditional meta being overly liberal, since it calculates P-values assuming it's meta-Z score has a variance of 1, instead of 1.4. The inflation is readily apparent in the Q-Q plot of Figure 1c, compared to 1d.

3.2. *Example Pleiotropy Scanning in the NHLBI Family Heart Study*

Table 2: Correlated vs. Traditional Meta-Analysis of GWAS SNPs to assess Pleiotropy of related traits in the NHLBI Family Heart Study

Table 2a: SNPS in the NEGR1 gene (chrom 1) for pleiotropy to BMI (Body Mass Index) and Waist Circumference (WC)

SNP	P-value BMI	Beta BMI	SE (beta) BMI	P-value WC	Beta WC	SE (beta) WC	P-value Traditional Meta	P-value Correlated Meta
rs577674	9.52E-06	0.58	0.13	8.38E-06	0.58	0.13	6.52E-10	1.59E-06
rs522451	9.96E-06	-0.59	0.13	9.90E-06	-0.59	0.13	8.01E-10	1.80E-06
rs473019	1.00E-05	-0.59	0.13	9.99E-06	-0.59	0.13	8.11E-10	1.81E-06
rs580626	1.00E-05	-0.59	0.13	1.00E-05	-0.59	0.13	8.16E-10	1.82E-06

Table 2b: SNPs in the ABCF2 gene (chrom 7) to assess Pleiotropy for pleiotropy to HOMA (a measure of Insulin Resistance) and Waist Circumference (WC)

SNP	P-value HOMA	Beta HOMA	SE (beta) HOMA	P-value WC	Beta WC	SE (beta) WC	P-value Traditional Meta	P-value Correlated Meta
rs12538823	8.96E-06	0.24	0.05	1.77E-04	0.18	0.05	1.36E-08	9.75E-08
rs7786151	9.84E-06	-0.24	0.05	1.92E-04	-0.18	0.05	1.61E-08	1.13E-07
rs12113924	4.15E-06	-0.25	0.05	2.28E-04	-0.18	0.05	8.97E-09	6.77E-08
rs3800794	4.04E-06	-0.25	0.05	2.29E-04	-0.18	0.05	8.84E-09	6.68E-08

Finally, we demonstrate the utility of our correlated meta-analysis procedure on a real data example, from a GWAS on the NHLBI Family Heart Study. In Table 2, we show the results of combining two GWASes from the same study on two highly correlated traits to look for pleiotropy. Note that this is an extreme example of overlapping subjects (ALL 2,767 of them overlap), so the Lin and Sullivan method would not be of much help here, but if we do the traditional meta-analysis, we are in real danger of having type I errors accumulate in this situation. For traits Body Mass Index (BMI) and Waist Circumference (WC), the tetrachoric correlation between the scans of these two highly correlated variables is estimated to be 0.70. Ignoring this correlation with the traditional meta-analysis, results in meta-P~10^{-10} for the SNPS listed which are far above the GW threshold for significance. However, the correlated meta-analysis method finds a more moderate level of evidence at P~10^{-6}, suggesting that the traditional analysis is very much inflating the evidence. On the other hand, for HOMA (a measure of insulin resistance based upon insulin/glucose levels) and WC, the estimated tetrachoric correlation is 0.14. Here the traditional and correlated meta-analyses are in better agreement that indeed, there appears to be genome-wide significant pleiotropy at this locus for these two traits.

4. Discussion

Our correlated meta-analysis method provides a simple, robust approach to integrate information across multiple OMIC scans so as to avoid inflation of type I error due to hidden dependencies. Our method makes few statistical assumptions. It first estimates empirically the degree of non-independence between the OMIC scans, and then uses this estimate to corrects the meta-inference. The method does not require any additional knowledge of numbers of overlapping subjects, nor any preliminary forensic analyses, and provides the correct type I error when scans are correlated, regardless of the number and character of its source causes, as part of the meta-analysis itself. It is applicable for combining OMIC scans on quantitative, qualitative or any combinations of phenotypes. It can even be used to scan for evidence of pleiotropic effects when subjects are completely overlapping. When OMIC scans actually are independent, it estimates this correctly, and becomes the same as the traditional meta-analysis test. Thus, the method can be safely used in any situation, when there is any doubt that there may be violations of study independence assumptions.

5. Acknowledgments

This work is partially supported by grants AG023746, HL088215, DK075681 and DK089256 from the USA National Institutes of Health

References

1. TW Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, NY ISBN-13: 978-0471360919 (2003)

2. Province MA. The significance of not finding a gene. Am J Hum Genet. 2001 Sep;69(3):660-3. Epub 2001 Jul 30. PubMed PMID: 11481587; PubMed Central PMCID: PMC1235495.

3. Lin DY, Sullivan PF. Meta-analysis of genome-wide association studies with overlapping subjects. Am J Hum Genet. 2009 Dec;85(6):862-72. PubMed PMID: 20004761; PubMed Central PMCID: PMC2790578.

4. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Med Res Methodol. 2007 Jan 12;7:3. PubMed PMID: 17222330; PubMed Central PMCID: PMC1800862

5. Conneely KN, Boehnke M. Meta-analysis of genetic association studies and adjustment for multiple testing of correlated SNPs and traits. Genet Epidemiol. 2010 Nov;34(7):739-46. PubMed PMID: 20878715; PubMed Central PMCID: PMC3070606.

6. Hsu YH, Zillikens MC, Wilson SG, Farber CR, Demissie S, Soranzo N, Bianchi EN, Grundberg E, Liang L, Richards JB, Estrada K, Zhou Y, van Nas A, Moffatt MF, Zhai G, Hofman A, van Meurs JB, Pols HA, Price RI, Nilsson O, Pastinen T, Cupples LA, Lusis AJ, Schadt EE, Ferrari S, Uitterlinden AG, Rivadeneira F, Spector TD, Karasik D, Kiel DP. An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits. PLoS Genet. 2010 Jun 10;6(6):e1000977. PubMed PMID:

7. Ioannidis JP, Rosenberg PS, Goedert JJ, O'Brien TR; International Meta-analysis of HIV Host Genetics. Commentary: meta-analysis of individual participants' data in genetic epidemiology. Am J Epidemiol. 2002 Aug 1;156(3):204-10. PubMed PMID: 12142254. 20548944; PubMed Central PMCID: PMC2883588.

8. Turchin MC, Hirschhorn JN. Gencrypt: one-way cryptographic hashes to detect overlapping individuals across samples. Bioinformatics. 2012 Mar 15;28(6):886-8. Epub 2012 Feb 1. PubMed PMID: 22302573; PubMed Central PMCID: PMC3307118.

9. Hartung J. A note on combining dependent tests of significance. Biometrical Journal, 1999, 41(7): 849-855.

10. Province MA (2005) Meta-Analyses of Correlated Genomic Scans, Genetic Epidemiology 29: 137

11. Moutselos K, Maglogiannis I, Chatziioannou A. Delineation and interpretation of gene networks towards their effect in cellular physiology- a reverse engineering approach for the identification of critical molecular players, through the use of ontologies. Conf Proc IEEE Eng Med Biol Soc. 2010;2010:6709-12. PubMed PMID: 21096082.

PHYLOGENOMICS AND POPULATION GENOMICS: MODDELS, ALGORITHMS, AND ANALYTICAL TOOLS

LUAY K. NAKHLEH
Department of Computer Science, Rice University, Houston, Texas, USA

NOAH A. ROSENBERG
Department of Biology, Stanford University, Stanford, California, USA

TANDY WARNOW
Department of Computer Science, University of Texas, Austin, Texas, USA

Advances in phylogenetics and population genetics have produced increasing awareness of the existence of problems of interest to both fields, and of problems for which approaches from one of the two areas can inform developments in the other. Phylogenetics and population genetics examine similar topics, but at different biological scales. Both fields analyze genetic similarities and differences among organisms, and the evolutionary processes that generate those similarities and differences. Whereas population genetics considers individuals and populations within species, phylogenetics focuses on relationships among species themselves. The two fields share a number of overlapping tools, as well as similar data in the form of biological sequences. Further, they come into direct contact in the analysis of populations that are sufficiently distantly related that they differ as much as distinct species, or species that are sufficiently closely related that they approach the level of population differences.

The increasing availability of genetic sequences within and among species has made the connections between phylogenetics and population genetics all the more apparent. *Phylogenomics* and *population genomics* have emerged as subjects concerned with phylogenetic and population-genetic problems at a genomic scale. The interface of phylogenetics, population genetics, and genomics has now generated significant research challenges, demanding new evolutionary *models* for linking population-genetic and phylogenetic time scales, new *algorithms* for analyzing data in contexts that involve both population-genetic and phylogenetic perspectives, and new *analytical tools* for assessing properties of the algorithms. We are pleased to present five papers that represent a range of topics that link phylogenomics and population genomics, and that demonstrate a range of ways in which models, algorithms, and analytical tools can be used to advance the subject.

The papers of James H. Degnan and Sebastien Roch address a traditional problem at the intersection of phylogenetics and population genetics: the modeling and inference of species phylogenies when incomplete lineage sorting generates discordance among gene trees. Roch takes a modeling approach to the comparison of several algorithms for species tree inference. In a three-species model, for each of the algorithms, Roch uses large-deviations theory to analytically determine the rate at which the probability of failing to infer the correct species tree decays for large numbers of loci. A higher decay rate indicates that a method has more favorable performance. The paper, which uncovers a substantial difference among methods, produces an innovative analytical framework that can potentially be used for studying methodological performance in greater generality.

Whereas Roch presents a general perspective for assessing multiple methods, the complementary paper of Degnan instead looks closely at a single species tree inference method, the STAR approach (for Species Tree estimation using Average Ranks). In this method, the internal nodes of each of a series of input gene trees are given discrete ranks. For two species, the rank of their most recent common ancestor is averaged across all gene trees, and a species tree is inferred from the matrix of average ranks for all pairs of species. Degnan determines that the node-numbering scheme of the original STAR method is only one among many in a family of sensible schemes. By allowing variations of STAR that select from among this family of numbering schemes, Degnan finds that STAR can be generalized and enhanced. Together, the Degnan and Roch papers provide advances in the development and evaluation of new methods for inferring species trees in the presence of incomplete lineage sorting.

Two additional papers, one by Md. Shamsuzzoha Bayzid, Siavash Mirarab, and Tandy Warnow, and the other by Yu Lin, Fei Hu, Jijun Tang, and Bernard M. E. Moret, study species tree inference under another form of gene tree discordance. Gene trees can disagree not only because of incomplete lineage sorting, but also as a consequence of gene duplications and losses that lead to errors in assigning orthology across species. Duplications and losses begin by a population-genetic process: a macromutation arises, carrying a duplication or loss, and that mutation eventually spreads in a population. Once fixed, the duplication or loss becomes a phylogenetic character useful for analyses of species relationships.

Bayzid *et al.* algorithmically investigate the inference of species trees in the presence of discordance due to gene duplication and loss. They study a pair of optimization problems, one considering only duplications, and the other also considering losses. These problems are known to be NP-hard. The authors formulate solutions via a max (or min) clique problem, and they provide theoretical results for solving certain constrained versions of the problems. They obtain exact solutions for the constrained problems, providing algorithms that run polynomially in the number of genes and the number of taxa. The max clique formulation has similarities to work of Than & Nakhleh in the context of incomplete lineage sorting, thereby highlighting connections among algorithms for different processes that generate gene tree discordance.

Lin *et al.* consider genome rearrangements and insertions in addition to duplications and losses. They introduce a method for inferring species trees from genome sequences, accounting for the various types of gene content and gene order differences observed across species. Their method relies on an encoding of the spatial orientation of genes, and it performs inference with maximum likelihood. Lin *et al.* test their method using simulations incorporating a variety of factors, such as different simulated species trees, different combinations of evolutionary events along the tree, and errors in genome assembly. After obtaining favorable performance in their simulations, Lin *et al.* apply their approach to obtain a phylogeny of 68 genomes. As was true for the Degnan and Roch papers, the Bayzid *et al.* and Lin *et al.* papers illustrate two complementary components of the development and evaluation of phylogenomic algorithms, with Bayzid *et al.* proving theorems pertaining to the performance of their method, and Lin *et al.* examining simulations and an empirical assessment.

The final paper, by Naama M. Kopelman, Lewi Stone, Olivier Gascuel, and Noah A. Rosenberg, presents an example of a different aspect of the intersection of phylogenomics and population genomics. Kopelman *et al.* investigate the consequences of using a method borrowed from phylogenetics—the neighbor-joining algorithm—in a specific population-genetic context, namely that of admixed populations. Motivated by peculiar observations seen for admixed populations in neighbor-joining trees, they study neighbor-joining applied to populations that follow an admixture model. Kopelman *et al.* provide a mathematical demonstration under special cases of the model that admixed populations are expected to appear toward the middle of neighbor-joining trees, often with short branch lengths. The paper illustrates how mathematical analysis of a phylogenetic algorithm can be performed in a population-genetic setting.

Together, this collection of five papers highlights both the variety of algorithmic, mathematical, and statistical approaches now under development for investigating the interface of phylogenomics and population genomics, and the variety of problems of interest to both subjects. The proliferation of phylogenomic and population-genomic data will only increase the attention given to these problems, and the importance of devising sound models, algorithms, and analytical tools for addressing them is only likely to increase.

Acknowledgments

We gratefully acknowledge the support of National Science Foundation Collaborative Research grants DBI-1062335 (TW), DBI-1062463 (LKN), and DBI-1146722 (NAR).

INFERRING OPTIMAL SPECIES TREES UNDER GENE DUPLICATION AND LOSS

M. S. BAYZID, S. MIRARAB and T. WARNOW*

Department of Computer Science, The University of Texas at Austin,
Austin, Texas 78712, USA
**E-mail: tandy@cs.utexas.edu*
www.cs.utexas.edu/users/tandy

Species tree estimation from multiple markers is complicated by the fact that gene trees can differ from each other (and from the true species tree) due to several biological processes, one of which is gene duplication and loss. Local search heuristics for two NP-hard optimization problems - minimize gene duplications (MGD) and minimize gene duplications and losses (MGDL) - are popular techniques for estimating species trees in the presence of gene duplication and loss. In this paper, we present an alternative approach to solving MGD and MGDL from rooted gene trees. First, we characterize each tree in terms of its "subtree-bipartitions" (a concept we introduce). Then we show that the MGD species tree is defined by a maximum weight clique in a vertex-weighted graph that can be computed from the subtree-bipartitions of the input gene trees, and the MGDL species tree is defined by a minimum weight clique in a similarly constructed graph. We also show that these optimal cliques can be found in polynomial time in the number of vertices of the graph using a dynamic programming algorithm (similar to that of Hallett and Lagergren[1]), because of the special structure of the graphs. Finally, we show that a constrained version of these problems, where the subtree-bipartitions of the species tree are drawn from the subtree-bipartitions of the input gene trees, can be solved in time that is polynomial in the number of gene trees and taxa. We have implemented our dynamic programming algorithm in a publicly available software tool, available at `http://www.cs.utexas.edu/users/phylo/software/dynadup/`.

Keywords: Gene Duplication and Loss; Incomplete Lineage Sorting; Clique.

1. Introduction

The estimation of species trees typically proceeds by concatenating multiple sequence alignments together for many genes and then estimating a tree on the resultant "super-matrix". These "combined analyses" require that all sequences be orthologous (hence each taxon should appear in each gene sequence alignment at most once), and assume that the true trees for the different genes are topologically identical. These two conditions can easily fail to hold when gene duplication and loss occurs, even when valiant efforts are made to estimate orthology. Thus, the estimation of species trees from gene trees that can differ due to gene duplication and loss,[2–6] especially when these gene trees contain more than a single copy of each taxon, requires more care.

Two of the most popular approaches for species tree estimation in the presence of gene duplication and loss are methods, such as iGTP[7] and DupTree,[8] that employ local search techniques to "solve" the NP-hard optimization problems MGD (Minimize Gene Duplication) and MGDL (Minimize Gene Duplication and Loss). For example, analyses based upon MGD and MGDL have been used in estimating species trees for snakes,[9] vertebrates,[10,11] *Drosophia*,[12] and plants.[13] These local search strategies are effective for relatively small numbers of taxa, but their utility for very large numbers of taxa has not been explored. In addition to local search techniques, exact solutions[14,15] and fixed-parameter tractable algorithms[1,16] have been proposed for addressing MGD and MGDL; however, to date these approaches have not been used as widely as the heuristic searches.

In this paper we will present a new approach for MGD and MGDL that does not use local search techniques or branch-and-bound techniques, but instead uses dynamic programming to produce an optimal solution within a user-specified subspace of the set of candidate species trees. Thus, by letting that subspace be all possible species trees we obtain a globally optimal solution for MGD or MGDL, while constraining the set allows us to obtain good (even if not globally optimal) solutions in polynomial time. While our dynamic programming approach is similar to that of Hallet and Lagergren,[1] our clique-based formulation of the problem is new, and many of our theoretical results are not explicitly proven in Hallett and Lagergren.[1]

The algorithmic technique we present is also related to the approach used in Than and Nakhleh[17] (see also Yu, Warnow, and Nakhleh[18]) for the MDC (Minimize Deep Coalescence) problem,[5] an optimization problem for species tree estimation in the presence of incomplete lineage sorting. In these papers, the optimal solution for MDC is characterized graph-theoretically, as follows. First, every binary rooted tree on n taxa can be represented by its set of "clusters", where a cluster is the set of taxa that appear below a node in the tree. Furthermore, two clusters are said to be "compatible" if and only if they can co-exist in a tree (equivalently, two clusters are compatible if and only if they are pairwise disjoint or one contains the other). To solve MDC, each possible cluster is represented by a node in a graph, and edges exist between pairs of nodes whose clusters are compatible. It is known that whenever a set of clusters is given that are all pairwise compatible, then a rooted tree exists with precisely that set of clusters. Thus, a set of $n-1$ pairwise compatible clusters, where n is the number of species, defines a binary rooted species tree for that set of clusters.

Than and Nakhleh[17] showed that it is possible to weight the nodes in the graph so that the total weight of any $(n-1)$-clique is the MDC score for the species tree defined by that clique, so that solving the MDC problem is equivalent to finding a minimum weight $n-1$ clique.

This problem formulation seems to be particularly expensive, since MaxClique is NP-hard and the graph has an exponential number of vertices, but Than and Nakhleh also showed that finding the minimum weight clique of size $n-1$ can be obtained in time that is polynomial in the number of nodes in the graph, using dynamic programming (DP). They also presented a "heuristic" version that only uses clusters that appear in the input gene trees, and so runs in polynomial time. This heuristic version produces highly accurate species trees,[17–19] suggesting that restricting the search space to clusters in the input trees is an effective strategy for MDC.

The approach we present here for optimizing MGD or MGDL builds on these ideas. We also build a graph, but the nodes of our graph correspond to "subtree-bipartitions", a generalization of clusters that we define in this paper. We show how to define weights on vertices in the graph so that the optimal solution to MGD is obtained by finding a minimum weight clique of size $n-1$, and we show how to find that clique using dynamic programming. This technique directly allows us to solve the constrained MGD problem, in which we constrain the species tree solution to have its subtree-bipartitions from a user-provided set; as with MDC, a DP algorithm solves this in polynomial time. We then show how to extend this to the MGDL problem, using the same graph but with different weights on the edges.

The rest of the paper is organized as follows. In Section 2, we present the theoretical foundations and terminology. We present theory and algorithms for solving MGD in Section 3, and results for MGDL in Section 4.

2. Basics

2.1. *Prior Terminology and Theory*

We begin by defining the MGD, MGDL, and MDC problems. The input to each problem is the same: a set $\mathcal{G} = \{t_1, t_2, \ldots, t_k\}$ of rooted binary gene trees, with leaves drawn from the set $\mathcal{X}$ of n taxa, and we allow the gene trees to have multiple copies of the taxa, and even to miss some taxa. The output of each problem is a species tree T on $\mathcal{X}$ minimizing $\sum_i d(t_i, T)$, where $d(t_i, T)$ is defined differently for each problem.

The original definitions for these problems assumed that the gene tree t_i had at least one copy of each taxon, and so these definitions need to be modified in order to handle incomplete gene trees, which have no copies of some taxon.

Handling incomplete gene trees: Most of the literature has handled the case of incomplete gene trees t_i as follows. Let T' be the tree obtained by restricting T to the leaf set of t_i and then suppressing all non-root nodes of degree two (i.e., T' is the homeomorphic subtree of T defined on the leafset of t_i). Then, T' is used instead of T when computing the MDC, MGD, or MGDL score. We call this the *restriction*-based approach, and hence define the restriction-based optimization problems MGD_r, $MGDL_r$, and MDC_r. (See Bayzid and Warnow[20] for another approach for handling incomplete gene trees.)

Optimal Embeddings for $MGD_r, MDGL_r$, ***and*** MDC_r.

An embedding of a rooted gene tree t into a species tree T is a mapping f from the nodes of the gene tree to the nodes of the species tree that has some natural properties: first, f maps leaves in the gene tree mapped to the unique leaf in the species tree with the same taxon label, and second, f maintains the order relationships in the gene tree. This second condition can be stated as follows: if v and w are nodes in the gene tree with v above w (meaning that v is on the path from w to the root of the gene tree), then $f(v)$ is above $f(w)$ within the species tree.

Let T be a rooted binary tree. We denote the set of vertices of a tree T by $V(T)$, the root by $root(T)$, the internal nodes by $V_{int}(T)$, and the set of taxa that appear at the leaves by $L(T)$. (Note that since T can have multiple copies of some taxa, it is possible for $|L(T)|$ to be smaller than the number of leaves in T.)

A *clade* in T is a subtree of T rooted at some node in T, and the set of leaves of the clade is called a *cluster*. We denote the cluster at v by $c_T(v)$; however, when the tree T is understood, we may also write $c(v)$. We denote the set of clusters of a tree T by $C(T)$.

The most recent common ancestor (MRCA) of a set A of leaves in T is denoted by $MRCA_T(A)$. Given a gene tree gt and a species tree ST, where $L(gt) \subseteq L(ST)$, we define $\mathcal{M} : V(gt) \rightarrow V(ST)$ by $\mathcal{M}(v) = MRCA_{ST}(c_{gt}(v)))$. In other words, $\mathcal{M}$ associates each node u of gt to the MRCA in ST of the cluster below u.

The optimal embedding for each of the three criteria we discuss (MDC_r, MGD_r, and

$MGDL_r$) is obtained using $\mathcal{M}$, even when the gene tree gt is incomplete (lacks some taxon) or contains more than one copy of some taxon.[5,6,17,21] Therefore, since the same reconciliation of a gene tree into a species tree optimizes all three criteria, we may refer to an "optimal reconciliation" without specifying the criterion. Also, for any given mapping, the calculation of the three scores can be performed in polynomial time. Therefore, given a set of rooted gene trees and a rooted species tree, we can calculate the $MGD_r, MGDL_r$, and MDC_r scores of the species tree in polynomial time.

Duplication nodes: For a rooted gene tree gt and a rooted species tree ST, where $L(gt) \subseteq L(ST)$, an internal node v in gt is called a *duplication node* if $\mathcal{M}(v) = \mathcal{M}(v')$ for some child v' of v, and otherwise v is a *speciation node*.[21–24]

Given a rooted, binary gene tree gt and a rooted, binary species tree ST such that $L(gt) \subseteq L(ST)$, $Dup(gt, ST)$ denotes the number of duplications needed to reconcile gt with ST under the $\mathcal{M}$ mapping. For a set $\mathcal{G}$ of rooted, binary gene trees, the notation $Dup(\mathcal{G}, ST)$ extends in the obvious way.

Gene losses: Let gt be a rooted, binary gene tree and ST a rooted, binary species tree such that $L(gt) \subseteq L(ST)$. The restriction of ST to $L(gt)$, denoted by $\mathcal{R}_{ST}(L(gt))$, is the smallest subtree of ST containing $L(gt)$ as its leaf set. The homeomorphic subtree $ST|_{L(gt)}$ of ST induced by $L(gt)$ is a tree obtained from $\mathcal{R}_{ST}(L(gt))$ by suppressing all nodes of $\mathcal{R}_{ST}(L(gt))$ with indegree and outdegree 1. We denote by r and l the two children of an internal node u. Then the number of gene losses for a given gene tree gt and species tree ST for a particular internal node u (under the restriction-based analysis), denoted by $loss_u$, can be calculated as follows:[21–24]

$$loss_u = \begin{cases} d(\mathcal{M}(r), \mathcal{M}(u)) + 1 & \text{if } \mathcal{M}(r) \subsetneq \mathcal{M}(u) = \mathcal{M}(l), \\ d(\mathcal{M}(r), \mathcal{M}(u)) + d(\mathcal{M}(l), \mathcal{M}(u)) & \text{if } \mathcal{M}(r) \subsetneq \mathcal{M}(u) \supsetneq \mathcal{M}(l), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Here $d(s, s')$ is the number of internal nodes in the path in $ST|_{L(gt)}$ from s to s'.

The number of gene losses (under the restriction-based analysis) is given by $loss(gt, ST) = \sum_{g \in V(gt)} loss_g$, while for a set $\mathcal{G}$ of rooted, binary gene trees, the number of losses is given by $loss(\mathcal{G}, ST) = \sum_{gt \in \mathcal{G}} loss(gt, ST)$. The number of duplications and losses (again, under the restriction-based analysis), denoted by $Duploss(\mathcal{G}, ST)$, is the sum of the number of duplication and losses, i.e., $Duploss(\mathcal{G}, ST) = Dup(\mathcal{G}, ST) + loss(\mathcal{G}, ST)$.

2.2. *New Data Structures*

Subtree-Bipartitions: Let T be a rooted binary tree and u an internal node in T. The *subtree-bipartition* of u, denoted by $\mathcal{SBP}_T(u)$, is the unordered pair $(c_T(l)|c_T(r))$, where l and r are the two children of u. Note that subtree-bipartitions are not defined for leaf nodes. The set of subtree-bipartitions of a tree T is denoted by $\mathcal{SBP}_T = \{\mathcal{SBP}_T(u) : u \in V_{int}(T)\}$.

Domination, containment, disjointness, and compatibility: Let $BP_i = (P_{i_1}|P_{i_2})$ and $BP_j = (P_{j_1}|P_{j_2})$ be two subtree-bipartitions. We say that BP_i is *dominated* by BP_j (and conversely that BP_j *dominates* BP_i) if either of the following two conditions holds: (1) $P_{i_1} \subseteq P_{j_1}$ and $P_{i_2} \subseteq P_{j_2}$, or (2) $P_{i_1} \subseteq P_{j_2}$ and $P_{i_2} \subseteq P_{j_1}$. We say that BP_i *contains* BP_j if $P_{j_1} \cup P_{j_2} \subseteq P_{i_1}$ or $P_{j_1} \cup P_{j_2} \subseteq P_{i_2}$, and that BP_i and BP_j are *disjoint* if $[P_{i_1} \cup P_{i_2}] \cap [P_{j_1} \cup P_{j_2}] = \emptyset$. We say that two subtree bipartitions are *compatible* if one contains the other, or they are disjoint.

The Compatibility Graph $CG(\mathcal{G})$***:*** Let $\mathcal{G}$ be a set of rooted binary gene trees on the set $\mathcal{X}$ of n taxa. The *compatibility graph* $CG(\mathcal{G})$ has one vertex for each possible subtree-bipartition defined on $\mathcal{X}$, and there is an edge between two vertices if and only if the associated subtree-bipartitions are compatible.

Note that if two subtree-bipartitions are compatible, then their associated clusters (produced by unioning the two parts of the bipartition) are also either disjoint or one contains the other.

Observation 2.1. A set $\mathcal{C}$ of $n-1$ subtree bipartitions is compatible (meaning all pairs of clusters are compatible) if and only if there exists a binary rooted tree whose set of subtree bipartitions is exactly $\mathcal{C}$.

Proof. Follows from the definition of subtree bipartition compatibility, and the fact that a set of $n-1$ compatible clusters on n taxa defines a binary tree with that set of clusters. □

We use the fact that $(n-1)$-cliques in the compatibility graph define rooted binary trees to develop solutions for the MGD_r and $MGDL_r$ problems. To do this, we define weights on nodes in the compatibility graph to characterize the solutions to these problems as $(n-1)$-cliques with maximum weight (for MGD_r or minimum weight (for $MGDL_r$). As was done by Than and Nakhleh[17] for the MDC_c problem, we will present a dynamic programming algorithm that finds an optimal $(n-1)$-clique in time that is polynomial in the number of nodes in the compatibility graph.

2.3. *Theorems*

All results here are for rooted binary gene trees and species trees. We assume that the species tree has exactly one copy of each taxon in $\mathcal{X}$, but that the gene trees can have any number (including zero) of each taxon in $\mathcal{X}$. The total number of taxa in $\mathcal{X}$ is n.

Lemma 2.1. *Let gt be a rooted binary gene tree, ST a rooted binary species tree, and u an internal node of gt. Suppose the subtree-bipartition for u is dominated by the subtree-bipartition of v in ST. Then* $\mathcal{M}(u) = v$.

Proof. Since $\mathcal{SBP}_{gt}(u)$ is dominated by $\mathcal{SBP}_{ST}(v)$, it follows that $c_{gt}(u) \subseteq c_{ST}(v)$. Let $w = \mathcal{M}(u)$. Hence, $c_{ST}(v) \cap c_{ST}(w) \neq \emptyset$, and so v and w are comparable (that is, either they are identical or one lies above the other in ST). Suppose by way of contradiction that $v \neq w$. Since $c_{gt}(u) \subseteq c_{ST}(v)$, it follows that v must lie above w. But then $c_{ST}(w)$ is a subset of the cluster of one of v's children, and so disjoint from the cluster for the other child. Hence, $\mathcal{SBP}_{gt}(u)$ is not dominated by $\mathcal{SBP}_{ST}(v)$, contradicting the initial assumption. □

The following corollary is then obvious:

Corollary 2.1. *Let gt be a rooted binary gene tree and ST a rooted binary species tree. Then every subtree-bipartition of gt is dominated by at most one subtree-bipartition in ST.*

Theorem 2.1. *Let ST be a rooted, binary species tree, gt be a rooted binary gene tree, and u an internal node in gt. Then the subtree-bipartition of u in gt is dominated by a subtree-bipartition in ST if and only if u is a speciation node.*

Proof. Suppose u is a node in gt such that its subtree-bipartition is dominated by a subtree bipartition in ST. Let l and r be the two children of u in gt. Then $\mathcal{SBP}_{gt}(u) = (c(l)|c(r))$. Let v be a node in ST such that $\mathcal{SBP}_{gt}(u)$ is dominated by $\mathcal{SBP}_{ST}(v)$. Let l' and r' be the children of v. Then, without loss of generality, $c(l) \subseteq c(l')$ and $c(r) \subseteq c(r')$. Therefore, under the MRCA mapping, l and r will be mapped to a node in the subtree rooted at l' and r', respectively. Moreover, by Lemma 2.1 $\mathcal{M}(u) = v$. Therefore, $\mathcal{M}(l) \neq \mathcal{M}(u)$, and $\mathcal{M}(r) \neq \mathcal{M}(u)$. Hence u is not a duplication node.

Next, assume that $\mathcal{SBP}_{gt}(u)$ is not dominated by any subtree-bipartition of ST, and let $\mathcal{SBP}_{ST}(\mathcal{M}(u)) = (p_1|p_2)$. Then at least one of the following holds (1) $c(l) \not\subset p_1$ and $c(l) \not\subset p_2$ or (2) $c(r) \not\subset p_1$ and $c(r) \not\subset p_2$. Without loss of generality, suppose (1) holds. Then l cannot map to a node strictly below v. However, it is also equally obvious that l cannot map to a node strictly above v, since $\mathcal{M}(u) = v$ and l is a child of u. Hence, it must be that $\mathcal{M}(l) = u$. But in this case, u is a duplication node. □

We now define some functions:

- $\mathit{dominated}(bp, ST) \in \{0, 1\}$, with $\mathit{dominated}(bp, ST) = 1$ if bp is dominated by a subtree-bipartition in $\mathcal{SBP}_{ST}$, and 0 otherwise.
- $\mathit{dom}(bp, bp') = 1$ if bp is dominated by bp' and 0 otherwise.

Corollary 2.2. *Let gt be a rooted binary gene tree and ST a rooted binary species tree. Then*

$$\mathit{Dup}(gt, ST) = |V_{int}(gt)| - \sum_{u \in V_{int}(gt)} \mathit{dominated}(\mathcal{SBP}_{gt}(u), ST).$$

Proof. Follows directly from Theorem 2.1. □

3. Algorithms for MGD_r on rooted binary gene trees

3.1. *Graph-theoretic characterization of optimal solution to MGD_r*

Let $\mathcal{G} = \{gt_1, gt_2, \ldots, gt_k\}$ be a set of rooted, binary gene trees on the set $\mathcal{X}$ of n taxa, and let n_i be the number of leaves in tree gt_i. Note that n_i does not refer to $|L(gt_i)|$, since $L(gt_i)$ is the set of taxa in $\mathcal{X}$ that appear at least once in gt_i, whereas n_i is the total number of leaves in gt_i. Since gt_i can have multiple copies of a taxon, n_i can be larger than $|L(gt_i)|$.

We construct the *compatibility graph* $CG(\mathcal{G})$ with one vertex for each possible subtree-bipartition defined on $\mathcal{X}$, as described in the previous section. We set the weight of each node v, denoted by $W_{dom}(v)$, to be the total number of subtree-bipartitions of $\mathcal{G}$ that are dominated by v. That is,

$$W_{dom}(v) = \sum_{gt \in \mathcal{G}} |\{bp : bp \in \mathcal{SBP}_{gt} \text{ and } dom(bp, v) = 1\}|.$$

We then find a clique $\mathcal{C}$ of size $n-1$ so as to maximize the weight $W_{dom}(\mathcal{C})$ of the clique $\mathcal{C}$, where $W_{dom}(\mathcal{C}) = \sum_{v \in \mathcal{C}} W_{dom}(v)$.

Theorem 3.1. *Let $\mathcal{G} = \{gt_1, gt_2, \ldots, gt_k\}$ be a set of binary, rooted gene trees on the n taxa in $\mathcal{X}$. Let $\mathcal{C}$ be an $(n-1)$-clique in $CG(\mathcal{G})$ maximizing $W_{dom}(\mathcal{C})$, and let ST be the species tree defined by the clique (so that $\mathcal{SBP}_{ST}$ corresponds to $\mathcal{C}$). Then ST is a binary species tree that optimizes MGD_r with respect to $\mathcal{G}$.*

Proof. Recall that any $(n-1)$-clique in the compatibility graph defines a rooted binary tree on $\mathcal{X}$. Let $\mathcal{C}$ be a clique of size $n-1$ and ST be the tree defined by $\mathcal{C}$. By Corollary 2.1, every subtree-bipartition in gt_i can be dominated by at most one node in $\mathcal{C}$. Therefore, each node of gt_i contributes either 1 (if the node is dominated) or 0 (if the node is not dominated) to the weight of $\mathcal{C}$. Let w_i be the amount contributed by gt_i to the weight of $\mathcal{C}$. Thus, w_i is the number of speciation nodes in gt_i with respect to the species tree corresponding to ST. Then

$$\sum_{v \in \mathcal{C}} W_{dom}(v) = \sum_{i=1}^{k} w_i = W_{dom}(\mathcal{C}).$$

Furthermore, by Corollary 2.2 and because a rooted binary tree with n_i leaves has $n_i - 1$ internal nodes, $Dup(gt_i, ST) = n_i - 1 - w_i$. Then,

$$Dup(\mathcal{G}, T) = \sum_{i=1}^{k} Dup(gt_i, ST) = \sum_{i=1}^{k} [n_i - 1 - w_i] = N - k - W_{dom}(\mathcal{C}),$$

where $\sum_{i=1}^{k} n_i = N$. Therefore, the clique with maximum weight defines a tree ST that minimizes $Dup(\mathcal{G}, ST)$.

□

3.2. *The Dynamic Programming Algorithm for MGD_r*

The graph-theoretic characterization of the optimal solution for MGD_r given in the previous section suggests an algorithm for finding the optimal solution, in which a max weight clique is sought in an exponentially large graph. However, we will show that this optimal solution can be found in time that is polynomial in the number of vertices in the graph, using dynamic programming. In addition, we will show that a constrained version of the MGD_r problem, in which the allowed subtree-bipartitions are given as input, can also be solved using the same basic dynamic programming algorithm. Finally, when the set of allowed subtree-bipartitions comes from the input set of gene trees, the result is an algorithm that runs in polynomial time.

The motivation to restrict the attention to a subset of the subtree-bipartitions comes from the observations made by Than and Nakhleh,[17] who noted that that clusters in the species tree that optimizes MDC tend to appear in at least one of the input gene trees. Therefore, we consider a constrained search problem, where instead of considering all possible subtree-bipartitions, we only consider the subtree-bipartitions of the gene trees. When we do this, instead of constructing a compatibility graph with one node for each subtree bipartition, the compatibility graph will only have nodes for the (at most) $N - k$ subtree bipartitions in the input gene trees (where $N = \sum_{i=1}^{k} n_i$). A clique of size $n - 1$ with the maximum weight will define an optimal solution to the constrained version of MGD_r where the species tree is only permitted to have subtree bipartitions from the input gene trees.

Let $\mathcal{SBP}$ be any set of subtree-bipartitions, and let $\mathcal{CLS}$ be the set of associated clusters (i.e. $\mathcal{CLS} = \{p \cup q : (p|q) \in \mathcal{SBP}\}$. We will define the constrained MGD_r problem by limiting the solution space to those rooted, binary trees, all of whose subtree-bipartitions are in the set $\mathcal{SBP}$. Thus, by setting $\mathcal{SBP}$ to be the set of all possible subtree-bipartitions we obtain the globally optimal solution, but setting $\mathcal{SBP}$ to be a proper subset of the set of all subtree-bipartitions is also possible.

By Theorem 3.1, the binary species tree with a maximum total weight (as defined by summing up the weights of its subtree bipartitions) has a minimum number of duplications, because the duplication nodes are exactly those nodes whose subtree-bipartitions are not dominated by any subtree-bipartition in the species tree.

We now show how to calculate that optimal binary species tree directly, using dynamic programming. The DP algorithm computes a rooted, binary tree T_A for every cluster $A \in \mathcal{CLS}$, such that T_A maximizes the sum, over all gene trees t, of the number of subtree-bipartitions in t that are dominated by some subtree-bipartition in T_A. We denote this total number by $value(A)$.

We preprocess the data as follows. First, we compute the set $\mathcal{CLS}$, and order its elements based on size. We also calculate $\mathcal{SBP}_{\mathcal{G}} = \bigcup_{i=1}^{k} \mathcal{SBP}_{gt_i}$, i.e. the set of all subtree bipartitions in all gene trees, and we set $count(x)$ for $x \in \mathcal{SBP}_{\mathcal{G}}$ to be the number of times x appears in any of the gene trees. Recall that for a subtree bipartition x, we define $W_{dom}(x)$ to be the number of subtree bipartitions of the gene trees that are dominated by x. We define a partial order for elements of $\mathcal{SBP}$ and $\mathcal{SBP}_{\mathcal{G}}$ based upon subtree-bipartition size. For every ordered pair $< x, y >$ such that $x \in \mathcal{SBP}_{\mathcal{G}}$ and $y \in \mathcal{SBP}$, we determine whether x is dominated by y; if y dominates x then $W_{dom}(y)$ is incremented by $count(x)$. At the end of this step, $W_{dom}(y)$ is calculated correctly for every $y \in \mathcal{SBP}$. All this preprocessing can be computed in $O(n|\mathcal{SBP}|^2)$.

We compute $value(A)$ in order, from the smallest cluster to the largest cluster $\mathcal{X}$. We set $value(A)$ as follows. For any cluster A with two taxa, we set $value(A) = W_{dom}(a_1|a_2)$, where $A = \{a_1, a_2\}$. For a cluster A with more than two taxa, we set $value(A)$ as follows:

$$value(A) = \max\{value(A_1) + value(A - A_1) + W_{dom}(A_1|A - A_1) : (A_1|A - A_1) \in \mathcal{SBP}\}$$

If there is no $(A_1|A - A_1) \in \mathcal{SBP}$, we set its $value(A)$ to $-\infty$, signifying that A cannot be further resolved. At the end of the algorithm, if $\mathcal{SBP}$ includes at least one clique of size $n - 1$, we have computed $value(\mathcal{X})$ as well as sufficient information to construct the species tree having the minimum number of duplications. If subtree bipartitions in $\mathcal{SBP}$ are not sufficient for building a fully resolved tree on $\mathcal{X}$, then $value(\mathcal{X})$ will be $-\infty$, and our algorithm returns FAIL. Note that for a specific cluster A, $value(A)$ can be computed in $O(|\mathcal{SBP}|)$ time, since at worst we need to look at every subtree-bipartition in $\mathcal{SBP}$. In other words, we have proven the following:

Theorem 3.2. *Let $\mathcal{G}$ be a set of rooted binary gene trees, $\mathcal{SBP}$ a set of subtree-bipartitions. Then, if subtree bipartitions of $\mathcal{SBP}$ define at least one binary tree on $\mathcal{X}$, then the DP algorithm finds the species tree ST minimizing the total number of duplications subject to the constraint that $\mathcal{SBP}_{ST} \subseteq \mathcal{SBP}$ in $O(n|\mathcal{SBP}|^2)$ time. Therefore, if $\mathcal{SBP}$ is all possible subtree-bipartitions, we have an exact but exponential time algorithm. However, if $\mathcal{SBP}$ contains only those subtree-bipartitions from the input gene trees, then the DP algorithm finds the optimal constrained species tree in $O(d^2n^3k^2)$ (since the number of subtree-bipartitions $|\mathcal{SBP}|$ in $\mathcal{G}$ is $O(dkn)$), where n is the number of species, k is the number of gene trees, and d the maximum number of times that any taxon appears in any gene tree.*

4. Algorithms for $MGDL_r$

4.1. *Graph-Theoretic Characterization*

We begin with some additional terminology and theorems. For any cluster A in gt and a cluster B in ST, we say that A is B-maximal if (1) $A \subseteq B$, and (2) for any cluster A' in gt, if $A \subseteq A'$, then $A' \nsubseteq B$. We define $k_B(gt)$ to be the number of B-maximal clusters within gt, and Finally, in a rooted tree T with cluster G, the unique edge e that separates G from the rest of the leaves in T is called the *parent edge* of the cluster G.

Theorem 4.1. *(From Than and Nakhleh[17] and Yu, Warnow, and Nakhleh[18]) Let gt be a rooted binary gene tree and ST a species tree on the same set of taxa. Let B be a cluster in ST and let e be the parent edge of B in ST. Then $k_B(gt)$ is equal to the number of lineages on e in an optimal reconciliation of gt within ST with respect to MDC_c. Therefore, $MDC_c(gt, ST) = \sum(k_B(gt) - 1)$, where B ranges over the clusters of ST.*

Theorem 4.2. *Let gt be a rooted binary gene tree and ST a species tree on the same set of leaves. Then $MDC_r(gt, ST) = \sum(k_B(gt) - 1)$, where B ranges over the clusters of $ST|_{L(gt)}$.*

Proof. By definition, $MDC_r(gt, ST) = MDC_c(gt, ST|_{L(gt)})$. However, gt and $ST|_{L(gt)}$ have the same set of taxa. Therefore, by Theorem 4.1, $MDC_c(gt, ST|_{L(gt)}) = \sum(k_B(gt) - 1)$, as B ranges over the clusters of $ST|_{L(gt)}$. □

Theorem 4.3. *(From Zhang[21]) Let gt be a rooted binary gene tree and ST a rooted binary species tree. Then, under the restriction-based analysis, $Duploss(gt, ST) = MDC_r(gt, ST) + 3 * Dup(gt, ST) + |V(gt)| - |V(\mathcal{R}_{ST}(L(gt)))|$.*

Let v be a vertex associated with the subtree-bipartition $(p|q)$, and let $B = p \cup q$ be the cluster associated with v. Define $W_{xl}(v, gt)$ to be 0 if $p \cap L(gt) = \emptyset$ or $q \cap L(gt) = \emptyset$, and otherwise to be $k_B(gt) - 1$. Set $W_{xl}(v) = \sum_{i=1}^{k} W_{xl}(v, gt_i)$. Then, for any species tree ST and set $\mathcal{G}$ of gene trees, $MDC_r(\mathcal{G}, ST) = \sum_{i=1}^{k} MDC_r(gt_i, ST) = \sum_{v \in \mathcal{C}} W_{xl}(v)$, where $\mathcal{C}$ is the clique in $CG(\mathcal{G})$ that corresponds to ST.

Theorem 4.4. *Let $\mathcal{G} = \{gt_1, gt_2, \ldots, gt_k\}$ be a set of binary rooted gene trees on set $\mathcal{X}$ of n species, and let $CG(\mathcal{G})$ be the compatibility graph with vertex weights defined by $W_{MGDL}(v) =$*

$W_{xl}(v) - 3W_{dom}(v)$. The set of bipartitions in an $(n-1)$-clique of minimum weight in $CG(\mathcal{G})$ defines a binary species tree ST that optimizes $MGDL_r$.

Proof. Let $\mathcal{C}$ be a clique of size $n-1$ and ST be the rooted binary tree defined by the subtree-bipartitions represented by the nodes in $\mathcal{C}$. Let $\mathcal{SBP}_{dom}(gt, ST)$ be the set of subtree-bipartitions in gt that are dominated by a subtree-bipartition in ST, i.e., $\mathcal{SBP}_{dom}(gt, ST) = \{bp : bp \in \mathcal{SBP}_{gt} \text{ and } dominated(bp, ST) = 1\}$. Note that $|\mathcal{SBP}_{dom}(gt, ST)|$ is the number of speciation nodes in gt with respect to ST. Therefore, the total number of speciation nodes in $\mathcal{G}$ is $\sum_{i=1}^{k} |\mathcal{SBP}_{dom}(gt_i, ST)| = \sum_{v \in V_{int}(ST)} W_{dom}(v)$. Let $N = \sum_{i=1}^{k} n_i$. Then,

$$\begin{aligned}
Duploss(\mathcal{G}, ST) &= \sum_{i=1}^{k} Duploss(gt_i, ST) \\
&= \sum_{i=1}^{k} [MDC_r(gt_i, ST) + 3 * Dup(gt_i, ST) - (|V(gt_i)| - |V(\mathcal{R}_{ST}(L(gt_i)))|)] \text{ (by Theorem 4.3)} \\
&= \sum_{i=1}^{k} [MDC_r(gt_i, ST) + 3 * Dup(gt_i, ST)] - \sum_{i=1}^{k} (|V(gt_i)| - |V(\mathcal{R}_{ST}(L(gt_i)))|) \\
&= \sum_{i=1}^{k} [MDC_r(gt_i, ST) + 3 * ((n_i - 1) - |\mathcal{SBP}_{dom}(gt_i, ST)|)] \\
&\quad - \sum_{i=1}^{k} (|V(gt_i)| - |V(\mathcal{R}_{ST}(L(gt_i)))|) \text{ (by Corollary 2.2)} \\
&= \sum_{v \in \mathcal{C}} W_{xl}(v) + \sum_{i=1}^{k} 3(n_i - 1) - 3 \sum_{v \in \mathcal{C}} W_{dom}(v) \\
&\quad - \sum_{i=1}^{k} (2n_i - 1) + \sum_{i=1}^{k} |V(\mathcal{R}_{ST}(L(gt_i)))| \text{ (since } |V(gt_i)| = 2n_i - 1) \\
&= \sum_{v \in \mathcal{C}} (W_{xl}(v) - 3W_{dom}(v)) + 3 \sum_{i=1}^{k} n_i \ - 3k - 2 \sum_{i=1}^{k} n_i + k + \sum_{i=1}^{k} |V(\mathcal{R}_{ST}(L(gt_i)))| \\
&= \sum_{v \in \mathcal{C}} W_{MGDL}(v) + \sum_{i=1}^{k} n_i - 2k + \sum_{i=1}^{k} |V(\mathcal{R}_{ST}(L(gt_i)))| \\
&= W_{MGDL}(\mathcal{C}) + N - 2k + \sum_{i=1}^{k} |V(\mathcal{R}_{ST}(L(gt_i)))|
\end{aligned}$$

Note that $|V(\mathcal{R}_{ST}(L(gt_i)))|$ does not depend on ST. Therefore, the clique $\mathcal{C}$ with minimum weight defines a tree ST that minimizes $Duploss(\mathcal{G}, ST)$.

□

4.2. *Dynamic Programming Approach for $MGDL_r$*

We now show how to use dynamic programming to find the optimal solution for $MGDL_r$ without having to explicitly search for the optimal clique. As we did for MGD_r, we generalize the problem to allow the user to provide a set $\mathcal{SBP}$ of subtree-bipartitions, and the solution space is restricted to those rooted, binary trees, all of whose subtree-bipartitions are in the set $\mathcal{SBP}$.

We compute $value(A)$ for all clusters A with at least two species as follows. If $|A| = 2$, we set $value(A) = W(a_1|a_2)$, where $A = \{a_1, a_2\}$. For set A with more than two taxa, we set $value(A)$ as follows:

$$value(A) = \min\{value(A_1) + value(A - A_1) + W_{xl}(A_1|A - A_1) - 3W_{dom}(A_1|A - A_1) : (A_1|A - A_1) \in \mathcal{SBP}\}.$$

The optimal number of duplications and losses is given by $value(\mathcal{X}) + N - 2k + \sum_{i=1}^{k} |V(\mathcal{R}_{ST}(L(gt_i))|$, where $N = \sum_{i=1}^{k} n_i$, and n_i is the number of leaves in gene tree gt_i. By backtracking, we can find the optimal set of compatible clusters and hence can construct the optimal tree. We now have the following theorem:

Theorem 4.5. *Let $\mathcal{G}$ be a set of k rooted binary gene trees on the set $\mathcal{X}$ of n taxa. Let $\mathcal{SBP}$ be an arbitrary set of subtree bipartitions on $\mathcal{X}$. Then the DP algorithm finds the species tree ST optimizing $MGDL_r$, subject to the constraint that $\mathcal{SBP}_{ST} \subseteq \mathcal{SBP}$, in $O(n|\mathcal{SBP}|^2)$ time. Therefore, for the case where $\mathcal{SBP}$ is the set of subtree-bipartitions from the k gene trees, the algorithm uses $O(d^2n^3k^2)$ time, where d is the maximum number of times any taxon appears in any gene tree.*

5. Acknowledgments

This research was supported by NSF (DEB 0733029 and DBI 1062335) to TW, NSERC (to SM), the Guggenheim Foundation (to TW), and the Fulbright Foundation (to MSB).

References

1. M. T. Hallett and J. Lagergren, New algorithms for the duplication-loss model, in *Proc RECOMB*, 2000.
2. W. Fitch and E. Margoliash, *Science* **155**, 279 (1967).
3. R. D. M. Page, *Syst. Biol.* **43**, 58 (1994).
4. M. Goodman, J. Czelusniak, G. Moore, E. Romero-Herrera and G. Matsuda, *Syst. Zool.* **28**, 132 (1979).
5. W. P. Maddison, *Syst. Biol.* **46**, 523 (1997).
6. L. Zhang, *J. Comp. Biol.* **4**, 177 (1997).
7. R. Chaudhary, M. S. Bansal, A. Wehe, D. Fernández-Baca and O. Eulenstein, *BMC Bioinf.* , 574 (2010).
8. A. Wehe, M. S. Bansal, J. G. Burleigh and O. Eulenstein, *Am. J. Bot.* **24**, 1540 (2008).
9. J. B. Slowinski, A. Knight and A. P. Rooney, *Mol. Phylog. Evol.* **8**, 349 (1997).
10. R. D. M. Page, *Mol. Phylog. Evol.* **14**, 89 (2000).
11. R. D. M. Page and J. A. Cotton, Vertebrate phylogenomics: reconciled trees and gene duplications, in *Proc Pacific Symposium on Biocomputing*, 2002.
12. J. Cotton and R. Page, *Tangled tales from multiple markers: reconciling conflict between phylogenies to build molecular supertrees*, in *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, ed. O. R. P. Bininda-Emonds 2004, pp. 107–125.
13. M. Sanderson and M. McMahon, *BMC Evol. Biol.* **7**, p. S3 (2007).

14. J. P. Doyon and C. Chauve, *Software Tools and Algorithms for Biological Systems (book series, Advances in Experimental Medicine and Biology)* , 287 (2011).
15. W.-C. Chang, G. J. Burleigh, D. F. Fernndez-Baca and O. Eulenstein, *BMC Bioinf.* **12**, p. S14 (2011).
16. U. Stege, Gene trees and species trees: The gene-duplication problem is fixed-parameter tractable, in *Proc International Workshop on Algorithms and Data Structures (WADS)*, 1999.
17. C. V. Than and L. Nakhleh, *PLoS Comp. Biol.* **5** (2009).
18. Y. Yu, T. Warnow and L. Nakhleh, Algorithms for MDC-based multi-locus phylogeny inference: Beyond rooted binary gene trees on single alleles, in *Proc RECOMB*, 2011.
19. J. Yang and T. Warnow, *BMC Bioinf.* **12(Suppl 9)** (2011).
20. M. S. Bayzid and T. Warnow, *J. Comp. Biol.* **19**, 591 (2012).
21. L. Zhang, *IEEE/ACM Trans. Comp. Biol. Bioinf.* **8**, 1685 (2011).
22. R. Guigo, I. Muchnik and T. Smith, *Mol. Phylog. Evol.* **6**, 189 (1996).
23. B. Ma, M. Li and L. Zhang, On reconstructing species trees from gene trees in terms of duplications and losses, in *Proc RECOMB*, 1998.
24. P. Gorecki, Reconciliation problems for duplication, loss and horizontal gene transfer, in *Proc RECOMB*, 2004.

EVALUATING VARIATIONS ON THE STAR ALGORITHM FOR RELATIVE EFFICIENCY AND SAMPLE SIZES NEEDED TO RECONSTRUCT SPECIES TREES

JAMES H. DEGNAN

Department of Mathematics and Statistics, University of Canterbury,
Christchurch, 8140, New Zealand
**E-mail: j.degnan@math.canterbury.ac.nz*
www.canterbury.ac.nz

Many methods for inferring species trees from gene trees have been developed when incongruence among gene trees is due to incomplete lineage sorting. A method called STAR (Liu et al, 2009), assigns values to nodes in gene trees based only on topological information and uses the average value of the most recent common ancestor node for each pair of taxa to construct a distance matrix which is then used for clustering taxa into a tree. This method is very efficient computationally, scaling linearly in the number of loci and quadratically in the number of taxa, and in simulations has shown to be highly accurate for moderate to large numbers of loci as well as robust to molecular clock violations and misestimation of gene trees from sequence data. The method is based on a particular choice of numbering nodes in the gene trees; however, other choices for numbering nodes in gene trees can also lead to consistent inference of the species tree. Here, expected values and variances for average pairwise distances and differences between average pairwise distances in the distance matrix constructed by the STAR algorithm are used to analytically evaluate efficiency of different numbering schemes that are variations on the original STAR numbering for small trees.

Keywords: Statistical consistency, phylogenetics, multispecies coalescent, incomplete lineage sorting, sample size

1. Introduction

Numerous methods have been developed in recent years for inferring species trees (trees describing the history of speciation events for a set of species) from gene trees (trees on which DNA sequences evolve).[1–5] Methods that explicitly model the multispecies coalescent and account for uncertainty in the gene trees due to the mutation process can be the most accurate when gene tree discordance is due to incomplete lineage sorting, but can also be computationally very slow, particularly in the number of genes. In practice researchers sometimes have difficulty with convergence of the MCMC algorithms for these methods due to the relatively large number of genes.[6] With whole genome sequencing becoming increasingly common, this problem with the methods being able to keep up with the data is likely to increase in the future and motivates the need for computationally more efficient methods that will still be powerful enough to make accurate inferences. Methods that do not explicitly model the multispecies coalescent (e.g., rooted triple consensus,[7] R*,[8] STEAC and STAR,[9,10] the quartet version of BUCKY,[11] and triplet MRP[12] can still be robust under the model and can have the advantages of performing well under model violations and being computationally efficient enough to handle genomic levels of data.

A particularly promising method in simulations has been STAR,[9] which stands for Species Tree inference using Average Ranks. The method assigns a value to each node in an input gene tree. The pairwise distance between two leaves of the tree is interpreted as twice the value of the node of their most recent common ancestor (MRCA) in the gene tree, and the pairwise distances for every pair of species is averaged over all loci. The resulting distance matrix can then be used to construct a tree using any clustering algorithm, for instance UPGMA or neighbor joining.

A key issue for the algorithm to work is how to assign the node values. The original STAR algorithm assigns a value of n to the root node, ρ, and the value of a node k is n minus the number of edges separating the node from the root. These node values are called "ranks" in Liu et al. (2009), where a higher rank means fewer edges separate the node from the root. (This usage of "rank" is slightly different from the usage of ranked trees elsewhere, where real-valued divergence times are sorted and their relative order is used to determine the rank of a node[13,14]) The node numbering used by STAR can also be interpreted as replacing all branch lengths on the gene trees with length 1 (extending external branch lengths as necessary to make trees ultrametric), and computing the average distance for each pair of species on these transformed gene trees. This numbering scheme leads to statistically consistent estimation of the species tree topology in the sense that as more independent loci (gene trees) are used, the probability that the method returns the correct species tree topology approaches 1.

Although the original numbering scheme used in STAR is statistically consistent, other numbering schemes also lead to consistent inference, as is shown in.[15] This naturally raises the question of whether other numbering schemes could be better or worse than STAR, and whether there is an optimal numbering scheme? This paper addresses this question by analytically determining expected values and variances of average distances between species in the distance matrix constructed by generalized versions of STAR for 4-taxon trees. An additional application of this approach is that sample sizes (numbers of independent loci) needed to confidently reconstruct certain inequalities in pairwise distances between taxa can be estimated.

2. Generalized STAR

To generalize STAR, let the value assigned to an internal node of a gene tree be a_j, where j is the number of edges separating the node from the root, ρ. Thus, the root node gets value a_0, the two daughter nodes of the root get value a_1 (assuming neither is a leaf), etc. There are at most $n-1$ distinct "ranks" in a gene tree, and each is only used if the gene tree is completely unbalanced (a *caterpillar* topology in which only one internal node has two leaves as its immediate descendants). Thus, a balanced four-taxon tree only uses a_0 for the root and a_1 for the two internal nodes. Thus a numbering scheme can be specified as an $(n-1)-tuple$, $(a_0, a_1, \cdots, a_{n-2})$. For the standard STAR algorithm, $a_0 = n$ and $a_i = a_{i-1} - 1$, $1 \leq i \leq n-2$. We define a *generalized STAR numbering scheme* for an n-taxon species tree to be any $(n-1)$-tuple $(a_0, \ldots, a_{n-2})$ satisfying $a_0 \geq a_1 \geq \cdots, \geq a_{n-2}$, where at least one of the inequalities is strict. The same numbering scheme is applied to each gene tree at each locus, and we assume that all gene trees have the same taxa, although these assumptions can be relaxed somewhat (see Allman et al. (2012)).

In the notation used in this paper, the STAR algorithm works by creating a distances matrix, where the (i,j)th entry is the average distance between taxa i and j, $\overline{D}_{ij}$. Letting $D_{ij}^{(\ell)}$

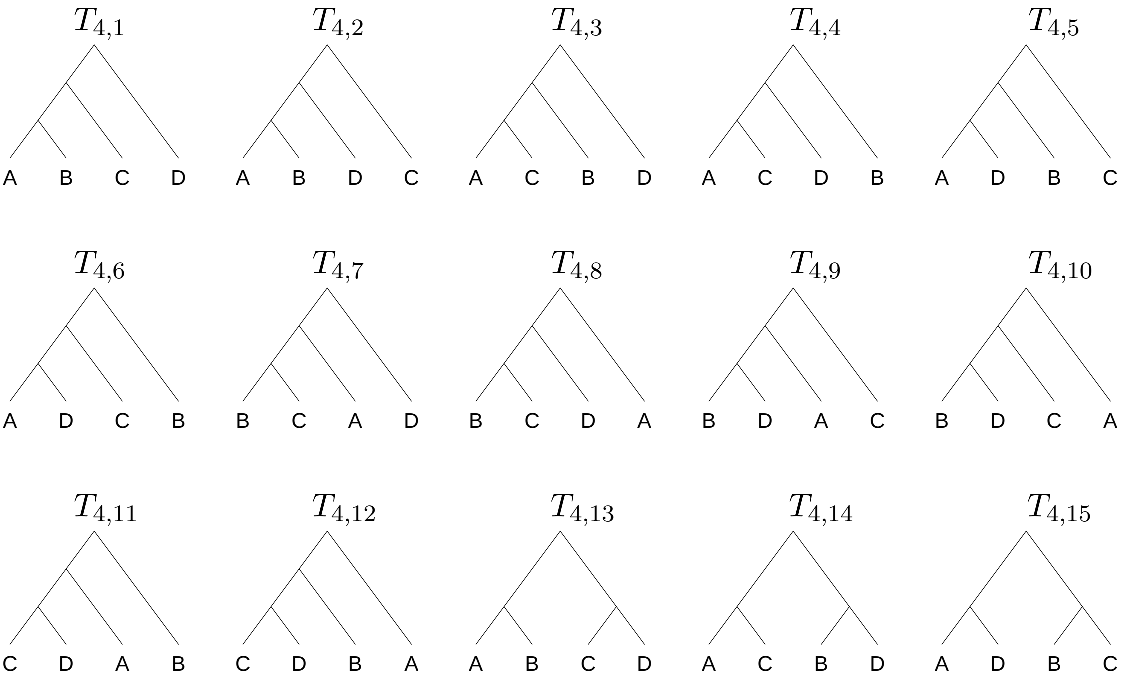


Fig. 1. Four-taxon trees used to determine expected values of the STAR distance matrix in the four-taxon case.

be the distance between taxa i and j at locus ℓ, if there are N loci, then $\overline{D}_{ij} = (1/N)\sum_{i=1}^{N} D_{ij}^{(\ell)}$.

For the 4-taxon case, the standard STAR algorithm uses $(a_0, a_1, a_2) = (4, 3, 2)$. In the standard STAR numbering scheme, all internal branches are equal in length and external branch lengths can be chosen to make the gene tree ultrametric (so that the distance from root to tip is constant). Translating the distances (adding a constant to each distance) or multiplying each by a constant factor should not affect the clustering applied to the distance matrix generated by STAR. Hence for the 4-taxon case, we can consider a generalized numbering scheme $(1, a, 0)$ and try to determine the optimal value of a, where $a = 1/2$ yields the same species tree estimate as the original STAR numbering scheme. More generally, we can consider a numbering scheme $\mathbf{a} = (a_0, \ldots, a_{n-2})$ to be equivalent to the numbering scheme $(\mathbf{a} - a_0)/(a_0 - a_{n-2})$, which fixes the smallest and largest values at 0 and 1, respectively. To determine consequences of different choices of a for $(1, a, 0)$, formulas for moments of STAR distances are shown next.

3. Expected values and variances of STAR distances

Explicit calculations of expected values, variances, and covariances of STAR distances can be used to estimate sample sizes necessary for the STAR tree to have certain relationships over others. For the 4-taxon species tree $\sigma_{4,1} = (((A, B){:}x, C){:}y, D)$, we are particularly interested in the sample size necessary for the STAR tree to have clade $\{ABC\}$ as opposed to clade $\{CD\}$. For notation, we let D_{ij} be the distance between taxa i and j on a single random gene tree occurring on the species tree. We let $\mathbb{E}[D_{ij}]$ be the expected distance between taxa i and j. Thus, as the number of loci goes to infinity STAR tree has clade $\{ABC\}$ as opposed to clade $\{CD\}$ for species tree $\sigma_{4,1}$ if $\mathbb{E}[D_{AB}] < \mathbb{E}[D_{AC}] = \mathbb{E}[D_{BC}] < \mathbb{E}[D_{CD}]$. The greatest difficulty is in being confident (having enough loci) that the last inequalities, $\mathbb{E}[D_{AC}], \mathbb{E}[D_{BC}] < \mathbb{E}[D_{CD}]$ hold.

We can determine expected values and higher moments for the random distances D_{ij} for a generalized star scheme by

$$\mathbb{E}[D_{ij}^k] = \sum_{y=1}^{(2n-3)!!} (d_{ij}(y))^k \, p_{n,y}(\lambda), \tag{1}$$

where y indexes the gene tree topology, $d_{ij}(y)$ is the observed value of the random variable D_{ij} ($d_{ij}(y)$ depends on the topology y), $p_{n,y}$ is the probability of gene tree topology y in some ordering of tree topologies for n taxa, and λ is the set of internal branch lengths on the species tree. Four-taxon tree topologies are listed and enumerated as $T_{4,y}$, $y = 1, \ldots, 15$, in Figure 1, so that $p_{4,y}$ is the probability that a gene tree has topology $T_{4,y}$. The probabilities $p_{n,y}$ can be computed symbolically using the software COAL.[16]

Additionally, we will need covariances, which can be obtained from

$$\mathbb{E}[D_{ij}D_{k\ell}] = \sum_{y=1}^{(2n-3)!!} d_{ij}(y) \, d_{k\,\ell}(y) \, p_{n,y}(\lambda) \tag{2}$$

where at least two of $\{i, j, k, \ell\}$ are distinct.

From the Central Limit Theorem, the random variables $\overline{D}_{BC}$, $\overline{D}_{CD}$, and $\overline{D}_{CD} - \overline{D}_{BC}$ converge in distribution to normal random variable as the number of loci goes to infinity. We know that $\mathbb{E}[D_{CD} - D_{BC}] > 0$, so that given enough loci, C will be likely to be clustered with B (and therefore also with A) rather than D. We therefore need the variance of $D_{CD} - D_{BC}$ to determine how many loci will be needed with a given probability for the inequality to be positive. Here we have

$$\mathbb{V}(D_{CD} - D_{BC}) = \mathbb{V}(D_{CD}) + \mathbb{V}(D_{BC}) - 2Cov(D_{CD}, D_{BC}), \tag{3}$$

where $\mathbb{V}$ and Cov are the variance and covariance, respectively. These can be evaluated using equations (1) and (2). Knowing the approximate normal distribution for $\overline{D}_{CD} - \overline{D}_{BC}$ as a function of the numbering scheme (a_0, a_1, a_2) also allows us to compare the relative efficiencies of different numbering schemes in terms of the sample size needed to have a high probability of obtaining the correct species tree estimate.

Although the Central Limit Theorem applies asymptotically, in practice, the distances $\overline{D}_{BC}$, $\overline{D}_{CD}$, $\overline{D}_{CD} - \overline{D}_{BC}$ have detectable deviations from normality with 10 loci, and are slightly left-skewed. Simulations were done with STAR to test the applicability of the Central Limit Theorem for finite samples of size 10, 50, 100, and 500 loci on the species tree $\sigma_{4,1}$. The normality of $\overline{D}_{CD} - \overline{D}_{BC}$ was tested using the Shapiro-Wilks test in R,[17] and results are listed in Table 1 for the numbering schemes (4,3,2) and (4,3,0). Statistically significant deviations are detectable with a sample size of 100 or less, but are difficult to detect with samples of size 500 loci. We note that although deviations from normality are detectable, the power to detect deviations is fairly high, since there are 1000 observations, and deviation from normality is difficult to detect by eye using histograms.

Table 1 also lists the c.o.v. (estimated from the simulations), and the proportion of estimated species trees that are correctly inferred using UPGMA implemented in Phybase[18] on the estimated distance matrix, both of which can be used as measures of the efficiency of the

Table 1. Expected values, variances, tests of normality for $\overline{D}_{CD} - \overline{D}_{BC}$ estimated from finite numbers of loci, and proportion of times the correct species tree was estimated under the STAR algorithm. The standard deviation and c.o.v. are based on the sample size, and are $\sqrt{v(a)/n}$ and $\sqrt{v(a)/n}/e(a)$, respectively. P-values are for the normality of $\overline{D}_{CD} - \overline{D}_{BC}$.

Branch lengths			$\overline{D}_{CD} - \overline{D}_{BC}$				proportion
(x, y)	(a_0, a_1, a_2)	loci	mean	sd	c.o.v.	p-value	correct
$(0.05, 0.05)$	$(4, 3, 2)$	10	0.047	0.325	6.919	0.023	0.170
$(0.05, 0.05)$	$(4, 3, 2)$	50	0.056	0.140	2.337	0.076	0.253
$(0.05, 0.05)$	$(4, 3, 2)$	100	0.061	0.098	1.619	0.190	0.363
$(0.05, 0.05)$	$(4, 3, 2)$	500	0.063	0.046	0.718	0.868	0.793
$(0.05, 0.05)$	$(4, 3, 0)$	10	0.107	0.570	5.350	0.000	0.145
$(0.05, 0.05)$	$(4, 3, 0)$	50	0.118	0.246	2.093	0.349	0.275
$(0.05, 0.05)$	$(4, 3, 0)$	100	0.120	0.173	1.438	0.555	0.394
$(0.05, 0.05)$	$(4, 3, 0)$	500	0.122	0.079	0.646	0.225	0.849
$(1.00, 0.05)$	(4,3,2)	10	0.052	0.273	5.234	0.000	0.452
$(1.00, 0.05)$	(4,3,2)	50	0.055	0.122	2.204	0.004	0.535
$(1.00, 0.05)$	(4,3,2)	100	0.053	0.088	1.651	0.069	0.619
$(1.00, 0.05)$	(4,3,2)	500	0.055	0.034	0.707	0.604	0.894
$(1.00, 0.05)$	(4,3,0)	10	0.076	0.380	5.022	0.000	0.452
$(1.00, 0.05)$	(4,3,0)	50	0.075	0.176	2.273	0.070	0.551
$(1.00, 0.05)$	(4,3,0)	100	0.077	0.125	1.617	0.137	0.652
$(1.00, 0.05)$	(4,3,0)	500	0.079	0.056	0.708	0.340	0.905

two numbering schemes. For the species tree with branches $(x, y) = (0.05, 0.05)$, for each given number of loci, the scheme $(4, 3, 2)$ has a higher c.o.v. than $(4, 3, 0)$, although proportions of correctly inferred trees are only statistically significantly better for $(4, 3, 0)$ when sample sizes reach 500 loci. Note, however, that both in simulation (Table 1) and based on theoretical sample size calculations in Table 2, $(4, 3, 2)$ and $(4, 3, 0)$ are approximately equally good for $(x, y) = (1.0, 0.05)$. We note that $(x, y) = (0.05, 1.0)$ leads to more gene tree discordance than $(1.0, 0.05)$

4. Evaluation of variations on STAR

4.1. *The 4-taxon case*

To evaluate generalized STAR in the 4-taxon case, we let the numbering scheme be $(1, a, 0)$. To find an optimal value of a, set $e(a) := \mathbb{E}_a[D_{CD} - D_{BC}]$ and $v(a) = \mathbb{V}_a[D_{CD} - D_{BC}]$, i.e., taking means and variances parameterized by a. Using the normal approximation, the probability that

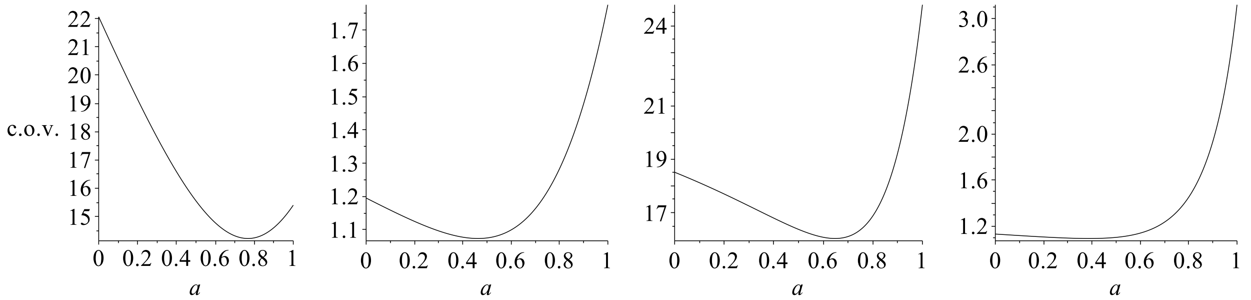


Fig. 2. Coefficient of variation for $D_{CD} - D_{BC}$ as a function of a using the STAR numbering scheme $(1, a, 0)$ for species tree $\sigma_{4,1}$ with $(x, y) = (0.05, 0.05), (0.05, 1.0), (1.0, 0.05), (1.0, 1.0)$.

$\overline{D}_{CD} - \overline{D}_{BC}$ is greater than 0 is approximately $\mathbb{P}_a[Z < (0 - e(a))/\sqrt{v(a)/n}] = \Phi(\sqrt{n}e(a)/\sqrt{v(a)})$, where Z is a standard normal random variable and Φ is the standard normal cumulative distribution function. Thus the sample size, N, needed to have confidence $1 - \alpha$ that $\mathbb{E}[D_{CD} - D_{BC}] > 0$ is approximately

$$N = \lceil (\Phi^{-1}(1 - \alpha)\text{c.o.v.}(a))^2 \rceil \tag{4}$$

where $\text{c.o.v.}(a) = \sqrt{v(a)}/e(a)$ is the coefficient of variation. We consider the optimal value of a is the value that minimizes N in equation (4), or equivalently, that minimizes the coefficient of variation, $\sqrt{v(a)}/e(a)$. For species tree $(((A, B){:}x, C){:}y, D)$, the coefficient of variation under the scheme $(1, a, 0)$ can be written analytically using

$$\begin{aligned}
v(a) = \Big(&- \mathrm{e}^{-2x} - 9\mathrm{e}^{-2y} - 7\,\mathrm{e}^{-x-3y} + 6\,\mathrm{e}^{-4y-x} + 15\,\mathrm{e}^{-y} + 2\,\mathrm{e}^{-2x-3y} + 3\,\mathrm{e}^{-x} \\
&- 3\,\mathrm{e}^{-x-y} - \mathrm{e}^{-2x-6y}\Big) a^2/9 + \Big(- 30\,\mathrm{e}^{-y} + 3\,\mathrm{e}^{-x-2y} - 1\,\mathrm{e}^{-2x-3y} + 18\,\mathrm{e}^{-2y} \\
&+ 10\,\mathrm{e}^{-x-3y} - 9\mathrm{e}^{-4y-x} + \mathrm{e}^{-2x-6y} + \mathrm{e}^{-2x-y} - \mathrm{e}^{-2x-4y}\Big) a/9 \\
&- 1/3\,\mathrm{e}^{-x-2y} + 1/3\,\mathrm{e}^{-4y-x} + 1/18\,\mathrm{e}^{-2x-4y} - \mathrm{e}^{-2y} - 1/36\,\mathrm{e}^{-2x-2y} - 1/36\,\mathrm{e}^{-2x-6y} \\
&+ 5/3\,\mathrm{e}^{-y} - 5/18\,\mathrm{e}^{-x-3y} + 1/6\,\mathrm{e}^{-x-y} \\
e(a) = \big(&1/3\,\mathrm{e}^{-x} - 1 + \mathrm{e}^{-y} - 1/3\,\mathrm{e}^{-x-3y}\big)\, a + 1 - \mathrm{e}^{-y} - 1/6\,\mathrm{e}^{-x-y} + 1/6\,\mathrm{e}^{-x-3y}
\end{aligned}$$

where these values were computed symbolically using equations (1)-(3), using COAL for the gene tree probabilities $p_{n,i}(\lambda)$, and simplifying in the software MAPLE.

The optimal value of a is difficult to find analytically as a function of x and y; however, for fixed x and y, one can equivalently find the optimal value of $v(a)/e^2(a)$, which is a rational function with both numerator and denominator being quadratic functions in a, and the minimum of this function can be found analytically. For $(x, y) = (0.05, 0.05)$, for example, the optimal value is $a \approx 0.767$. This value is close to $a = 3/4$, which is equivalent to the numbering scheme $(4, 3, 0)$. The coefficient of variation as a function of a is shown in Figure 2 for a few choices of (x, y) and for species trees $\sigma_{4,1}$.

We compute sample sizes required to get a 95% chance that a random sample of N loci results in $D_{CD} - D_{BC} > 0$ for two choices of (x, y) in Table 2. In the table, the root is difficult to resolve, and for $x = 1.0$, the fact that A and B form a clade is less to difficult to infer. We note that for $(x, y) = (0.05, 0.05)$, the numbering scheme $(4, 3, 1)$ does best among those listed, while for $(x, y) = (0.05, 1.0)$, the numbering scheme $(4, 3, 0)$ does best amongst the same set of numbering schemes.

We note that choosing a to maximize the probability that $D_{CD} - D_{BC} > 0$ does not necessarily maximize the probability that the STAR tree matches the species tree. In particular, for (x, y), if x is small and y is large, then $D_{CD} - D_{BC} > 0$ with high probability, and the more difficult relationships to resolve will be those between taxa A, B, and C. In this case, it might make sense to find a that maximizes the probability that $D_{BC} - D_{AB} > 0$, and sample sizes sufficient for $\overline{D}_{CD} - \overline{D}_{BC} > 0$ are unlikely to be sufficient for $\overline{D}_{BC} - \overline{D}_{AB} > 0$ to obtain.

The sample sizes here are only for being 95% confident that $\overline{D}_{CD} - \overline{D}_{BC} > 0$, which does not guarantee that the correct species tree will be estimated, although in practice, this is often the case. For the scheme $(4, 3, 0)$, a sample size of 548 is needed for 95% confidence that $\overline{D}_{CD} - \overline{D}_{BC} > 0$ when $(x, y) = (0.05, 0.05)$. In simulation, a sample size of 500 recovered the species tree only 84.9% of the time, although by formula (4), a sample size of 500 should have a 94% $(= \Phi(1.571))$ that $\overline{D}_{CD} - \overline{D}_{BC} > 0$. It is not surprising that sample sizes needed to recover the entire tree are somewhat larger than what is needed to estimate the inequality, as for example, $\overline{D}_{CD} - \overline{D}_{BC} > 0$ does not guarantee that $\overline{D}_{AB}$ is the smallest estimated distance, although this is necessary to correctly estimate the species tree.

An alternative approach to guaranteeing that a particularly difficult inequality is estimated correctly with high probability is to guarantee that all pairwise inequalities are estimated correctly. Given the lack of independence between pairwise distances, this is difficult to do exactly. However, using Bonferroni's inequality, k events (not necessarily independent or equiprobable), that each have probability at least $1 - \varepsilon/k$, all occur with probability at least $1 - \varepsilon$.[19] Thus, one could choose, for example, the sample size needed to correctly determine $D_{CD} - D_{BC} > 0$ with probability $1 - \alpha = 0.99$, and conclude that all $\binom{4}{2} = 6$ pairwise relationships (and therefore the correct tree) will be inferred with probability at least $1 - 6\alpha = 0.94$. In general, this approach will be quite conservative (i.e., will overestimate the number of loci needed) if it is based on the most difficult pairwise inequality. Sample sizes needed for 99% confidence can be obtained from 95% values by multiplying by $[\Phi^{-1}(1 - 0.99)/\Phi^{-1}(1 - 0.95)]^2 = (2.326/1.645)^2 \approx 2.00$. Thus, this approach suggests that samples sizes being doubled (for the 4-taxon case) would give approximately at least as much confidence that the entire tree was estimated correctly as well as the inequality $D_{CD} - D_{BC} > 0$.

From the 4-taxon examples, the branch lengths $(x, y) = (0.05, 0.05)$ are in the *anomaly zone*, in which the most likely gene tree topology is $((AB)(CD))$ rather than $(((AB)C)D)$.[20] However, $(x, y) = (1.0, 0.05)$ is not in the anomaly zone (i.e., the most likely gene tree topology matches the species tree topology) but requires similarly large samples (hundreds of loci) to recover the species tree with high probability (Table 1). The results are similar to other studies that have shown that hundreds of loci might be needed to accurately reconstruct the species tree from gene tree topologies when gene tree discordance is this high.[9,21]

Table 2. Samples sizes and c.o.v. needed for approximate 95% confidence that $\overline{D}_{CD} - \overline{D}_{BC} > 0$. The c.o.v. is based on $\sqrt{v(a)}/e(a)$ for a single locus.

(x, y)	(a_0, a_1, a_2)	$(1, a, 0)$	number of loci needed	c.o.v.
$(0.05, 0.05)$	$(4, 3, 2)$	$(1, 0.5, 0)$	655	15.553
$(0.05, 0.05)$	$(4, 3, 1)$	$(1, 0.67, 0)$	564	14.428
$(0.05, 0.05)$	$(4, 3, 0)$	$(1, 0.75, 0)$	548	14.230
$(0.05, 0.05)$	$(4, 3.5, 0)$	$(1, 0.875, 0)$	567	14.474
$(0.05, 0.05)$	$(4, 2, 1)$	$(1, 0.33, 0)$	817	17.375
$(1.00, 0.05)$	$(4, 3, 2)$	$(1, 0.5, 0)$	726	16.371
$(1.00, 0.05)$	$(4, 3, 1)$	$(1, 0.67, 0)$	697	16.038
$(1.00, 0.05)$	$(4, 3, 0)$	$(1, 0.75, 0)$	725	16.358
$(1.00, 0.05)$	$(4, 3.5, 0)$	$(1, 0.875, 0)$	919	18.428
$(1.00, 0.05)$	$(4, 2, 1)$	$(1, 0.33, 0)$	791	17.097

4.2. *A 5-taxon example*

Another example of using different numbering schemes to distinguish difficult-to-resolve relationships is for the two species trees $\sigma_{5,1} = (((A, B){:}x, C){:}y, (D, E){:}z)$ and $\sigma_{5,2} = ((A, B){:}u, (C, (D, E){:}v){:}w)$. For $\sigma_{5,1}$, if x and y are small while z is relatively large, the most likely gene tree could have the same topology as $\sigma_{5,2}$. Similarly, if v and w are small, while u is relatively large, a gene tree with the same topology as $\sigma_{5,1}$ could be the most likely gene tree when $\sigma_{5,2}$ is the species tree. This example with these two candidate species trees is actually the smallest example of a "wicked forest", in which for each of two or more candidate species trees, the most likely gene tree topology matches a different species tree.[20,22] In this example, the clades $\{AB\}$ and $\{DE\}$ might not be very difficult to estimate, and the greatest difficulty is in deciding on which side of the root taxon C belongs. We note that this example was also one of the more difficult cases for estimating rooted species trees from unrooted gene trees.[23]

To get a sense of sample sizes that might be needed to correctly place taxon C, and to find an optimal numbering scheme (a_0, a_1, a_2, a_3) to use with STAR, we consider $D_{CD} - D_{BC}$. Here we map the smallest and largest values of the numbering scheme to 0 and 1, respectively, and consider schemes $(1, a_1, a_2, 0)$ with $1 > a_1 > a_2 > 0$. A plot of the coefficient of variation is given in Figure 3 for the species tree $(((A, B){:}x, C){:}y, (D, E){:}z)$ with $(x, y, z) = (0.05, 0.05, 1.0)$, which shows that larger values of a_1 tend to be more efficient, although some efficiency is lost with value of a_1 too close to 1, and that the choice of a_1 is more important than the choice of a_2.

Sample size calculations can be done as in the 4-taxon case, using $\mathbf{a} = (a_0, a_1, a_2, a_3)$ in place of a in equation (4). Here, a near optimal choice for $\mathbf{a}$ is $(1.0, 0.88, 0.50, 0.0)$. This is equivalent to $(5.00, 4.64, 3.5, 2.00)$ when the smallest and largest values are fixed at 2.0 and 5.0. Similarly, the standard STAR numbering scheme of $(5, 4, 3, 2)$ is equivalent to $(1, 2/3, 1/3, 0)$. Estimated expected values, standard deviations, c.o.v. (both estimated and theoretical), and proportion

Table 3. Expected values standard deviation, c.o.v., and for $\mathbb{E}[D_{CD} - D_{BC}]$ estimated from finite numbers of loci, and proportion of times the correct species tree was estimated under the STAR algorithm using species tree $(((A, B){:}x, C){:}y, (D, E){:}z)$. The theoretical c.o.v. is $\sqrt{\mathbb{V}([D_{CD} - D_{BC}]/n)}/\mathbb{E}[D_{CD} - D_{BC}]$.

Branch lengths (x, y, z)	numbering scheme	loci	mean	sd	c.o.v. (theoretical)	proportion correct
$(0.05, 0.05, 1.0)$	(5,4,3,2)	10	0.0691	0.336	4.861 (4.841)	0.144
$(0.05, 0.05, 1.0)$	(5,4,3,2)	50	0.071	0.150	2.105 (2.165)	0.255
$(0.05, 0.05, 1.0)$	(5,4,3,2)	100	0.073	0.104	1.434 (1.531)	0.375
$(0.05, 0.05, 1.0)$	(5,4,3,2)	500	0.068	0.046	0.670 (0.684)	0.800
$(0.05, 0.05, 1.0)$	(5,4.64,3.5,2)	10	0.062	0.266	4.308 (4.439)	0.152
$(0.05, 0.05, 1.0)$	(5,4.64,3.5,2)	50	0.054	0.106	1.964 (1.985)	0.273
$(0.05, 0.05, 1.0)$	(5,4.64,3.5,2)	100	0.058	0.080	1.376 (1.403)	0.405
$(0.05, 0.05, 1.0)$	(5,4.64,3.5,2)	500	0.055	0.034	0.611 (0.628)	0.865

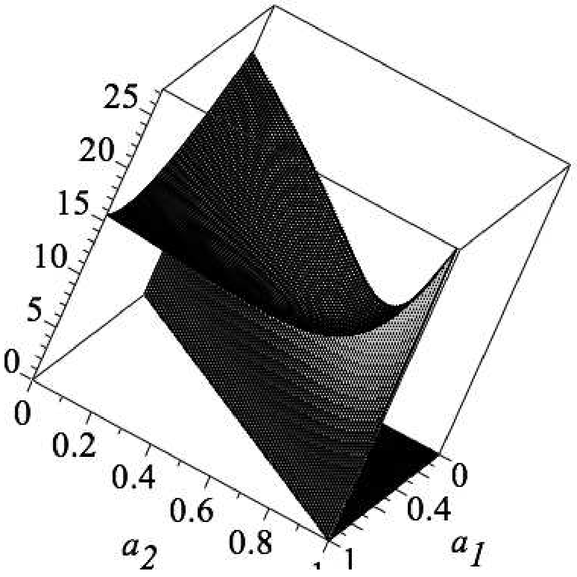


Fig. 3. C.o.v. as a function of a_1 and a_2 for the numbering scheme $(1, a_1, a_2, 0)$ for the species tree $(((A, B){:}x, C){:}y, (D, E){:}z)$ with $(x, y, z) = (0.05, 0.05, 1.0)$. The drop along the plane $a_1 = a_2$ occurs because of the assumption that $a_1 > a_2$.

of STAR trees matching the species tree are shown in Table 3. The sample size needed to determine $D_{CD} - D_{BC} > 0$ with 95% confidence is roughly $N = 534$ with $(a_0, a_1, a_2, a_3) = (5.00, 4.64, 3.50, 2.00)$ and $N = 634$ with $(a_0, a_1, a_2, a_3) = (5, 4, 3, 2)$.

5. Discussion

This paper has shown a framework for investigating variations on the STAR numbering scheme for the purpose of evaluating the relative efficiency of different schemes. The original STAR numbering scheme is well-chosen in that it is simple and works well in a wide variety of situations – i.e., for both long and short branches in the species trees investigated in this paper, the original STAR numbering of equally spaced branches often had a relatively low coefficient of variation, and optimal values for given species tree branch lengths are not necessarily optimal for other branch lengths. Overall, there is no numbering scheme that is uniformly optimal — that performs better than any other scheme for all species tree branch lengths.

If there is some knowledge of the species tree topology, in particular nodes that might be especially difficult to resolve, alternatives to the original STAR numbering scheme can perform better in some situations. In particular, if a node in the species tree is not very resolved, then making genes more star-like in the sense of making internal nodes closer to the root than under the standard STAR algorithm, can lead to improvements in estimating species trees in terms of the number of loci needed. For a fixed number of loci, this could result in improved bootstrap support for the problematic nodes. The sample size calculations used in this paper assume approximately normal distributions for the distances between taxa averaged over many loci. The normality assumption is more reasonable with large numbers of loci; thus, for branch lengths for which equation (4) returns a small number of loci, the normality assumption is less plausible. Instead, equation (4) is intended for use with difficult species trees for which large sample sizes might be required, making the normality assumption more reasonable.

In this paper, only known gene trees have been used, although in practice gene trees are estimated with some error. Because topologies can typically be estimated more reliably than branch lengths, however, STAR and its variations should be less sensitive to misestimation of gene trees than methods that use branch lengths.[9] Although the effects of misestimation on species tree inference can be simulated directly, we note that theoretical expected values, variances, and covariances, and therefore sample size calculations do not assume that gene tree probabilities are obtained directly from the multispecies coalescent. Instead, the probabilities $p_{n,i}$ used in equations (1) and (2) can come from any model for the gene tree topologies, including a model that includes error in the gene trees. In particular, if a distribution on estimated gene trees is obtained, say $\{\widehat{p_i}\}$, then this distribution can be used in equations (1) and (2), and the relative efficiency of different numbering schemes can be compared on different distributions of estimated trees. Similarly, effects of other processes, such as horizontal gene transfer,[24] gene duplication,[25,26] and hybridization[27,28] can be studied as long as distributions of gene tree topologies can be obtained (either theoretically or estimated through simulations).

Some unanswered questions raised by this study is whether the original STAR numbering scheme performs best “on average”, perhaps averaged over species trees generated on a Yule model, and whether one STAR numbering scheme can dominate another — that is, could one STAR numbering scheme always perform better than another for all possible topologies and branch lengths in the species tree? The framework used in this paper of using expected pairwise distances as well as their variances and covariances could be used to investigate these questions further.

Acknowledgments

This work was supported by the New Zealand Marsden Fund. The author is grateful for comments from the anonymous reviewers.

References

1. B. Rannala and Z. Yang, *Annu. Rev. Genom. Human Genet.* **9**, 217 (2008).
2. J. H. Degnan and N. A. Rosenberg, *Trends Ecol. Evol.* **24**, 332 (2009).
3. S. V. Edwards, *Evolution* **63**, 1 (2009).
4. L. Liu, L. Yu, L. S. Kubatko, D. K. Pearl and S. V. Edwards, *Mol. Phylogenet. Evol.* **53**, 320 (2009).
5. L. L. Knowles and L. S. Kubatko, *Estimating species trees: practical and theoretical aspects* (Wiley-Blackwell, Hoboken, NJ, 2010).
6. K. A. Cranston, B. Hurwitz, D. Ware, L. Stein and R. A. Wing, *Syst. Biol.* **58**, 489 (2009).
7. G. B. Ewing, I. Ebersberger, H. A. Schmidt and A. von Haeseler, *BMC Evol. Biol.* **8**, p. 118 (2008).
8. J. H. Degnan, M. DeGiorgio, D. Bryant and N. A. Rosenberg, *Syst. Biol.* **58**, 35 (2009).
9. L. Liu, L. Yu, D. K. Pearl and S. V. Edwards, *Syst. Biol.* **58**, 468 (2009).
10. L. Liu and L. Yu, *Syst. Biol.* **60**, 661 (2011).
11. B. R. Larget, S. K. Kotha, C. N. Dewey and C. Ané, *Bioinformatics* **26**, 2910 (2010).
12. Y. Wang and J. H. Degnan, *Stat. Appl. Genet. Mol.* **10**, p. 21 (2011).
13. T. Gernhard, D. Ford, R. Vos and M. Steel, *Evolutionary Bioinformatics Online* **2**, 285 (2006).
14. J. H. Degnan, N. Rosenberg and T. Stadler, *Math. Biosci.* **235**, 45 (2012).
15. E. S. Allman, J. H. Degnan and J. A. Rhodes, *http://www.arxiv/abs/1204.4413* , 23 (2012).
16. J. H. Degnan and L. A. Salter, *Evolution* **59**, 24 (2005).
17. R Development Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, (2012). ISBN 3-900051-07-0.
18. L. Liu and L. Yu, *Bioinformatics* **26**, 962 (2010).
19. S. Ross, *A First Course in Probability*, 5th edn. (Prentice-Hall, Upper Saddle River, NJ, 1998).
20. J. H. Degnan and N. A. Rosenberg, *PLoS Genet.* **2**, 762 (2006).
21. Y. Wu, *Evolution* **66**, 763 (2012).
22. N. A. Rosenberg and R. Tao, *Syst. Biol.* **57**, 131 (2008).
23. E. S. Allman, J. H. Degnan and J. A. Rhodes, *J. Math. Biol.* **62**, 833 (2011).
24. Y. Chung and C. Ané, *Syst. Biol.* **60**, 261 (2011).
25. O. Åkerbord, B. Sennblad, L. Arvestad and J. Lagergren, *Proc. Natl. Acad. Sci. USA* **106**, 5714 (2009).
26. M. Rasmussen and M. Kellis, *Genome Res.* **22**, 755 (2012).
27. C. Meng and L. S. Kubatko, *Theor. Popul. Biol.* **75**, 35 (2009).
28. Y. Yu, J. H. Degnan and L. Nakhleh, *PLoS Genet.* **8**, p. e1002660 (2012).

THE BEHAVIOR OF ADMIXED POPULATIONS IN NEIGHBOR-JOINING INFERENCE OF POPULATION TREES

NAAMA M. KOPELMAN and LEWI STONE

Porter School of Environmental Studies, Department of Zoology, Tel Aviv University, Ramat Aviv, Israel

OLIVIER GASCUEL

Méthodes et Algorithmes pour la Bioinformatique, LIRMM-CNRS, Montpellier, France

NOAH A. ROSENBERG*

Department of Biology, Stanford University, Stanford, California, USA
**E-mail: noahr@stanford.edu*

Neighbor-joining is one of the most widely used methods for constructing evolutionary trees. This approach from phylogenetics is often employed in population genetics, where distance matrices obtained from allele frequencies are used to produce a representation of population relationships in the form of a tree. In phylogenetics, the utility of neighbor-joining derives partly from a result that for a class of distance matrices including those that are *additive* or tree-like—generated by summing weights over the edges connecting pairs of taxa in a tree to obtain pairwise distances—application of neighbor-joining recovers exactly the underlying tree. For populations within a species, however, migration and admixture can produce distance matrices that reflect more complex processes than those obtained from the bifurcating trees typical in the multispecies context. Admixed populations—populations descended from recent mixture of groups that have long been separated—have been observed to be located centrally in inferred neighbor-joining trees, with short external branches incident to the path connecting their source populations. Here, using a simple model, we explore mathematically the behavior of an admixed population under neighbor-joining. We show that with an additive distance matrix, a population admixed among two source populations necessarily lies on the path between the sources. Relaxing the additivity requirement, we examine the smallest nontrivial case—four populations, one of which is admixed between two of the other three—showing that the two source populations never merge with each other before one of them merges with the admixed population. Furthermore, the distance on the constructed tree between the admixed population and either source population is always smaller than the distance between the source populations, and the external branch for the admixed population is always incident to the path connecting the sources. We define three properties that hold for four taxa and that we hypothesize are satisfied under more general conditions: *antecedence of clustering*, *intermediacy of distances*, and *intermediacy of path lengths*. Our findings can inform interpretations of neighbor-joining trees with admixed groups, and they provide an explanation for patterns observed in trees of human populations.

Keywords: admixture; neighbor-joining; phylogenetics; population genetics

1. Introduction

Distance matrix methods in phylogenetics construct trees of taxa using algorithms applied to matrices that tabulate pairwise evolutionary distances between the taxa.[1,2] Among these methods, neighbor-joining[3,4] is one of the most popular.[5–7] One of its key features is its consistency: if the distance matrix is *additive*, such that a tree of taxa exists that generates the distances in the matrix, then neighbor-joining recovers this exact tree.[5,8,9] Further, neighbor-

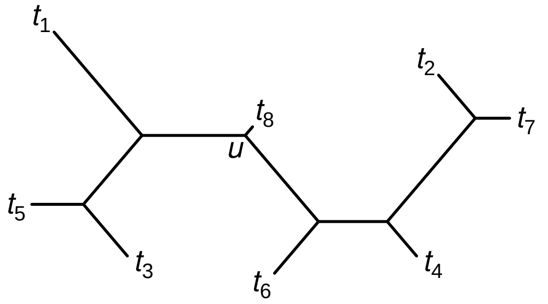


Fig. 1. Properties observed for admixed taxa in neighbor-joining trees. Taxon t_8 represents an admixture of source populations t_1 and t_2. The admixed taxon appears on a short external branch incident to the path connecting the source populations. Denoting distances on the tree by $\hat{d}$ and topological path lengths that count edges separating pairs of taxa by $\hat{b}$, the tree illustrates the properties of *intermediacy of distances* ($\hat{d}_{t_1,t_8} < \hat{d}_{t_1,t_2}$ and $\hat{d}_{t_2,t_8} < \hat{d}_{t_1,t_2}$, or equivalently, $\hat{d}_{u,t_8} < \hat{d}_{u,t_1}$ and $\hat{d}_{u,t_8} < \hat{d}_{u,t_2}$, where u is the unique node that places t_1, t_2, and t_8 in different subtrees), and *intermediacy of path lengths* ($\hat{b}_{t_1,t_8} \leq \hat{b}_{t_1,t_2}$ and $\hat{b}_{t_2,t_8} \leq \hat{b}_{t_1,t_2}$).

joining is robust in that theoretical and simulation-based studies have found it to infer sensible trees under a broad range of mathematical and biological conditions.[7,9–13]

As trees have long been used in population genetics to describe relationships among populations,[14,15] the neighbor-joining algorithm has been applied extensively as a population clustering tool, using distance matrices calculated from population-level allele frequencies. In humans, neighbor-joining trees have been and continue to be a regular feature of studies of population relationships.[16–20] In population-genetic studies, because migration and admixture sometimes generate evolutionary histories that cannot easily be described by a bifurcating tree of populations, a neighbor-joining tree is treated as a type of population clustering diagram rather than a precise representation of the evolutionary history of the populations.

When neighbor-joining has been used with admixed populations—populations recently descended from two or more source groups that have long been separated—particular characteristics of the inferred trees have often been observed (Fig. 1). For example, one simulation study based on human data identified a reduction in the external branch length leading to an admixed population as the strength of gene flow with other populations was increased.[21] It has also been suggested on the basis of observed human population trees that a short external branch for a population on a constructed neighbor-joining tree can imply recent admixture of the population, and that admixed populations often appear in the "middle" of a neighbor-joining tree, on branches incident to paths connecting possible source populations.[21–25] This pattern is evident in Fig. 2, in which admixed Mestizo populations from Latin America lie on branches incident to the path connecting Native American and European populations. Here, we seek to understand these results on the behavior of admixed populations in the application of the neighbor-joining algorithm. We therefore apply neighbor-joining to populations that satisfy a simple admixture model, first considering the case in which the distance matrix is additive. Next, for the case of $n = 4$ taxa, we use a mechanistic mathematical investigation to examine three specific properties of neighbor-joining trees involving an admixed population.

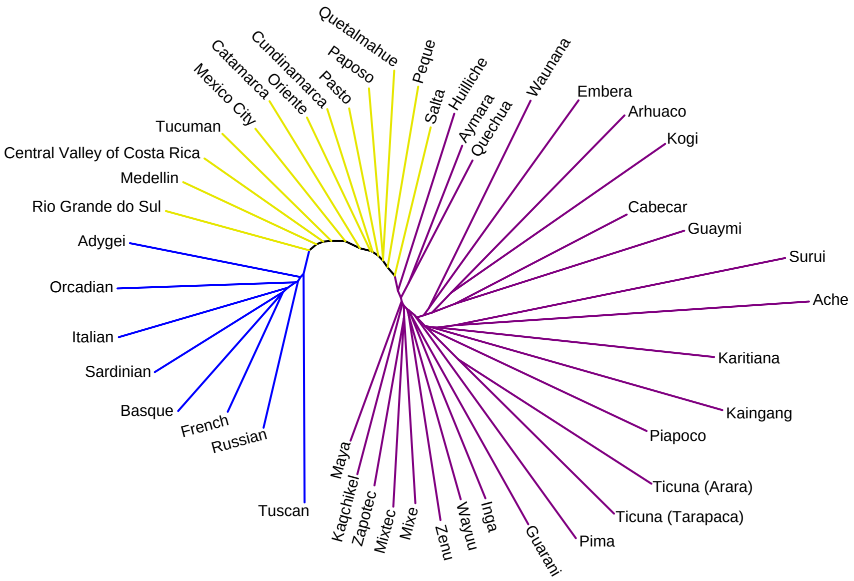


Fig. 2. Neighbor-joining tree of admixed Mestizo populations together with Native American and European populations that represent ancestral source regions for the admixed populations. The tree, obtained using `Neighbor` and `Drawtree` in the `Phylip` package,[26] uses data on 678 microsatellite loci in 13 Mestizo, 26 Native American, and 8 European populations.[27–29] Allele frequencies were computed from 872 individuals—249 Mestizo, 463 Native American, and 160 European—and distances were computed with `Microsat`[30] using one minus the proportion of shared alleles.[31] Mestizo, Native American, and European branches appear in yellow, purple, and blue, respectively. Mestizos lie in the "middle" of the tree, connecting to the path that links the Native Americans and Europeans. External branches for Mestizo populations are shorter on average (0.102) than for Native American (0.146) and European populations (0.109); Mestizo populations have 9 of the 15 shortest external branches.

2. The neighbor-joining algorithm

We briefly review the neighbor-joining algorithm.[3,4] Consider a set of n taxa, together with a distance function d computed for each pair of taxa, such that the distance between taxa i and j is denoted d_{ij}. The algorithm takes as input the distance matrix D containing entries d_{ij}, with i and j ranging from 1 to n, and it outputs a bifurcating unrooted tree. D is symmetric ($d_{ij} = d_{ji}$), with zeroes on the diagonals ($d_{ii} = 0$) and nonnegative real entries ($d_{ij} \geq 0$).

As in other agglomerative algorithms that construct bifurcating trees,[2,32] at each of a series of steps, the two nearest taxa according to a selection criterion are connected to a new interior node, becoming "neighbors" on the constructed tree. Branch lengths from the new node to the nodes it agglomerates, as well as the distances to all remaining nodes, are then calculated, and a new distance matrix is obtained. This procedure is repeated iteratively until the last three nodes remain, and these three nodes are then connected to a final interior node. Because the last three nodes are always joined, the number of taxa must exceed three for neighbor-joining to have a nontrivial decision at the first step.

At each step, the key decision is the choice of the pair of taxa that are agglomerated. Neighbor-joining uses an $n \times n$ matrix Q, containing entries q_{ij} for pairs of taxa (i, j):

$$q_{ij} = (n-2)d_{ij} - \sum_{k=1}^{n} d_{ik} - \sum_{k=1}^{n} d_{jk}. \qquad (1)$$

The two taxa that are agglomerated are those with the minimal value of q_{ij} (choosing randomly in case of ties). If taxa i and j are agglomerated, then their distances to the new node u become

$$d_{iu} = \frac{1}{2} d_{ij} + \frac{1}{2(n-2)} \left(\sum_{k=1}^{n} d_{ik} - \sum_{k=1}^{n} d_{jk} \right) \tag{2}$$

$$d_{ju} = \frac{1}{2} d_{ij} + \frac{1}{2(n-2)} \left(\sum_{k=1}^{n} d_{jk} - \sum_{k=1}^{n} d_{ik} \right). \tag{3}$$

The distances of all remaining nodes k to node u are computed as

$$d_{ku} = (d_{ik} + d_{jk} - d_{ij})/2. \tag{4}$$

The next agglomeration then proceeds from an $(n-1) \times (n-1)$ distance matrix that replaces distances involving nodes i and j with those involving the single node u.

3. An admixture scenario

We examine a scenario in which one of the taxa is admixed among two of the others. This taxon can be viewed as having been formed from its two source taxa, such that individual members of the taxon have ancestors in both source groups. We label the taxa $t_1, t_2, \ldots, t_n$. Without loss of generality, let taxon t_n be the admixed group, and suppose that it is an admixture of taxa t_1 and t_2. The relationships among the remaining $n-3$ taxa $(t_3, t_4, \ldots, t_{n-1})$ and between these taxa and t_1, t_2, and t_n are not specified; we do not consider any additional admixture relationships that might exist among these taxa. We assume $n \geq 4$, so that at least one taxon is considered in addition to t_1, t_2, and t_n.

In a standard statistical model of admixture used in population genetics, allele frequencies in an admixed taxon are given by linear combinations of the allele frequencies of the source taxa.[33–37] We denote by λ the proportion of the ancestry of taxon t_n arising from t_1 and by $1-\lambda$ the corresponding proportion arising from t_2, where $0 < \lambda < 1$. For any allelic type, if p_{t_i} denotes the frequency of the specified allele in taxon t_i, then

$$p_{t_n} = \lambda p_{t_1} + (1-\lambda) p_{t_2}. \tag{5}$$

It follows that if for each of the taxa in a pair, a distance function d is linear in each component of the allele frequency vector at a locus, then the distances between the admixed taxon and other taxa are obtained as linear combinations of corresponding distances involving taxa t_1 and t_2. Therefore, for $1 \leq i \leq n-1$,

$$d_{t_n,t_i} = \lambda d_{t_1,t_i} + (1-\lambda) d_{t_2,t_i}. \tag{6}$$

Eq. 6 continues to hold if for a series of loci, the distance function d is linear for each taxon in each component of the allele frequency vector *at each locus*, as would occur if the distance between a pair of taxa at a set of loci were computed as the mean of locus-wise distances that were each linear in the components of the allele frequency vector at the specified locus.

We assume that the distance function supplied to neighbor-joining satisfies eq. 6, and that it is symmetric, nonnegative, and zero if and only if it is computed between a taxon and itself; we otherwise do not concern ourselves with the form of the function. While typical population-genetic distance functions often involve nonlinear relationships with allele frequencies and do not necessarily follow eq. 6—consider the nonlinear graphs in Figure 3 of Boca & Rosenberg,[38] which illustrate that for the F_{ST} measure and an admixed population t_n whose frequencies are linear combinations of those of populations t_1 and t_2, $F_{ST}(t_n, t_1) \neq \lambda F_{ST}(t_1, t_1) + (1 - \lambda) F_{ST}(t_2, t_1)$—eq. 6 is a natural extension of the ubiquitous eq. 5 from allele frequencies to distance functions. For F_{ST}, it can be shown from eqs. 1 and 7 of Boca & Rosenberg[38] that for small λ, $F_{ST}(t_n, t_1) \approx \lambda F_{ST}(t_1, t_1) + (1 - \lambda) F_{ST}(t_2, t_1)$. Thus, we view eq. 6 as a reasonable first approximation for examining properties of neighbor-joining in an admixture scenario.

4. The neighbor-joining algorithm in an admixture scenario

Our goal is to construct a distance matrix according to the admixture rule in eq. 6, mechanistically apply neighbor-joining to the matrix, and characterize the properties of the inference process and the resulting inferred tree. We examine two settings. In the first, arbitrarily many taxa are considered, and their distances produce an additive distance matrix (and therefore satisfy a *tree metric*[39]). In the second, a general matrix is investigated, with distances that do not necessarily follow a tree metric, but the matrix includes only four taxa.

4.1. *The additive case for n taxa*

We first assume that the distance matrix is additive. In this case, by the consistency property of the neighbor-joining algorithm,[5,8,9] distances between taxa on the constructed neighbor-joining tree exactly equal those of the input matrix. Denote by $\hat{d}$ the distance function computed for pairs of nodes in the inferred neighbor-joining tree, such that for taxa t_i and t_j, $\hat{d}_{t_i,t_j}$ is the sum of the lengths of the branches on the path connecting t_i and t_j. Recalling that d represents distance in the input distance matrix, if the matrix is additive, then for all (t_i, t_j),

$$\hat{d}_{t_i,t_j} = d_{t_i,t_j}. \tag{7}$$

Because $d_{t_1,t_n} = (1 - \lambda) d_{t_1,t_2}$ and $d_{t_2,t_n} = \lambda d_{t_1,t_2}$ by eq. 6,

$$\hat{d}_{t_1,t_n} = (1 - \lambda) \hat{d}_{t_1,t_2} \tag{8}$$

$$\hat{d}_{t_2,t_n} = \lambda \hat{d}_{t_1,t_2}. \tag{9}$$

It follows that $\hat{d}_{t_1,t_n} + \hat{d}_{t_2,t_n} = \hat{d}_{t_1,t_2}$, from which we can infer that taxa t_1, t_2, and t_n are collinear in the inferred neighbor-joining tree, with t_n in the interior of the path from t_1 to t_2.

We can obtain an even stronger result. Consider a case with at least four taxa: t_1, t_2, t_n, and, without loss of generality, t_3 (Fig. 3A). In the inferred neighbor-joining tree, a path of length c connects taxon t_3 to some point P on the path from t_1 to t_2 (including the endpoints). Without loss of generality, we can assume that P lies on the path from t_1 to t_n (including the endpoints). We denote the distances $\hat{d}_{t_1,P}$ and $\hat{d}_{P,t_n}$ by nonnegative values y and z, respectively. We denote $\hat{d}_{t_1,t_2} = d_{t_1,t_2} = x$, for some nonnegative x.

By eq. 7, $\hat{d}_{t_1,t_n} = y + z = d_{t_1,t_n} = (1 - \lambda)x$. By eqs. 6 and 7,

$$d_{t_3,t_n} = \lambda d_{t_3,t_1} + (1 - \lambda) d_{t_3,t_2} \tag{10}$$

$$\hat{d}_{t_3,t_n} = \lambda \hat{d}_{t_3,t_1} + (1 - \lambda) \hat{d}_{t_3,t_2}. \tag{11}$$

In other words, $c + z = \lambda(c + y) + (1 - \lambda)(c + x - y)$. Together with the relationship $y + z = (1 - \lambda)x$ and the assumption that $\lambda > 0$, eq. 11 implies that $y = 0$. It then follows that taxon t_3 lies on

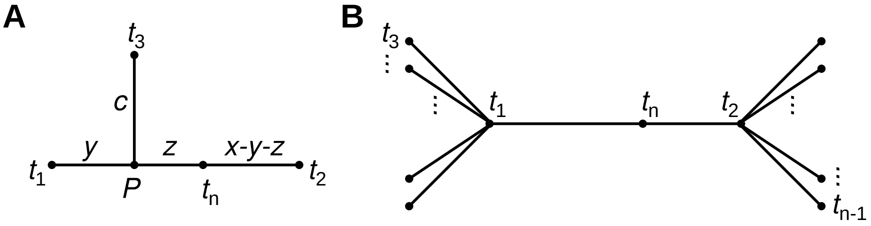


Fig. 3. The case of an additive distance matrix for n taxa. (A) Illustration of distances in the model. In the text it is shown that $y = 0$. (B) The structure required for a tree. Taxa t_1 and t_2 lie at multifurcating nodes, with t_n on the line connecting them. The leaves connected to the multifurcations have labels $t_3, t_4, \ldots, t_{n-1}$.

a line with taxa t_1, t_2, and t_n. Further, taxon t_3 lies on the side of taxon t_1 opposite to taxa t_2 and t_n; otherwise, by eq. 11, we would have $(1-\lambda)x - c = \lambda c + (1-\lambda)(x-c)$, which requires $c = 0$. In turn, $c = 0$ implies $\hat{d}_{t_3,t_1} = 0$, and hence, $d_{t_3,t_1} = 0$, contradicting the assumption that all pairs of taxa are separated by positive distances in the distance matrix.

We have therefore shown that for an additive tree with taxon t_n admixed between t_1 and t_2, any additional taxon beyond t_1, t_2, and t_n must be collinear with t_1, t_2, and t_n, and must lie exterior to the path connecting t_1 and t_2. Thus, each additional taxon $t_3, t_4, \ldots, t_{n-1}$ is connected to t_1 or t_2 by an external branch. The admixture model together with the assumption of an additive distance matrix imposes such a strong restriction on the set of allowed distance matrices that it forces all taxa onto a highly constrained tree (Fig. 3B). When we consider the placement of each taxon $t_3, t_4, \ldots, t_{n-1}$, we find that this tree has two multifurcating nodes separated by a line that joins taxa t_1 and t_2, with t_n as the only intervening taxon.

The additive case can assist in explaining phenomena observed empirically with admixed populations in the application of neighbor-joining:[21–25] in the additive case, t_n has external branch length 0, a result compatible with the short external branches detected for admixed taxa. Further, t_n lies on the path connecting t_1 and t_2, compatible with the observation that admixed taxa lie in the "middle" of inferred neighbor-joining trees, with external branches incident to the paths connecting their source taxa. We can thus see that the empirical Fig. 2 resembles Fig. 3B, as the short internal branches among Native Americans and Europeans give rise to a shape with near multifurcations on each side of the admixed Mestizo groups.

4.2. *The case of $n = 4$ taxa, not necessarily additive*

The additive case is restrictive and atypical of the population-genetic context, in which migration and admixture generate non-tree-like evolution. We can then consider the more general setting of arbitrary genetic distance matrices with positive entries, examining the smallest nontrivial case, with $n = 4$ taxa. In this case, the admixed taxon is t_4, with source taxa t_1 and t_2. We set the distances among taxa t_1, t_2, and t_3 to be $d_{t_1,t_2} = x$, $d_{t_1,t_3} = y$ and $d_{t_2,t_3} = z$, for some positive x, y, and z. Employing eq. 6, the distance matrix D has the form:

$$D = \begin{pmatrix} 0 & x & y & (1-\lambda)x \\ x & 0 & z & \lambda x \\ y & z & 0 & \lambda y + (1-\lambda)z \\ (1-\lambda)x & \lambda x & \lambda y + (1-\lambda)z & 0 \end{pmatrix}. \tag{12}$$

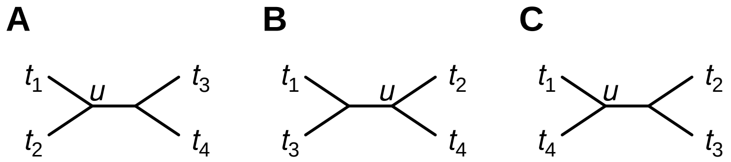


Fig. 4. The three possible topologies for $n = 4$ taxa. Node u is the unique node that places t_1, t_2, and t_4 in different subtrees.

Using eq. 1 to calculate the matrix Q used in deciding which taxa will agglomerate, we obtain

$$Q = \begin{pmatrix} 0 & q_1 & q_2 & q_3 \\ q_1 & 0 & q_3 & q_2 \\ q_2 & q_3 & 0 & q_1 \\ q_3 & q_2 & q_1 & 0 \end{pmatrix}, \tag{13}$$

where

$$\begin{aligned} q_1 &= -(x+y+z) && (14)\\ q_2 &= -(2-\lambda)x - \lambda y - (2-\lambda)z && (15)\\ q_3 &= -(1+\lambda)x - (1+\lambda)y - (1-\lambda)z. && (16) \end{aligned}$$

Examining the relationships among q_1, q_2, and q_3, we have that

$$\begin{aligned} q_1 < q_2 &\Leftrightarrow x + z < y && (17)\\ q_1 < q_3 &\Leftrightarrow x + y < z && (18)\\ q_2 < q_3 &\Leftrightarrow y < (1-2\lambda)x + z. && (19) \end{aligned}$$

As in the work of Eickmeyer & Yoshida,[40] we partition the four-dimensional space of possible values of (λ, q_1, q_2, q_3) according to the tree topologies produced by neighbor-joining.

Three tree topologies are possible with the four taxa (Fig. 4). In Fig. 4A, taxon t_4 is separated by three edges from taxa t_1 and t_2, which themselves are separated by only two edges. In Fig. 4B, t_4 is separated by two edges from t_2 and by three edges from t_1; t_1 and t_2 are separated by three edges. Taxa t_1 and t_2 are also separated by three edges in Fig. 4C, but t_4 is instead separated by two edges from t_1 and by three edges from t_2.

Seven possibilities exist for the smallest entry of Q: (1) q_1, (2) q_2, (3) q_3, (4) q_1 and q_2 (tied), (5) q_1 and q_3 (tied), (6) q_2 and q_3 (tied), and (7) q_1, q_2, and q_3 (all tied). Each choice leads to a particular outcome among the three tree topologies in Fig. 4, with two or more topologies being possible outcomes in cases that involve ties. For each value among q_1, q_2, and q_3, two pairs of taxa produce the same value in the matrix Q. It can be shown that in each case, either choice of which pair is first to agglomerate leads to the same inferred tree. Without loss of generality, we choose the pair that does not include taxon t_3.

Four of the seven cases are not possible. In case 1, summing $x + z < y$ and $x + y < z$ in eqs. 17 and 18, we obtain $x < 0$. In case 4, setting $q_1 = q_2$ in eq. 17, $x + z = y$, from which we obtain $x < 0$ using $x + y < z$ in eq. 18. Similarly, in case 5, $q_1 = q_3$ in eq. 18 produces $x < 0$ using $x + z < y$ in eq. 17. In case 7, eqs. 17-19 become equalities, leading to $x = 0$ when $x + z = y$ is substituted into eq. 19. All of these cases contradict the assumption that $x > 0$.

We consider the three allowable cases (cases 2, 3, and 6). For each of the possible inferred neighbor-joining trees, denote by u the unique interior node that places taxa t_1, t_2, and t_4 in distinct subtrees (Fig. 4). Denote by $\hat{d}$ the distance between nodes on the inferred tree. In case 2, q_2 is smallest, taxa t_2 and t_4 agglomerate first, and using eqs. 2-4, we obtain

$$\hat{d}_{u,t_2} = (\lambda/4)(3x - y + z) \tag{20}$$

$$\hat{d}_{u,t_4} = (\lambda/4)(x + y - z) \tag{21}$$

$$\hat{d}_{u,t_1} = (1 - \lambda)x. \tag{22}$$

We can show that $\hat{d}_{u,t_4} < \hat{d}_{u,t_2}$ and $\hat{d}_{u,t_4} < \hat{d}_{u,t_1}$. The first of these two inequalities is equivalent to $\lambda y < \lambda(x+z)$, which holds because $\lambda > 0$, and because $y < x + z$ by eq. 17. For the second inequality, note first that $y < (1-2\lambda)x + z$ by eq. 19. Substituting the right-hand side in place of y in eq. 21, $\hat{d}_{u,t_4}$ is less than $[2\lambda(1-\lambda)]x/4$, which in turn is less than $\hat{d}_{u,t_1}$ because $0 < \lambda < 1$.

In case 3, q_3 is smallest, taxa t_1 and t_4 agglomerate first, and using eqs. 2-4, we obtain

$$\hat{d}_{u,t_1} = [(1 - \lambda)/4](3x + y - z) \tag{23}$$

$$\hat{d}_{u,t_4} = [(1 - \lambda)/4](x - y + z) \tag{24}$$

$$\hat{d}_{u,t_2} = \lambda x. \tag{25}$$

Similarly to case 2, we show $\hat{d}_{u,t_4} < \hat{d}_{u,t_1}$ and $\hat{d}_{u,t_4} < \hat{d}_{u,t_2}$. The first inequality is equivalent to $(1-\lambda)z < (1-\lambda)(x+y)$, which holds because $\lambda < 1$, and because $z < x + y$ by eq. 18. For the second equality, $(1-2\lambda)x + z < y$ by eq. 19. Substituting the left-hand side in place of y in eq. 24, $\hat{d}_{u,t_4}$ is less than $[(2\lambda(1-\lambda)]x/4$, which in turn is smaller than $\hat{d}_{u,t_2}$ because $0 < \lambda < 1$.

Finally, in case 6, q_2 and q_3 are tied with the smallest values, and either t_2 and t_4 agglomerate first as in case 2, or t_1 and t_4 agglomerate first as in case 3. Neighbor-joining produces the tree in Fig. 4C with probability 1/2, and the tree in Fig. 4B with probability 1/2. With either choice, the same arguments used to demonstrate $\hat{d}_{u,t_4} < \hat{d}_{u,t_1}$ and $\hat{d}_{u,t_4} < \hat{d}_{u,t_2}$ in cases 2 and 3 apply, except that y is equal to (instead of greater than or less than) $(1-2\lambda)x + z$.

This collection of results demonstrates three phenomena for four-taxon trees built from distance matrices formed according to our admixture model. (1) The admixed taxon agglomerates with one of its two source taxa before the sources agglomerate with each other. Cases 2, 3, and 6 are the only ones allowable, and in these cases, the first neighbor-joining step agglomerates the admixed taxon t_4 with one of the sources. (2) Denoting by u the unique node for which the admixed taxon and its source taxa all lie in different subtrees, the distance on the neighbor-joining tree of the admixed taxon to u is smaller than the distances to u of both source taxa. We demonstrated this result in each of the allowed cases, and it therefore holds in general. (3) The number of edges separating the source taxa on the inferred neighbor-joining tree, for each source taxon, is greater than or equal to the number of edges separating the admixed taxon from the source taxon. Only the trees in Figs. 4B and 4C are possible outcomes of neighbor-joining in our model, and the result holds for each of these trees.

5. Properties

Using the four-taxon results, we can formally define three properties of a distance matrix and its resulting neighbor-joining tree. The properties are well-defined for arbitrary n, and it is possible to evaluate whether a given n-taxon distance matrix satisfies them when neighbor-joining is applied. All three properties are possessed by all matrices generated by the four-taxon case of our admixture model.

Property 1: antecedence of clustering. The admixed taxon clusters with one of its source taxa before the source taxa cluster together. Stated precisely, some clade containing t_1 but not t_2 or t_n merges with some clade containing t_n but not t_1 or t_2, or, some clade containing t_2 but not t_1 or t_n merges with some clade containing t_n but not t_1 or t_2, before any clade containing t_1 but not t_2 or t_n merges with any clade containing t_2 but not t_1 or t_n.

Here we allow a clade to have any size, and potentially only a single taxon. In identifying the steps at which t_1, t_2, and t_n merge into the neighbor-joining tree, as in our four-taxon case, to ensure that these taxa do not all merge simultaneously at the final stage, we adopt the convention that if a four-taxon stage is reached in which t_1, t_2, and t_n lie in separate subtrees, we choose to agglomerate two among these three subtrees rather than agglomerating the third one with the unique available subtree that does not contain t_1, t_2, or t_n.

Property 2: intermediacy of distances. The distance on the constructed neighbor-joining tree between the admixed taxon and either of its source taxa is smaller than the corresponding distance between the two source taxa. That is, $\hat{d}_{t_1,t_n} < \hat{d}_{t_1,t_2}$ and $\hat{d}_{t_2,t_n} < \hat{d}_{t_1,t_2}$. Equivalently, if u is the unique node in the constructed neighbor-joining tree for which t_1, t_2, and t_n lie in different subtrees, then $\hat{d}_{u,t_n} < \hat{d}_{u,t_1}$ and $\hat{d}_{u,t_n} < \hat{d}_{u,t_2}$.

Property 3: intermediacy of path lengths. The number of edges separating the source taxa in the constructed neighbor-joining tree is greater than or equal to the number of edges separating the admixed taxon and either source taxon. If we define $\hat{b}_{ij}$ as the number of edges in the path separating nodes i and j in the inferred tree, then $\hat{b}_{t_1,t_2} \geq \hat{b}_{t_1,t_n}$ and $\hat{b}_{t_1,t_2} \geq \hat{b}_{t_2,t_n}$.

We have already demonstrated that in our admixture model, Properties 2 and 3 hold for all distance matrices in the n-taxon additive case; for Property 2, using eqs. 8 and 9 and $0 < \lambda < 1$, $\hat{d}_{t_1,t_n} < \hat{d}_{t_1,t_2}$ and $\hat{d}_{t_2,t_n} < \hat{d}_{t_1,t_2}$. For Property 3, we have shown that for an n-taxon additive distance matrix, taxon t_n lies on the interior of the path connecting t_1 and t_2, and it is the only taxon so located. Thus, $\hat{b}_{t_1,t_2} = 2$, while $\hat{b}_{t_1,t_n} = \hat{b}_{t_2,t_n} = 1$, and Property 3 holds.

6. Discussion

We have examined neighbor-joining in a model in which an admixed taxon is produced from two source taxa, finding that for a four-taxon scenario, distance matrices and their resulting trees possess three properties: *antecedence of clustering*, in which the admixed population clusters with one of the sources before the sources cluster with each other; *intermediacy of distances*, in which the distance on the constructed tree between the admixed taxon and either source taxon is less than the distance between the sources; and *intermediacy of path lengths*, in which the number of edges separating the admixed taxon and either source taxon is no larger than the number of edges separating the sources. We have further shown that for an arbitrary number of taxa, the latter two properties hold when the distance matrix is additive.

By a mechanistic examination, we have found that our model has features seen in empirical

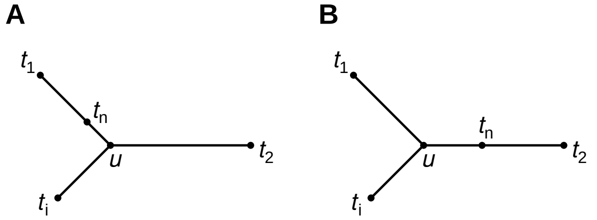


Fig. 5. Two possible placements of t_n with respect to t_1, t_2, and t_i, illustrating how $t_1, t_2, \ldots, t_n$ can have an additive distance matrix when $t_1, t_2, \ldots, t_{n-1}$ have an additive distance matrix. For convenience, we illustrate only one representative taxon t_i from among $\{t_3, t_4, \ldots, t_{n-1}\}$.

observations of neighbor-joining trees that involve admixed populations. In particular, the placement of admixed populations on short external branches incident to the paths connecting their source populations[21–25] matches the demonstration in the additive and four-taxon cases of the *intermediacy of distances* and *intermediacy of path lengths* properties. The theoretical approach validates the view that populations that are centrally located on neighbor-joining trees and that possess short external branches might be recently admixed.

Our results suggest a broader investigation of the extent to which the three properties hold with an arbitrary number of taxa. We have not reported a result regarding *antecedence of clustering* in the n-taxon additive case, nor have we commented on any of the properties for general n-taxon distance matrices that are not necessarily additive. However, we expect that Properties 1-3 will be satisfied by our admixture model considerably more often than in a model in which no special constraints are imposed on distances that involve the nth taxon. As our model also involves an nth taxon with special features, the general analysis of the model might benefit from the "rogue taxon" framework of Cueto & Matsen,[41] in which the addition of an nth taxon alters the tree produced for an initial group of $n-1$ taxa.

An additional direction is to study alternative admixture models. Distance methods are most sensible when a distance is nearly additive; however, eq. 6 severely restricts the distance matrix, as it forces a structure with two multifurcating nodes. This aspect of the model can be relaxed by assuming that the distance is additive for taxa $t_1, t_2, \ldots, t_{n-1}$, and that only distances involving t_n satisfy eq. 6. For $1 \leq i \leq n-1$, we can then apply the distance

$$d_{t_n,t_i} \quad = \quad \begin{cases} d_{t_1,t_i} - (1-\lambda)d_{t_1,t_2} & \text{if} \quad (1-2\lambda)d_{t_1,t_2} \leq d_{t_1,t_i} \\ d_{t_2,t_i} - \lambda d_{t_1,t_2} & \text{if} \quad (1-2\lambda)d_{t_1,t_2} \geq d_{t_1,t_i}. \end{cases} \tag{26}$$

With this distance function, t_n is simply placed on the path from t_1 to t_2 in a preexisting tree relating taxa $t_1, t_2, \ldots, t_{n-1}$ (Fig. 5). Properties 2 and 3 continue to hold.

To obtain eq. 26, we first suppose that $t_1, t_2, \ldots, t_{n-1}$ have an additive distance matrix. We wish to place taxon t_n on the tree that generates the matrix so that the matrix for $t_1, t_2, \ldots, t_n$ is additive. First, given λ, t_n is placed on the path from t_1 to t_2 such that eqs. 8 and 9 are satisfied. It remains to compute $\hat{d}_{t_n,t_i}$ for $i = 3, 4, \ldots, n-1$. Denote by u the unique node of the tree that places t_1, t_2, and t_i in distinct subtrees (Fig. 5). Then

$$\hat{d}_{t_1,u} \quad = \quad (\hat{d}_{t_1,t_2} + \hat{d}_{t_1,t_i} - \hat{d}_{t_2,t_i})/2 \tag{27}$$

$$\hat{d}_{t_2,u} \quad = \quad (\hat{d}_{t_1,t_2} + \hat{d}_{t_2,t_i} - \hat{d}_{t_1,t_i})/2 \tag{28}$$

$$\hat{d}_{t_i,u} \quad = \quad (\hat{d}_{t_1,t_i} + \hat{d}_{t_2,t_i} - \hat{d}_{t_1,t_2})/2. \tag{29}$$

If $\hat{d}_{t_1,t_n} \leq \hat{d}_{t_1,u}$, then t_n lies on the path from t_1 to u (Fig. 5A), and

$$\hat{d}_{t_n,t_i} = \hat{d}_{u,t_i} + \hat{d}_{t_1,t_2} - \hat{d}_{t_1,t_n}. \tag{30}$$

If, on the other hand, $\hat{d}_{t_1,t_n} \geq \hat{d}_{t_1,u}$, then t_n lies on the path from t_2 to u (Fig. 5B), and

$$\hat{d}_{t_n,t_i} = \hat{d}_{u,t_i} + \hat{d}_{t_1,t_2} - \hat{d}_{t_2,t_n}. \tag{31}$$

Applying eqs. 8, 9, and 27-29 together with the fact that $\hat{d} = d$ for additive distance matrices, we produce the relationship in eq. 26.

Analysis of the three properties using this modified form for the admixture model, or more generally using specific distance functions commonly employed in population genetics, will further illuminate the features of neighbor-joining in admixed populations. Such analyses might also facilitate investigations of the behavior with admixed populations of other tree-building methods, or of phylogenetic network methods[42] that are more directly designed to accommodate taxa with non-tree-like evolutionary histories.

Acknowledgments

We thank M. Feldman and four reviewers for comments. This work has been supported by National Institutes of Health grant R01 GM081441, by National Science Foundation grants BCS-1024627 and DBI-1146722, and by the Burroughs Wellcome Fund.

References

1. D. L. Swofford, G. J. Olsen, P. J. Waddell and D. M. Hillis, Phylogenetic inference, in *Molecular Systematics*, eds. D. M. Hillis, C. Moritz and B. K. Mable (Sinauer, Sunderland, MA, 1996) pp. 407–514.
2. J. Felsenstein, *Inferring Phylogenies* (Sinauer, Sunderland, MA, 2004).
3. N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).
4. J. A. Studier and K. J. Keppler, *Mol. Biol. Evol.* **5**, 729 (1988).
5. D. Bryant, *J. Classif.* **22**, 3 (2005).
6. O. Gascuel and M. Steel, *Mol. Biol. Evol.* **23**, 1997 (2006).
7. R. Mihaescu, D. Levy and L. Pachter, *Algorithmica* **54**, 1 (2009).
8. O. Gascuel, Concerning the NJ algorithm and its unweighted version, UNJ, in *Mathematical Hierarchies and Biology*, eds. B. Mirkin, F. R. McMorris, F. S. Roberts and A. Rzhetsky (American Mathematical Society, Providence, 1997) pp. 149–170.
9. K. Atteson, *Algorithmica* **25**, 251 (1999).
10. N. Saitou and T. Imanishi, *Mol. Biol. Evol.* **6**, 514 (1989).
11. M. K. Kuhner and J. Felsenstein, *Mol. Biol. Evol.* **11**, 459 (1994).
12. C. A. M. Russo, N. Takezaki and M. Nei, *Mol. Biol. Evol.* **13**, 525 (1996).
13. S. T. Kalinowski, *Heredity* **102**, 506 (2009).
14. A. W. F. Edwards and L. L. Cavalli-Sforza, Reconstruction of evolutionary trees, in *Phenetic and Phylogenetic Classification*, eds. V. H. Heywood and J. McNeill (Systematics Association, London, 1964) pp. 67–76.
15. L. L. Cavalli-Sforza and A. W. F. Edwards, *Evolution* **21**, 550 (1967).
16. A. M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd and L. L. Cavalli-Sforza, *Nature* **368**, 455 (1994).
17. J. K. Pritchard, M. Stephens and P. Donnelly, *Genetics* **155**, 945 (2000).

18. M. Jakobsson, S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H.-C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg and A. B. Singleton, *Nature* **451**, 998 (2008).
19. G. Atzmon, L. Hao, I. Pe'er, C. Velez, A. Pearlman, P. F. Palamara, B. Morrow, E. Friedman, C. Oddoux, E. Burns and H. Ostrer, *Am. J. Hum. Genet.* **86**, 850 (2010).
20. K. Hunley and M. Healy, *Am. J. Phys. Anthropol.* **146**, 530 (2011).
21. A. Ruiz-Linares, E. Minch, D. Meyer and L. L. Cavalli-Sforza, Analysis of classical and DNA markers for reconstructing human population history, in *The Origin and Past of Modern Humans as Viewed from DNA*, eds. S. Brenner and K. Hanihara (World Scientific, Singapore, 1995) pp. 123–148.
22. A. M. Bowcock, J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto, K. K. Kidd and L. L. Cavalli-Sforza, *Proc. Natl. Acad. Sci. USA* **88**, 839 (1991).
23. J. L. Mountain, A. A. Lin, A. M. Bowcock and L. L. Cavalli-Sforza, *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **337**, 159 (1992).
24. A. A. Lin, J. M. Hebert, J. L. Mountain and L. L. Cavalli-Sforza, *Gene Geog.* **8**, 191 (1994).
25. J. L. Mountain and L. L. Cavalli-Sforza, *Proc. Natl. Acad. Sci. USA* **91**, 6515 (1994).
26. J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.6 (Department of Genome Sciences, University of Washington, Seattle, 2005).
27. N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard and M. W. Feldman, *PLoS Genet.* **1**, 660 (2005).
28. S. Wang, C. M. Lewis Jr., M. Jakobsson, S. Ramachandran, N. Ray, G. Bedoya, W. Rojas, M. V. Parra, J. A. Molina, C. Gallo, G. Mazzotti, G. Poletti, K. Hill, A. M. Hurtado, D. Labuda, W. Klitz, R. Barrantes, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, L. T. Tsuneto, E. Llop, F. Rothhammer, L. Excoffier, M. W. Feldman, N. A. Rosenberg and A. Ruiz-Linares, *PLoS Genet.* **3**, 2049 (2007).
29. S. Wang, N. Ray, W. Rojas, M. V. Parra, G. Bedoya, C. Gallo, G. Poletti, G. Mazzotti, K. Hill, A. M. Hurtado, B. Camrena, H. Nicolini, W. Klitz, R. Barrantes, J. A. Molina, N. B. Freimer, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, L. T. Tsuneto, J. E. Dipierri, E. L. Alfaro, G. Bailliet, N. O. Bianchi, E. Llop, F. Rothhammer, L. Excoffier and A. Ruiz-Linares, *PLoS Genet.* **4**, e1000037 (2008).
30. E. Minch, A. Ruiz Linares, D. B. Goldstein, M. W. Feldman and L. L. Cavalli-Sforza, MICROSAT (version 1.5d): a program for calculating statistics on microsatellite data (Department of Genetics, Stanford University, Stanford, CA, 1998).
31. J. L. Mountain and L. L. Cavalli-Sforza, *Am. J. Hum. Genet.* **61**, 705 (1997).
32. O. Gascuel, *Mol. Biol. Evol.* **11**, 961 (1994).
33. J. C. Long and P. E. Smouse, *Am. J. Phys. Anthropol.* **61**, 411 (1983).
34. D. A. Fournier, T. D. Beacham, B. E. Riddell and C. A. Busack, *Can. J. Fish. Aquat. Sci.* **41**, 400 (1984).
35. N. A. Rosenberg, L. M. Li, R. Ward and J. K. Pritchard, *Am. J. Hum. Genet.* **73**, 1402 (2003).
36. H. Tang, J. Peng, P. Wang and N. J. Risch, *Genet. Epidemiol.* **28**, 289 (2005).
37. D. H. Alexander, J. Novembre and K. Lange, *Genome Res.* **19**, 1655 (2009).
38. S. M. Boca and N. A. Rosenberg, *Theor. Pop. Biol.* **80**, 208 (2011).
39. C. Semple and M. Steel, *Phylogenetics* (Oxford University Press, Oxford, 2003).
40. K. Eickmeyer and R. Yoshida, *Lect. Notes Comp. Sci.* **5147**, 81 (2008).
41. M. A. Cueto and F. A. Matsen, *Bull. Math. Biol.* **73**, 1202 (2011).
42. D. H. Huson, R. Rupp and C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications* (Cambridge University Press, Cambridge, 2010).

MAXIMUM LIKELIHOOD PHYLOGENETIC RECONSTRUCTION FROM HIGH-RESOLUTION WHOLE-GENOME DATA AND A TREE OF 68 EUKARYOTES

YU LIN*
Laboratory for Computational Biology and Bioinformatics, EPFL,
Lausanne VD, CH-1015, Switzerland
**E-mail: yu.lin@epfl.ch*

FEI HU and JIJUN TANG
Department of Computer Science and Engineering, University of South Carolina,
Columbia, SC 29208, USA
E-mail: {hu5,jtang}@cse.sc.edu

BERNARD M.E. MORET
Laboratory for Computational Biology and Bioinformatics, EPFL,
Lausanne VD, CH-1015, Switzerland
E-mail: bernard.moret@epfl.ch

The rapid accumulation of whole-genome data has renewed interest in the study of the evolution of genomic architecture, under such events as rearrangements, duplications, losses. Comparative genomics, evolutionary biology, and cancer research all require tools to elucidate the mechanisms, history, and consequences of those evolutionary events, while phylogenetics could use whole-genome data to enhance its picture of the Tree of Life. Current approaches in the area of phylogenetic analysis are limited to very small collections of closely related genomes using low-resolution data (typically a few hundred syntenic blocks); moreover, these approaches typically do not include duplication and loss events. We describe a maximum likelihood (ML) approach for phylogenetic analysis that takes into account genome rearrangements as well as duplications, insertions, and losses. Our approach can handle high-resolution genomes (with 40,000 or more markers) and can use in the same analysis genomes with very different numbers of markers. Because our approach uses a standard ML reconstruction program (RAxML), it scales up to large trees. We present the results of extensive testing on both simulated and real data showing that our approach returns very accurate results very quickly. In particular, we analyze a dataset of 68 high-resolution eukaryotic genomes, with from 3,000 to 42,000 genes, from the eGOB database; the analysis, including bootstrapping, takes just 3 hours on a desktop system and returns a tree in agreement with all well supported branches, while also suggesting resolutions for some disputed placements.

Keywords: Maximum likelihood; Phylogenetic reconstruction; Genome rearrangement; Gene duplication; Gene loss

1. Introduction

1.1. *Overview*

Phylogenetic analysis is one of the main tools of evolutionary biology. Most of it to date has been carried out using sequence data (or, more rarely, morphological data). Sequence data can be collected in large amounts at very low cost and, at least in the case of coding genes, is relatively well understood, but it requires accurate determination of orthologies and gives us only local information—and different parts of the genome may evolve at different rates or according to different models. Events that affect the structure of an entire genome may hold the key to building a coherent picture of the past history of contemporary organisms. Such events occur at a much larger scale than sequence mutations—entire blocks of a genome may be permuted (rearrangements), duplicated, or lost. As whole genomes are sequenced at increasing rates, using whole-genome data for phylogenetic analyses is attracting increasing interest, especially as researchers uncover links between large-scale genomic events (rearrangements, duplications leading to increased copy numbers) and various diseases (such as cancer) or health conditions (such as autism). However, using whole-genome data in phylogenetic reconstruction has proved far more challenging than using sequence data and numerous problems plague existing methods: oversimplified models, poor accuracy, poor scaling, lack of robustness, lack of statistical assessment, etc.

In this paper, we describe a new approach that resolves these problems and promises to open the way to widespread use of whole-genome data in phylogenetic analysis.

1.2. *Prior work*

Rearrangement data was first used in phylogenetic analysis 80 years ago by Sturtevant and Dobzhansky,[1] but largely ignored for the next 45 years, until revived by Palmer and Thompson[2,3] and Day and Sankoff.[4] In the last 30 years, models of whole-genome evolution, their corresponding distance measures, and algorithms for reconstructing phylogenies under such models, have been the subject of intense research, for which see the text of Fertin *et al.*[5] As in sequence-based phylogenetic reconstruction, approaches based on whole-genome data can be classified in three main categories.

Parsimony-based approaches seek the tree and internal genomes that minimize the total number of events needed to produce the given genomes from a common ancestor. Blanchette *et al.* introduced the first algorithmic approach to the reconstruction of a phylogenetic tree to minimize the total number of *breakpoints*—adjacencies present in one genome, but absent in the other.[6] Moret *et al.* reimplemented this approach in their GRAPPA tool and extended it to *inversion distances*—inversions being the best documented of the hypothesized mechanisms of genomic rearrangements.[7] GRAPPA focused on unichromosomal genomes; to handle multichromosomal genomes, Bourque and Pevzner proposed MGR,[8] based on GRAPPA's distance computations. Whereas BPAnalysis and GRAPPA search all trees and report the one with the best score (an approach that limits GRAPPA to trees of 15 taxa unless combined with the DCM approach of Tang and Moret[9]), MGR uses a heuristic sequential addition method to grow the tree one species at a time. This heuristic approach trades accuracy for scalability, yet MGR does not scale well—in particular, it cannot be used to infer a phylogeny from modern high-resolution data. These various methods are all limited to rearrangements—extensions to handle gene[a] duplications, insertions and losses appear extremely complex and would further limit their scalability.

[a] We use the word "gene" as this is in fact a common form of whole-genome data, but other kinds of markers could be used; more generally, the constituents are syntenic blocks.

Distance-based approaches first estimate the pairwise distances between every pair of leaves, then apply a method such as Neighbor-Joining[10] or FastME[11] to reconstruct the phylogeny from the matrix of pairwise distances. For unichromosomal genomes under inversions, transpositions, and inverted transpositions, Wang and Warnow showed how to estimate a true evolutionary distance from the number of breakpoints.[12,13] For unichromosomal genomes evolving under inversions only, an experimental approach was used by Moret *et al.* to derive an estimate from the inversion edit distance, yielding greatly increased accuracy in tree estimation under both distance and parsimony methods.[14] For multichromosomal genomes, rearrangement operations can be modeled by a single operation called "Double-Cut-and-Join (DCJ)".[15] Lin and Moret developed a procedure to estimate the true evolutionary distance between two genomes under the DCJ model;[16] Lin *et al.* then refined the estimator to include gene duplication and loss events,[17] although that estimator requires knowledge of the direction of time, something usually missing in phylogenetic estimation. The accuracy of distance methods depends entirely on the accuracy of distance estimation and any distance estimator suffers from the saturation problem: as the measured distance increases beyond a certain threshold, the variance in the estimator grows significantly.

Maximum-likelihood (ML) approaches seek the tree and associated model parameters that maximize the probability of producing the given set of leaf genomes. Theoretically, such approaches are much more computationally expensive than both distance-based and parsimony-based approaches, but their accuracy has long been a major attraction in sequence-based phylogenetic analysis. Moreover, in the last few years, packages such as RAxML[18] have largely overcome computational limitations and allowed reconstructions of large trees (with thousands of taxa) and the use of long sequences (to a hundred thousand characters). It was not until last year, however, that the first successful attempt to use ML reconstruction based on whole-genome data was published;[19] results from this study on bacterial genomes were promising, but somewhat difficult to explain, while the method appeared too time-consuming to handle eukaryotic genomes.

2. Methods

Our approach encodes the whole-genome data into binary sequences using both gene adjacencies and gene content, then estimates the transition parameters for the resulting binary sequence data, and finally uses sequence-based ML reconstruction to infer the tree. We call our new approach *Maximum Likelihood on Whole-genome Data (MLWD).*

2.1. *Encoding genomes into binary sequences*

We represent the genome in terms of adjacency information and gene content as follows. Denote the tail of a gene g by g^t and its head by g^h. We write $+g$ to indicate an orientation from tail to head ($g^t \rightarrow g^h$), $-g$ otherwise ($g^h \rightarrow g^t$). Two consecutive genes a and b can be connected by one *adjacency* of one of the following four types: $\{a^t, b^t\}$, $\{a^h, b^t\}$, $\{a^t, b^h\}$, and $\{a^h, b^h\}$. If gene c lies at one end of a linear chromosome, then we have a corresponding singleton set, $\{c^t\}$ or $\{c^h\}$, called a *telomere*. A *genome* can then be represented as a multiset of adjacencies and telomeres. For example, a toy genome composed of one linear chromosome,

	adjacency information						content information			
	$\{a^h, a^h\}$	$\{a^t, b^h\}$	$\{a^t, c^h\}$	$\{b^t, c^t\}$	$\{a^h, d^h\}$	$\{b^t, d^t\}$	a	b	c	d
Genome 1	1	1	1	1	0	0	1	1	1	0
Genome 2	0	1	0	0	1	1	1	1	0	1

(+a,+b,-c,+a,+b,-d,+a), and one circular one, (+e,-f), can be represented by the multiset of adjacencies and telomeres $\{\{a^t\}$, $\{a^h, b^t\}$, $\{b^h, c^h\}$, $\{c^t, a^h\}$, $\{a^h, b^t\}$, $\{b^h, d^h\}$, $\{d^t, a^h\}$, $\{a^h\}$, $\{e^h, f^h\}$, $\{e^t, f^t\}\}$. In the presence of duplicated genes, there is no one-to-one correspondence between genomes and multisets of genes, adjacencies, and telomeres. For example, the genome composed of the linear chromosome (+a, +b, -d, +a, +b, -c, +a) and the circular one (+e, -f), would have the same multisets of adjacencies and telomeres as our toy example.

For data limited to rearrangements (i.e. for genomes with identical gene content), we encode only the adjacency information. For a possible adjacency or telomere, we write 1 (or 0) to indicate its presence (or absence) in a genome. We consider only those adjacencies and telomeres that exist in at least one of the input genomes. If the total number of distinct genes among the input genomes is n, then the total number of distinct adjacencies and telomeres is $\binom{2n+2}{2}$, but the number of adjacencies and telomeres that appear in at least one input genome is typically far smaller—in fact, it is usually linear in n rather than quadratic. For the general model, which includes gene duplications, insertions, and losses in addition to rearrangements, we extend the encoding of adjacencies by also encoding the gene content. For each gene, we write 1 (or 0) to indicate the presence (or absence) of this gene in a genome. For the two toy genomes of Figure 1, the resulting binary sequences and their derivation are shown in Table 1.

2.2. *Estimating transition parameters*

Since our encodings are binary sequences, the parameters of the model are simply the transition probability from presence (1) to absence (0) and that from absence (0) to presence (1). Let us first look at adjacencies. Every DCJ operation will select two adjacencies (or telomeres) uniformly at random, and (if adjacencies) break them to create two new adjacencies. Each genome has $n + O(1)$ adjacencies and telomeres ($O(1)$ is the number of linear chromosomes in the genome, viewed as a small constant). Thus the transition probability from 1 to 0 at some fixed index in the sequence is $\frac{2}{n+O(1)}$ under one DCJ operation. Since there are up to $\binom{2n+2}{2}$ possible adjacencies and telomeres, the transition probability from 0 to 1 is $\frac{2}{2n^2+O(n)}$. Thus the transition from 0 to 1 is roughly $2n$ times less likely than that from 1 to 0. Despite the restrictive assumption that all DCJ operations are equally likely, this result is in line with general opinion about the probability of eventually breaking an ancestral adjacency (high) vs. that of creating

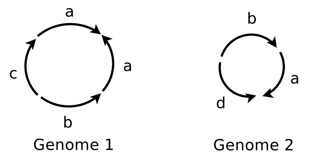


Fig. 1. Two toy genomes.

a particular adjacency along several lineages (low)—a version of homoplasy for adjacencies.

In the general model, we also have transitions for gene content. Once again, the probability of losing a gene independently along several lineages is high, whereas the probability of gaining the same gene independently along several lineages (the standard homoplasy) is low. However, there is no simple uniformity assumption that would enable us to derive a formula for the respective probabilities—there have been attempts to reconstruct phylogenies based on gene content only,[20–22] but they were based on a different approach—so we experimented with various values of the ratio between the probability of a transition from 1 to 0 and that of a transition from 0 to 1.

2.3. *Reconstructing the phylogeny*

Once we have the binary sequences encoding the input genomes and have computed the transition parameters, we use the ML reconstruction program RAxML[18] (version 7.2.8 was used to produce the results given in this paper) to build a tree from these sequences. Because RAxML uses a time-reversible model, it estimates the transition parameters directly from the input sequences by computing the base frequencies. In order to set up the $2n$ ratio, we simply add a direct assignment of the two base frequencies in the code.

3. Results

3.1. *Experimental Design*

We ran a series of experiments on simulated datasets in order to evaluate the performance of our approach against a known "ground truth" under a wide variety of settings. We then ran our reconstruction algorithm on a dataset of 68 eukaryotic genomes, from unicellular parasites to mammalians, obtained from the *Eukaryotic Gene Order Browser (eGOB)* database.[23]

Our simulation studies follow standard practice in phylogenetic reconstruction.[24] We generate model trees under various parameter settings, then use each model tree to evolve an artificial root genome from the root down to the leaves, by performing randomly chosen evolutionary events on the current genome, finally obtaining datasets of leaf genomes for which we know the complete evolutionary history. We then reconstruct trees for each dataset by applying different reconstruction methods and compare the results against the model tree.

3.1.1. *Simulating phylogenetic trees*

A model tree consists of a rooted tree topology and corresponding branch lengths. The trees are generated by a three-step process. We first generate birth-death trees using the tree generator (from the geiger library) in the software R[25] (with a birth rate of 0.001 and a death rate of 0), which simulates the development of a model tree under a uniform, time-homogeneous birth-death process. The branch lengths in such trees are ultrametric (the root-to-leaf paths all have the same length), so, in the second step, the branch lengths are modified as follows. We choose a parameter c; for each branch we sample a number s uniformly from the interval $[-c, +c]$ and multiply the original branch length by e^s (for the experiments in this paper, we set $c = 2$). Thus, each branch length is multiplied by a possibly different random number.

Finally, we rescale all branch lengths to achieve a target diameter D (the length of the longest path, defined as the sum of the edge lengths along that path) for the model tree. (Note that the unit of "length" is one expected evolutionary operation.)

Our experiments are conducted by varying three main parameters: the number of taxa , the number of genes, and the target diameter. We used two values for each of the first two parameters: 50 and 100 taxa, and $1,000$ and $5,000$ genes. For the third parameter, the diameter of the tree, we varied it from n to $4n$, where n is the number of genes. For each setting of the parameters, we generated 100 datasets; data presented below are averages over these 100 datasets.

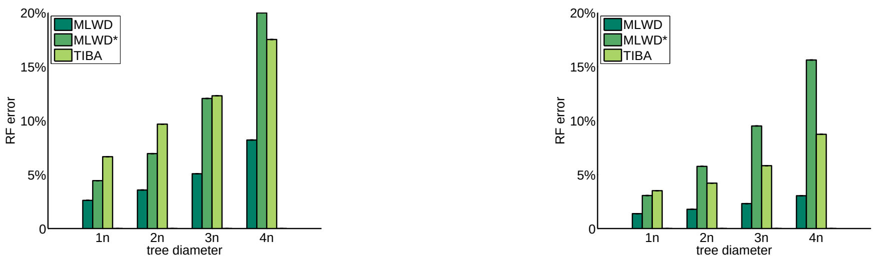
3.1.2. *Simulating evolutionary events along branches in the trees*

In the rearrangement-only model, all evolutionary events along the branches are DCJ operations. The next event is then chosen uniformly at random among all possible DCJ operations.

In the general model, an event can be a DCJ operation or one of a gene duplication, gene insertion, or gene loss. Thus we randomly sample three parameters for each branch: the probability of occurrence of a gene duplication, p_d, the probability of occurrence of a gene insertion, p_i and the probability of occurrence of a gene loss, p_l. (The probability of occurrence of a DCJ operation is then just $p_r = 1 - p_d - p_i - p_l$.) The next evolutionary event is chosen randomly from the four categories according to these parameters. For gene duplication, we uniformly select a position to start duplicating a short segment of chromosomal material and place the new copy to a new position within the genome. We set $L_{\max}$ as the maximum number of genes in the duplicated segment and assume that the number of genes in that segment is a uniform random number between 1 and $L_{\max}$. In our simulations, we used $L_{\max} = 5$. For gene insertion, we tested two different possible scenarios, one for genomes of prokaryotic type and the other for genomes of eukaryotic type. For the former, we uniformly select one position and insert a new gene; for the latter, we uniformly select one existing gene and mutate it into a new gene. Finally, for gene loss, we uniformly select one gene and delete it.

3.2. *Results for simulations under rearrangements*

We compared the accuracy of three different approaches, MLWD, MLWD* and TIBA. MLWD (Maximum Likelihood on Whole-genome Data) is our new approach; MLWD* follows the same procedure as MLWD, but does not use our computation of transition probabilities—instead,



(a) 1,000 genes

(b) 5,000 genes

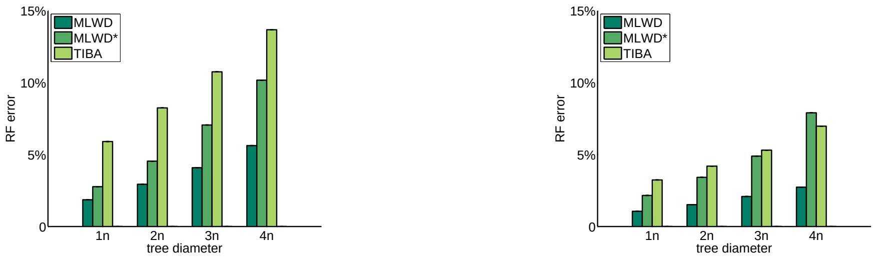
Fig. 2. RF error rates for different approaches for trees with 50 species, with genomes of $1,000$ and $5,000$ genes and tree diameters from one to four times the number of genes, under the rearrangement model.

it allows RAxML to estimate and set them; finally, TIBA is a fast distance-based tool to reconstruct phylogenies from rearrangement data,[26] which combines a pairwise distance estimator[16] and the FastME[11] distance-based reconstruction method. We did not compare with the approaches of Hu *et al.*[19] or those of Cosner *et al.*,[27] because both are too slow and because the former is also limited by their character encodings to a maximum of 20 taxa. Figures 2 and 3 show error rates for different approaches; the x axis indicates the error rates and the y axis indicates the tree diameter. Error rates are RF error rates,[28] the standard measure of error for phylogenetic trees—the RF rate expresses the percentage of edges in error, either because they are missing or because they are wrong.

These representative simulations show that our MLWD approach can reconstruct much more accurate phylogenies from rearrangement data than the distance-based approach TIBA, in line with experience in sequence-based reconstruction. MLWD also outperforms MLWD*, underlining the importance of estimating and setting the transition parameters before applying the sequence-based ML method.

3.3. *Results for simulations under the general model*

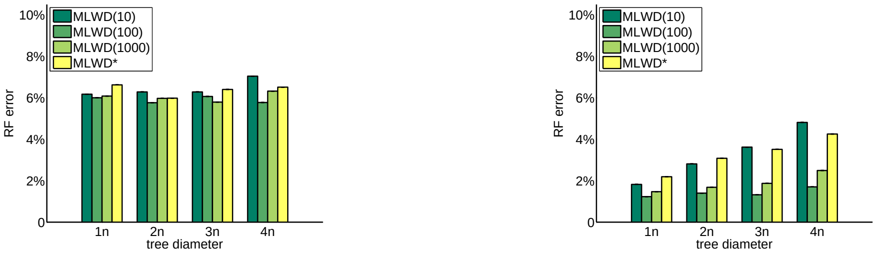
Here we generated more complex datasets than for the previous set of experiments. For example, among our simulated eukaryotic genomes, the largest genome has more than 20,000 genes, and the biggest gene family in a single genome has 42 members. In our approach, the encoded sequence of each genome combines both the adjacency and gene content information, which makes it difficult to compute optimal transition probabilities, as discussed in Section 2.2. Thus we set different bias values and compare them under simulation results. If the transition probability of any gene or adjacency from 0 to 1 in MLWD is set to be m times less than that in the opposite direction, we name it MLWD(m) $(m = 10, 100, 1000)$. Figure 4 summarizes the RF error rates. Whereas the best ratio in the rearrangement model was $2n$ (as derived in Section 2.2), the best ratio under the general model is much smaller. This difference can be attributed to the relatively modest change in gene content compared to the change in adjacencies: since we encode presence or absence of a gene, but not the number of copies of the gene, not only rearrangements, but also many duplication and loss events will not alter the encoded gene content.



(a) 1,000 genes

(b) 5,000 genes

Fig. 3. RF error rates for different approaches for trees with 100 species, with genomes of $1,000$ and $5,000$ genes and tree diameters from one to four times the number of genes, under the rearrangement model.



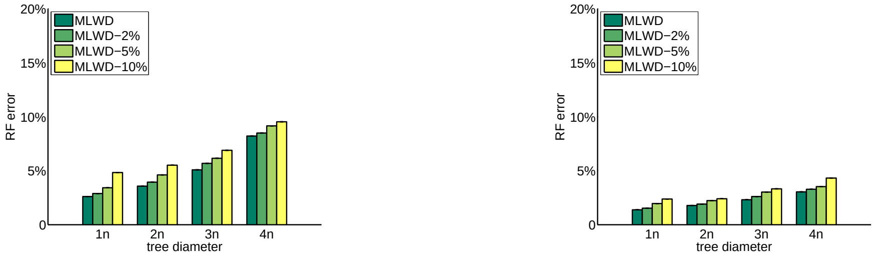
(a) 1,000 genes *(b) 5,000 genes*

Fig. 4. RF error rates for different approaches for trees with 50 species, with initial genomes of size 1,000 and 5,000 and tree diameters from one to four times the number of genes in the initial genome, under the general model of evolution.

3.4. *Results for simulated poor assemblies*

High-throughput sequencing has made it possible to sequence many genomes, but the finishing steps—producing a good assembly from the sequence data—are time-consuming and may require much additional laboratory work. Thus many sequenced genomes remain broken into a number of contigs, thereby inducing a loss of adjacencies in the source data. In addition, some assemblies may have errors, thereby producing spurious adjacencies while losing others. We designed experiments to test the robustness of our approach in handling genomes with such assembly defects. We introduce artificial breakages in the leaf genomes by "losing" adjacencies, which correspondingly breaks chromosomes into multiple contigs. For example, MLWD-x% represents the cases of losing x% of adjacencies, that is, x% of the adjacencies are selected uniformly at random and discarded for each genome.

Figure 5 shows RF error rates for MLWD on different quality of genome assemblies under the rearrangement model. Our approach is relatively insensitive to the quality of assembly, especially when the tree diameter is large, that is, when it includes highly diverged taxa. Note that this finding was to be expected in view of the good results of our approach using an encoding that, as observed earlier, does not uniquely identify the ordering of the genes along the chromosomes.



(a) 1,000 genes *(b) 5,000 genes*

Fig. 5. RF error rates for MLWD on different qualities of genome assemblies, for trees with 50 species, with genomes of size 1,000 and 5,000. with tree diameters from one to four times the number of genes, under the rearrangement model.

3.5. *Results for a dataset of high-resolution eukaryotic genomes*

Figure 6 shows the reconstructed phylogeny of 68 eukaryotic genomes from the eGOB (Eukaryotic Gene Order Browser) database.[23] The database contains the order information of orthologous genes (identified by OrthoMCL[29]) of 74 different eukaryotic species. The total number of different gene markers in eGOB is around $100,000$. We selected 68 genomes for their size (the number of gene markers) varying from $3k$ to $42k$; the remaining 6 genomes in the database have too few adjacencies (fewer than $3,000$). We encode the adjacency and gene content information of all 68 genomes into 68 binary sequences of length $652,000$. We set the bias ratio to be 100, according to the result of our simulation studies from Section 3.3. Building this phylogeny (using RAxML with fast bootstrapping) took under 3 hours of computing time on a desktop computer.

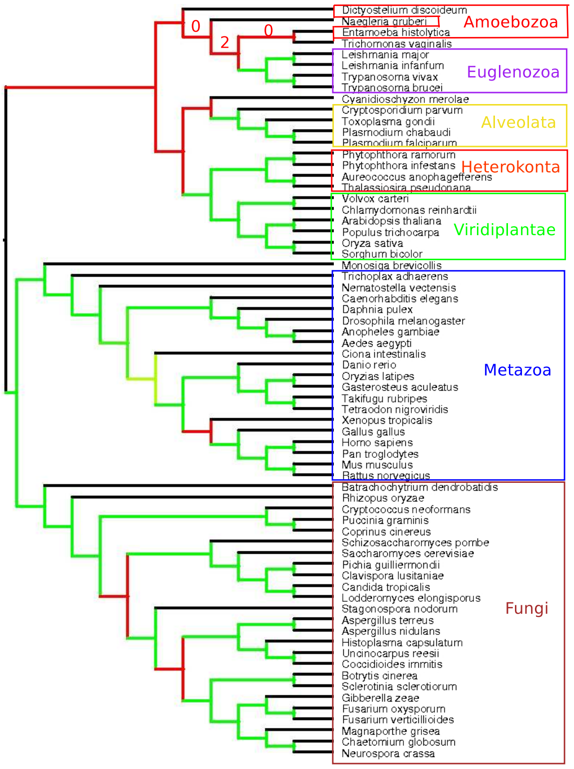


Fig. 6. The reconstructed phylogeny of 68 eukaryotic genomes

The tree is drawn by the tool iTOL;[30] the internal branches are colored into green, yellow and red, indicating, respectively, strong support (bootstrap value > 90), medium support (bootstrap value between 60 and 90), and weak support (bootstrap value < 60). As shown in Figure 6, all major groups in those 68 eukaryotic genomes are correctly identified, with the exception of Amoebozoa. But those incorrect branches with respect to Amoebozoa do receive extremely low bootstrap values (0 and 2), indicating that they are very likely to be wrong. For the phylogeny of Metazoa, the tree is well supported from existing studies.[31,32] For the phylogeny of model fish species (D. rerio, G. aculeatus, O. latipes, T. rubripes, and

T. nigroviridis), two conflicting phylogenies have been published, using different choices of alignment tools and reconstruction methods for sequence data.[33] Our result supports the second phylogeny, which is considered as the correct one by the authors in their discussion.[33] For the phylogeny of Fungi, our results agree with most branches for common species in recent studies.[34,35] It is worth mentioning that among three Chytridiomycota species C. cinereus, P. gramnis, and C. neoformans, our phylogeny shows that C. cinereus and P. gramnis are more closely related, which conflicts with the placement of C. cinereus and C. neoformans as sister taxa, but with very low support value (bootstrapping score 35).[35] C. merolae, a primitive red algae, has been the topic of a longrunning debate over its phylogenetic position.[36] Our result suggests that C. merolae is closer to Alveolata than to Viridiplantae, in agreement with a recent finding obtained by sequencing and comparing expressed sequence tags from different genomes.[37]

Finally, in order to explore the relationship between gene content and gene order, we ran MLWD* on the 68 eukaryotic genomes using only adjacency information as well as using only content information. The tree reconstructed from adjacency information only is poor, with even major clades getting mixed—an unsurprising result in view of the huge variation in gene content among these 68 genomes. The tree reconstructed from gene-content information only correctly identifies all major groups except Amoebozoa; however, it suffers from some major discrepancies with our current understanding of several clades. For example, X. tropicalis is thought to be closer to mammals than to fishes.[38] H. capsulatum, U. reesii, and C. immitis are considered to be in the same order (Onygenales); together with A. nidulans and A. terreus they are considered to be in the same class (Eurotiomycetes), but S. nodorum is thought to belong to a different class (Dothideomycetes).[35] In this particular dataset, which is a sparse sampling of the entire eukaryotic branch of the Tree of Life, most genomes differ significantly in gene content, so that we would expect the tree based on gene-content information to be close to that obtained with both gene adjacencies and gene content. For a denser sampling or for a tree of closely related genomes, adjacency information becomes crucial. A distinguishing feature of MLWD is that it uses both at once and to good effect.

4. Conclusion

In spite of many compelling reasons for using whole-genome data in phylogenetic reconstruction, practice to date has continued to use selected sequences of moderate length using nucleotide-, aminoacid-, or codon-level models. Such models are of course much simpler and much better studied than models for the evolution of genomic architecture. Mostly though, it is the lack of suitable tools that has prevented more widespread use of whole-genome data: previous tools all suffered from serious problems, usually combinations of oversimplified models, poor accuracy, poor scaling, lack of robustness against errors in the data, and lack of any bootstrapping or other statistical assessment procedures.

The approach we presented is the first to overcome all of these difficulties: it uses a fairly general model of genomic evolution (rearrangements plus duplications, insertions, and losses of genomic regions), is very accurate, scales as well as sequence-based approaches, is quite robust against typical assembly errors and omissions of genes, and supports standard bootstrapping methods. Our analysis of a 68-taxon collection of eukaryotic genomes, ranging from parasitic unicellular organisms with simple genomes to mammals and from around 3,000 genes to over 40,000 genes, could not have been conducted, regardless of computational resources, with any other tools without accepting severe compromises in the data (e.g., equalizing gene content) or the quality of the analysis (by using a distance-based reconstruction method). Our analysis also helps make the case for phylogenetic reconstruction based on whole-genome data. We did not need to choose particular regions of genomes nor to process the data from the eGOB database in any manner; in particular, we did not need to perform a multiple sequence alignment. We were able to run a complete analysis on a "Tree of Life" of all main branches of the Eukaryota, with very divergent genomes (and hence very large pairwise distances), without taking any special precautions and without preinterpreting the data (and thus possibly biasing the output). We could do all of this in a few hours on a desktop machine—in spite of the very long sequences produced by our encoding. We could run the identical software on a collection of organellar genomes or of bacterial genomes with equal success (and in much less time).

Naturally, much work remains to be done. In particular, given the complexity of genomic architecture, current evolutionary models (such as the one we used) are too simple, although even at that level, we need to elucidate simple parameters, such as the ratio of the transition probabilities between loss and gain of a given gene. Using different transition probabilities for adjacencies and for content, by running a compartmentalized analysis, should prove beneficial on large datasets. Larger issues of data preparation also loom. For instance, moving from an assembled genome to the type of data we used continues to require manual intervention—gene-finding, or syntenic block decomposition, are too complex for fully automated procedures. Determination of orthologies, necessary to the identification of syntenic blocks, should be done on the basis of a known phylogeny: that is, the same interdependence exists at the whole-genome level between reconstruction and preprocessing (orthology) as at the sequence level, where it is between reconstruction and alignment. Indeed, most of the methodological questions that the phylogenetic community has been studying in the context of sequence-based reconstruction also arise, in suitably modified terms, in the context of whole-genome data. Our new method provides a first means of empirical enquiry into these questions.

References

1. A. Sturtevant and T. Dobzhansky, *Proc. Nat'l Acad. Sci., USA* **22**, 448 (1936).
2. J. Palmer and W. Thompson, *Proc. Nat'l Acad. Sci., USA* **78**, 5533 (1981).
3. J. Palmer and W. Thompson, *Cell* **29**, 537 (1982).
4. W. Day and D. Sankoff, *J. Theor. Biol.* **127**, 213 (1987).
5. G. Fertin, A. Labarre, I. Rusu, E. Tannier and S. Vialette, *Combinatorics of Genome Rearrangements* (MIT Press, 2009).
6. M. Blanchette, G. Bourque and D. Sankoff, Breakpoint phylogenies, in *Genome Informatics*, eds. S. Miyano and T. Takagi (Univ. Academy Press, Tokyo, 1997) pp. 25–34.
7. B. Moret, S. Wyman, D. Bader, T. Warnow and M. Yan, A new implementation and detailed study of breakpoint analysis, in *Proc. 6th Pacific Symp. on Biocomputing (PSB'01)*, (World Scientific Pub., 2001).
8. G. Bourque and P. Pevzner, *Genome Res.* **12**, 26 (2002).
9. J. Tang and B. Moret, Scaling up accurate phylogenetic reconstruction from gene-order data, in

Proc. 11th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'03), , Bioinformatics Vol. 19 (Oxford U. Press, 2003).

10. N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).
11. R. Desper and O. Gascuel, *J. Comput. Biol.* **9**, 687 (2002).
12. L.-S. Wang and T. Warnow, Estimating true evolutionary distances between genomes, in *Proc. 1st Workshop Algs. in Bioinf. (WABI'01)*, Lecture Notes in Comp. Sci.(2149) (Springer Verlag, Berlin, 2001).
13. L.-S. Wang, Exact-IEBP: a new technique for estimating evolutionary distances between whole genomes, in *Proc. 33rd Ann. ACM Symp. Theory of Comput. (STOC'01)*, (ACM Press, New York, 2001).
14. B. Moret, J. Tang, L.-S. Wang and T. Warnow, *J. Comput. Syst. Sci.* **65**, 508 (2002).
15. S. Yancopoulos, O. Attie and R. Friedberg, *Bioinformatics* **21**, 3340 (2005).
16. Y. Lin and B. Moret, Estimating true evolutionary distances under the DCJ model, in *Proc. 16th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'08)*, , Bioinformatics Vol. 24(13)2008.
17. Y. Lin, V. Rajan, K. Swenson and B. Moret, Estimating true evolutionary distances under rearrangements, duplications, and losses, in *Proc. 8th Asia Pacific Bioinf. Conf. (APBC'10)*, , BMC Bioinformatics Vol. 11 (Suppl. 1)2010. S54.
18. A. Stamatakis, *Bioinformatics* **22**, 2688 (2006).
19. F. Hu, N. Gao, M. Zhang and J. Tang, Maximum likelihood phylogenetic reconstruction using gene order encodings, in *Proc. IEEE Symp. Comput. Intell. in Bioinf. & Comput. Biol. (CIBCB'11)*, (IEEE, 2011).
20. B. Snel, P. Bork and M. Huynen, *Nature Genetics* **21**, 108 (1999).
21. D. Huson and M. Steel, *Bioinformatics* **20**, 2044 (2004).
22. H. Zhang, Y. Zhong, B. Hao and X. Gu, *Gene* **441**, 163 (2009).
23. M. López and T. Samuelsson, *Bioinformatics* (2011).
24. D. Hillis and J. Huelsenbeck, *Science* **267**, 255 (1995).
25. R Development Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, (2009).
26. Y. Lin, V. Rajan and B. Moret, *J. Comput. Biol.* **18**, 1130 (2011).
27. M. Cosner, R. Jansen, B. Moret, L. Raubeson, L. Wang, T. Warnow and S. Wyman, An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae, in *Comparative Genomics*, eds. D. Sankoff and J. Nadeau (Kluwer Academic Publishers, Dordrecht, NL, 2000) pp. 99–122.
28. D. Robinson and L. Foulds, *Mathematical Biosciences* **53**, 131 (1981).
29. F. Chen, A. Mackey, J. Vermunt and D. Roos, *PLoS ONE* **2**, p. e383 (2007).
30. I. Letunic and P. Bork, *Nucl. Acids Res.* **39**, W475 (2011).
31. C. Ponting, *Nat. Rev. Genet.* **9**, 689 (2008).
32. M. Srivastava *et al.*, *Nature* **454**, 955 (2008).
33. E. Negrisolo, H. Kuhl, C. Forcato, N. Vitulo, R. Reinhardt, T. Patarnello and L. Bargelloni, *Mol. Biol. Evol.* **27**, 2757 (2010).
34. D. Fitzpatrick, M. Logue, J. Stajich and G. Butler, *BMC Evolutionary Biology* **6**, p. 99 (2006).
35. H. Wang, Z. Xu, L. Gao and B. Hao, *BMC Evolutionary Biology* **9**, p. 195 (2009).
36. H. Nozaki, M. Matsuzaki, M. Takahara, O. Misumi, H. Kuroiwa, M. Hasegawa, T. Shin-i, Y. Kohara, N. Ogasawara and T. Kuroiwa, *J. Mol. Evol.* **56**, 485.
37. F. Burki, K. Shalchian-Tabrizi, M. Minge, A. Skjveland, S. Nikolaev, K. Jakobsen and J. Pawlowski, *PLoS ONE* **2**, p. e790 (2007).
38. U. Hellsten *et al.*, *Science* **328**, 633 (2010).

AN ANALYTICAL COMPARISON OF MULTILOCUS METHODS UNDER THE MULTISPECIES COALESCENT: THE THREE-TAXON CASE

SEBASTIEN ROCH

Department of Mathematics
University of Wisconsin–Madison
Madison, WI
roch@math.wisc.edu

Incomplete lineage sorting (ILS) is a common source of gene tree incongruence in multilocus analyses. Numerous approaches have been developed to infer species trees in the presence of ILS. Here we provide a mathematical analysis of several coalescent-based methods. The analysis is performed on a three-taxon species tree and assumes that the gene trees are correctly reconstructed along with their branch lengths. It suggests that maximum likelihood (and some equivalents) can be significantly more accurate in this setting than other methods, especially as ILS gets more pronounced.

Keywords: Multispecies coalescent, incomplete lineage sorting, gene tree/species tree.

1. Introduction

Incomplete lineage sorting (ILS) is an important confounding factor in phylogenetic analyses based on multiple genes or loci.[1,2] ILS is a population-level phenomenon that is caused by the failure of two lineages to coalesce in a population, leading to the possibility that one of the lineages first coalesces with a lineage from a less closely related population. As a result, it can produce extensive gene tree incongruence that must be accounted for appropriately in multilocus analyses.[3]

A large number of methods have been developed to address this source of incongruence.[4] Several such methods rely on a statistical model of ILS known as the multispecies coalescent. In this model, populations are connected by a phylogeny. Independent coalescent processes are performed in each population and assembled to produce gene trees. Several methods have been shown to be statistically consistent under the multispecies coalescent, that is, they are guaranteed to return the correct species tree given enough loci.[5–8]

The performance and accuracy of coalescent-based multilocus methods have been the subject of numerous simulation studies.[5,9] In this paper, we complement such studies with a detailed analytical comparison in a tractable test case, a three-taxon species tree. We analyze 7 methods: maximum likelihood (ML),[10] GLASS/Maximum Tree (MT),[7,11] R^*,[12] STAR,[5] minimizing deep coalescences (MDC),[1] STEAC,[5] and shallowest coalescences (SC).[1] Under the assumption that gene trees are reconstructed without estimation error, we derive the exponential decay rate of the failure probability as the internal branch length of the species tree varies. The analysis, which relies on large-deviations theory, reveals that ML and GLASS/MT are more accurate in this setting than the other methods—especially in the regime where ILS is more common.

2. Materials and Methods

2.1. *Multispecies coalescent: Three-taxon case*

We first describe the statistical model under which our analysis is performed, the *multispecies coalescent*. We only discuss the three-taxon case. For more details, see Ref. 2 and references therein.

A weighted rooted tree is called ultrametric if each leaf is exactly at the same distance from the root. For a three-leaf ultrametric tree G with leaves a, b, and c, we denote by $ab|c$ the topology where a and b are closer to each other than to c, and similarly for $ac|b$, $bc|a$. The topology of G is denoted by $\mathcal{T}[G]$.

Let S be an ultrametric species phylogeny with three taxa. We assume that all haploid populations in S have population size N. We denote the current populations by A, B and C (which we identify with the leaves of S) and we assume that S has topology $\mathrm{AB}|\mathrm{C}$. The ancestral populations are AB (corresponding to the immediate ancestor to populations A and B) and ABC (corresponding to the ancestor of populations A, B and C). The corresponding divergence times (backwards in time from the present) are denoted by τ_{AB} and τ_{ABC} with the assumption $\tau_{\mathrm{AB}} \leq \tau_{\mathrm{ABC}}$. All times are given in units of N generations. For a population X, we let $\tau_{\mathrm{X}}^{\mathrm{P}}$ be the divergence time of the parent population of X. Let $\mathbb{X} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{AB}, \mathrm{ABC}\}$ be the set of all populations in S.

We consider L loci $\ell = 1, \ldots, L$ and, for each locus, we sample one lineage from each population at time 0. For locus ℓ, we denote by $I_{\mathrm{X}}^{(\ell)}$ the number of lineages entering population X and by $O_{\mathrm{X}}^{(\ell)}$ the number of lineages exiting population X (backwards in time), where necessarily $I_{\mathrm{X}}^{(\ell)} \geq O_{\mathrm{X}}^{(\ell)}$. Similarly, for $k = O_{\mathrm{X}}^{(\ell)} + 1, \ldots, I_{\mathrm{X}}^{(\ell)}$, the time of the coalescent event bringing the number of lineages from k to $k-1$ in population X and locus ℓ is $T_{\mathrm{X}}^{(\ell,k)}$. We denote by $G_1, \ldots, G_L$ the corresponding ultrametric gene trees (including both topology and branch lengths).

Then, under the multispecies coalescent, assuming the loci are unlinked, the likelihood of the gene trees is given by

$$f(G_1, \ldots, G_L|S) = \prod_{\ell=1}^{L} \exp\Bigg(- \sum_{\mathrm{X}\in\mathbb{X}} \Bigg\{ \binom{O_{\mathrm{X}}^{(\ell)}}{2} \left(\tau_{\mathrm{X}}^{\mathrm{P}} - T_{\mathrm{X}}^{(\ell, O_{\mathrm{X}}^{(\ell)}+1)}\right) - \sum_{k=O_{\mathrm{X}}^{(\ell)}+1}^{I_{\mathrm{X}}^{(\ell)}} \binom{k}{2} \left(T_{\mathrm{X}}^{(\ell,k+1)} - T_{\mathrm{X}}^{(\ell,k)}\right) \Bigg\}\Bigg), \tag{1}$$

where we let $T_{\mathrm{X}}^{(\ell, I_{\mathrm{X}}^{(\ell)}+1)} = \tau_{\mathrm{X}}$ for convenience.[13]

The parameter governing the extent of incomplete lineage sorting is the length of the internal branch of S

$$t = \tau_{\mathrm{ABC}} - \tau_{\mathrm{AB}}.$$

The probability that the lineages from A and B fail to coalesce in branch AB, an event we denote by $\textsc{Fail}_\ell$ for locus ℓ (and its complement by $\textsc{Success}_\ell$), is

$$1 - p = e^{-t}.$$

Note that, in that case, all three gene-tree topologies are equally likely. Of course, $1-p \to 1$ as $t \to 0$.

2.2. *Multilocus methods*

A basic goal of multilocus analyses is to reconstruct a species phylogeny (including possibly estimates of the divergence times) from a collection of gene trees. Here we assume that the data consist of L gene trees $G_1, \ldots, G_L$ corresponding to L unlinked loci generated under the multispecies coalescent. We assume further that the gene trees are ultrametric and that their topologies and branch lengths are estimated without error.

We consider several common multilocus methods. In our setting, several of these methods are in fact equivalent and we therefore group them below. Note further that we only consider statistically consistent methods, that is, methods that are guaranteed to converge on the right species phylogeny (at least its topology) as the number of loci L increases to $+\infty$ in the test case we described above. We briefly describe these methods. For more details, see e.g. Ref. 4 and references therein.

ML/GLASS/MT Under the multispecies coalescent, maximum likelihood (ML) selects the topology and divergence times that maximizes the likelihood (1). ML is implemented in the software package STEM.[10]

In the GLASS method,[7] the species phylogeny is reconstructed from a distance matrix in which the entries are the minimum gene coalescence times across loci. The equivalent Maximum Tree (MT) method was introduced and studied in Refs. 8,11,14.

A key result in Ref. 8 is that, in the constant-population case, the term inside the exponential in the likelihood (1) is monotonically decreasing in the divergence times. As a result, because GLASS and MT select the phylogeny with the largest possible divergence times, maximum likelihood is equivalent to GLASS and MT in this context. See Ref. 8 for details.

***R**/STAR/MDC** In the R^* consensus method,[6,12] for each three-taxon set (here, we only have one such set), we include the topology that appears in highest frequency among the loci (breaking ties uniformly at random) and we reconstruct the most resolved phylogeny that is compatible with these three-taxon topologies.

In the STAR method,[5] the species phylogeny is reconstructed from a distance matrix in which the entries are the average ranks of gene coalescence times across loci. Here the root has the highest rank and the rank decreases by one as one goes from the root to the leaves.

The minimizing deep coalescences (MDC) method[1,15] selects the species phylogeny that requires the smallest number of "extra lineages," that is, lineages that fail to coalesce in a branch of the species phylogeny (breaking ties uniformly at random).

On a three-taxon phylogeny, there are only three distinct rooted topologies. In each case, the most recent divergence is assigned rank 1 in STAR and the other divergence is assigned rank 2. Hence selecting the topology corresponding to the lowest average rank is equivalent to selecting the most common topology among all loci—which is what R^* does. A similar argument shows that MDC also selects the R^* consensus tree in our test case.

Other common topology-based methods fall in this class, for instance, Rooted Triple Consensus.[16]

STEAC/SC In the STEAC method,[5] the species phylogeny is reconstructed from a distance matrix in which the entries are the average coalescence times across loci. The shallowest coalescences (SC) method is similar to STEAC in that it uses average coalescence times. The difference between the two methods is in how they deal with multiple alleles per population. Since we only consider the single-allele case, the two methods are equivalent here.

2.3. *Large-deviations approach*

As mentioned above, we consider estimation methods that are statistically consistent in the sense that they are guaranteed to converge on the correct species phylogeny as the number of loci L increases to $+\infty$. To compare different methods, we derive the rate of exponential decay of the probability of failure. Let S be a species phylogeny with internal branch length t and assume that $G_1, \ldots, G_L$ are unlinked gene trees generated under the multispecies coalescent. As we explain next, large-deviations theory (see e.g. Ref. 17) allows us to compute the (exponential) decay rate

$$\alpha_{\mathbb{M}}(t) = -\lim_{L \to +\infty} \frac{1}{L} \ln \mathbb{P}[\text{Method } \mathbb{M} \text{ fails given } L \text{ loci from } S],$$

that is, roughly

$$\mathbb{P}[\text{Method } \mathbb{M} \text{ fails given } L \text{ loci from } S] \approx e^{-L\alpha_{\mathbb{M}}(t)},$$

for large L. As the notation indicates, the key parameter that influences the decay rate is the length of the internal branch t of the species phylogeny. In particular, we expect that $\alpha_{\mathbb{M}}(t)$ is increasing in t as a larger t makes the reconstruction problem easier.

To derive $\alpha_{\mathbb{M}}(t)$, we first need to express the probability of failure as a large deviation event of the form

$$\mathbb{P}[\text{Method } \mathbb{M} \text{ fails given } L \text{ loci from } S] = \mathbb{P}\left[\sum_{\ell=1}^{L} Y_\ell > yL\right],$$

where y is a constant and $\{Y_\ell\}_{\ell=1}^{L}$ are independent identically distributed random variables. The particular choice of random variables depends on the method, as we describe below. Let

$$\phi(s) = \mathbb{E}[e^{sY_\ell}],$$

be the moment-generating function of Y_ℓ (which does not depend on ℓ by assumption). Then large-deviations theory stipulates (see e.g. Theorem 2.6.3 in Ref. 17) that the decay rate is given by

$$\alpha_{\mathbb{M}}(t) = ys_* - \ln \phi(s_*), \tag{2}$$

where $s_* > 0$ is the solution (if it exists) to

$$\frac{\phi'(s_*)}{\phi(s_*)} = y,$$

provided there is an $s > 0$ such that $\phi(s) < +\infty$, $y > \mathbb{E}[Y_\ell]$ and Y_ℓ is not a point mass at $\mathbb{E}[Y_\ell]$.

3. Results: Derivations of decay rates

3.1. *A domination result*

We first argue that, given perfectly reconstructed unlinked gene trees under the multispecies coalescent, ML/GLASS/MT always has a greater probability of success than R^*/STAR/MDC and STEAC/SC—or, in fact, any other method. Indeed note that the probability of success can be divided into two cases:

(1) The case where $\textsc{Success}_\ell$ occurs for at least one locus ℓ, an event of probability $(1-(1-p)^L)$. In that case, ML/GLASS/MT necessarily succeeds whereas the other two methods succeed with probability < 1.
(2) The case where $\textsc{Fail}_\ell$ occurs for all loci ℓ, an event of probability $(1-p)^L$. In that case, all methods succeed with probability $1/3$ by symmetry. For instance, for ML/GLASS/MT, any pair of populations is equally likely to lead to the smallest inter-species distance. A similar argument applies to the other two methods.

Hence, overall ML/GLASS/MT succeeds with greater probability.

3.2. *Decay rates*

We derive the decay rates for the methods above. The results are plotted in Figure 1. The asymptotic regimes are highlighted in Figures 2 and 3. For lack of space, all proofs can be found in Ref. 18.

ML/GLASS/MT In this case, the decay rate can be derived directly without using (2). Following the derivation in Ref. 7 (see also Ref. 8 for a similar argument), ML/GLASS/MT fails with probability

$$1-\left[(1-(1-p)^L)+\frac{1}{3}(1-p)^L\right]=\frac{2}{3}(1-p)^L=\frac{2}{3}e^{-tL}.$$

Then we get the following:

Claim 3.1 (ML/GLASS/MT). *The decay rate of ML/GLASS/MT on S is*

$$\alpha_{\mathrm{ML}}(t)=t.$$

R*/STAR/MDC** For a locus ℓ, we let $Z_{\mathrm{AB}}^{(\ell)}$ be 1 if $\textsc{Fail}_\ell$ occurs and $\mathcal{T}[G_\ell]=\mathrm{AB}|\mathrm{C}$, and 0 otherwise (where recall that $\mathcal{T}[G_\ell]$ is the topology of G_ℓ). We let

$$\mathcal{Z}_{\mathrm{AB}}=\sum_{\ell=1}^{L}Z_{\mathrm{AB}}^{(\ell)}.$$

Similarly, we define $Z_{\mathrm{AC}}^{(\ell)}$, $Z_{\mathrm{BC}}^{(\ell)}$, $\mathcal{Z}_{\mathrm{AC}}$ and $\mathcal{Z}_{\mathrm{BC}}$. Then R^*/STAR/MDC fails if

$$\mathcal{Z}_{\mathrm{AB}}+(L-\mathcal{Z}_{\mathrm{AC}}-\mathcal{Z}_{\mathrm{BC}}-\mathcal{Z}_{\mathrm{AB}})<\max\{\mathcal{Z}_{\mathrm{AC}},\mathcal{Z}_{\mathrm{BC}}\}.$$

It can be shown that

$$\alpha_{\mathrm{R}^*}(t)=-\lim_{L\to+\infty}\frac{1}{L}\ln\mathbb{P}[2\mathcal{Z}_{\mathrm{AC}}+\mathcal{Z}_{\mathrm{BC}}>L].$$

Then we get the following:

Claim 3.2 (R^*/STAR/MDC). *The decay rate of R^*/STAR/MDC on S is*

$$\alpha_{\mathrm{R}^*}(t)=-\ln\left(2\sqrt{\frac{1}{3}e^{-t}\left(1-\frac{2}{3}e^{-t}\right)}+\frac{1}{3}e^{-t}\right).$$

As $t\to 0$,

$$\alpha_{\mathrm{R}^*}(t)=\frac{3}{4}t^2+O(t^3),$$

and, as $t \to +\infty$*,*

$$\alpha_{\mathrm{R}^*}(t) \approx \frac{t}{2} - \frac{1}{2}\ln\frac{4}{3}.$$

STEAC/SC For a locus ℓ, we let $D_{\mathrm{AB}}^{(\ell)}$ be the time to the most recent common ancestor of A and B in G_ℓ (in unit of N generations). We let

$$\mathcal{D}_{\mathrm{AB}} = \sum_{\ell=1}^{L} D_{\mathrm{AB}}^{(\ell)}.$$

Similarly, we define $D_{\mathrm{AC}}^{(\ell)}$, $D_{\mathrm{BC}}^{(\ell)}$, $\mathcal{D}_{\mathrm{AC}}$ and $\mathcal{D}_{\mathrm{BC}}$. Then STEAC/SC fails if

$$\mathcal{D}_{\mathrm{AB}} > \min\{\mathcal{D}_{\mathrm{AC}}, \mathcal{D}_{\mathrm{BC}}\}.$$

It can be shown that

$$\alpha_{\mathrm{STEAC}}(t) = \lim_{L\to+\infty} -\frac{1}{L}\ln\mathbb{P}[\mathcal{D}_{\mathrm{AB}} - \mathcal{D}_{\mathrm{AC}} > 0].$$

Then we get the following:

Claim 3.3 (STEAC/SC). *The decay rate of STEAC/SC on* S *is*

$$\alpha_{\mathrm{STEAC}}(t) = -\ln\left(\frac{3e^{-s_* t} - s_*^2 e^{-t}}{3(1-s_*^2)}\right),$$

where $0 < s_* < 1$ *is the unique solution to the fixed-point equation*

$$s_* = \frac{1}{2}[6s_* - 3t(1-s_*^2)]e^{(1-s_*)t}.$$

Further, as $t \to 0$*,*

$$\alpha_{\mathrm{STEAC}}(t) = \frac{3}{8}t^2 + O(t^3),$$

and, as $t \to +\infty$*,*

$$\alpha_{\mathrm{STEAC}}(t) \approx t - \ln t - 0.1656.$$

4. Discussion

As can be seen from Figures 1 and 3 as well as from the asymptotics, ML/GLASS/MT does indeed give a larger decay rate for all t. In fact, the decay rate of ML/GLASS/MT is significantly higher, especially as $t \to 0$ that is, under high levels of incomplete lineage sorting. For instance, to be concrete, if $L = 500$ loci and $t = 0.1$ (in units of N generations), the probability of failure is approximately: 1.9×10^{-22} for ML/GLASS/MT; 0.038 for R^*/STAR/MDC; 0.16 for STEAC/SC. Intuitively, this difference in behavior arises from the fact that ML/GLASS/MT requires only *one* successful locus, whereas R^*/STAR/MDC and STEAC/SC rely on an *average* over all loci.

Comparing R^*/STAR/MDC and STEAC/SCin Figures 1, 2 and 3, note that $\alpha_{\mathrm{R}^*}(t)$ is higher than $\alpha_{\mathrm{STEAC}}(t)$ for small t but that the situation is reversed for large t. In fact, in the limit $t \to +\infty$, $\alpha_{\mathrm{STEAC}}(t)$ grows at roughly the same rate as $\alpha_{\mathrm{ML}}(t)$ (which is optimal by the domination result). At large t, STEAC/SC has somewhat of an advantage in that the expectation gap in the failure event increases linearly with t, whereas it saturates under R^*/STAR/MDC.

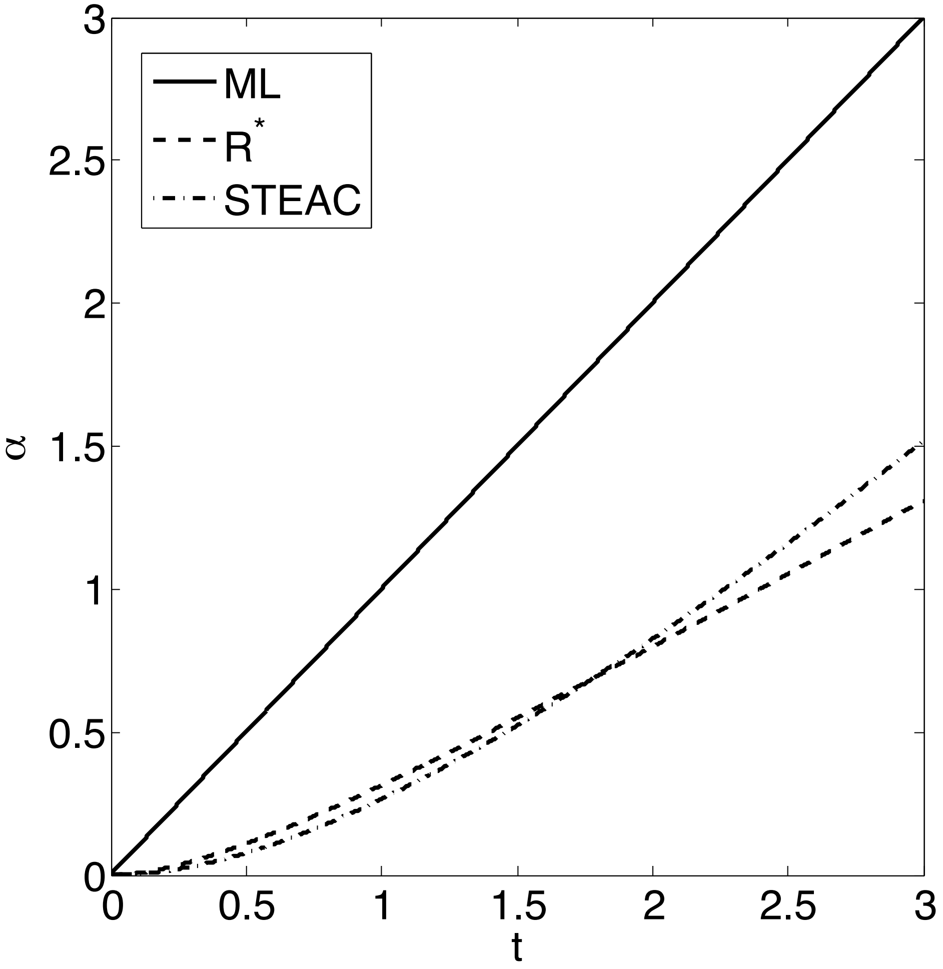


Fig. 1. Decay rates.

The analysis described here ignores several features that influence the accuracy of species tree reconstruction. Notably we have assumed that gene trees, including their branch lengths, are reconstructed without error. On real sequence datasets, the uncertainty arising from gene-tree estimation plays an important role. For instance, although GLASS/MT achieves the optimal decay rate in our setting, these methods are in fact sensitive to sequence noise because they rely on the computation of a minimum over loci—the very feature that leads to their superior performance here. See Ref. 5 for simulation results. Extending our analysis to incorporate gene tree estimation error is an important open problem which should help in the design of multilocus methods. It is important to note that, under appropriate modeling of sequence data, ML is *not* in general equivalent to GLASS/MT and comparing the sensitivity of these methods to estimation error is an interesting problem.

Other extensions deserve further study. Often many alleles are sampled from each population. Note that the benefit of multiple alleles is known to saturate as the number of alleles increases.[19] This is because the probability of observing any number of alleles at the top of a branch is uniformly bounded in the number alleles existing at the bottom.

Further, the molecular clock assumption, although it may be a reasonable first approximation in

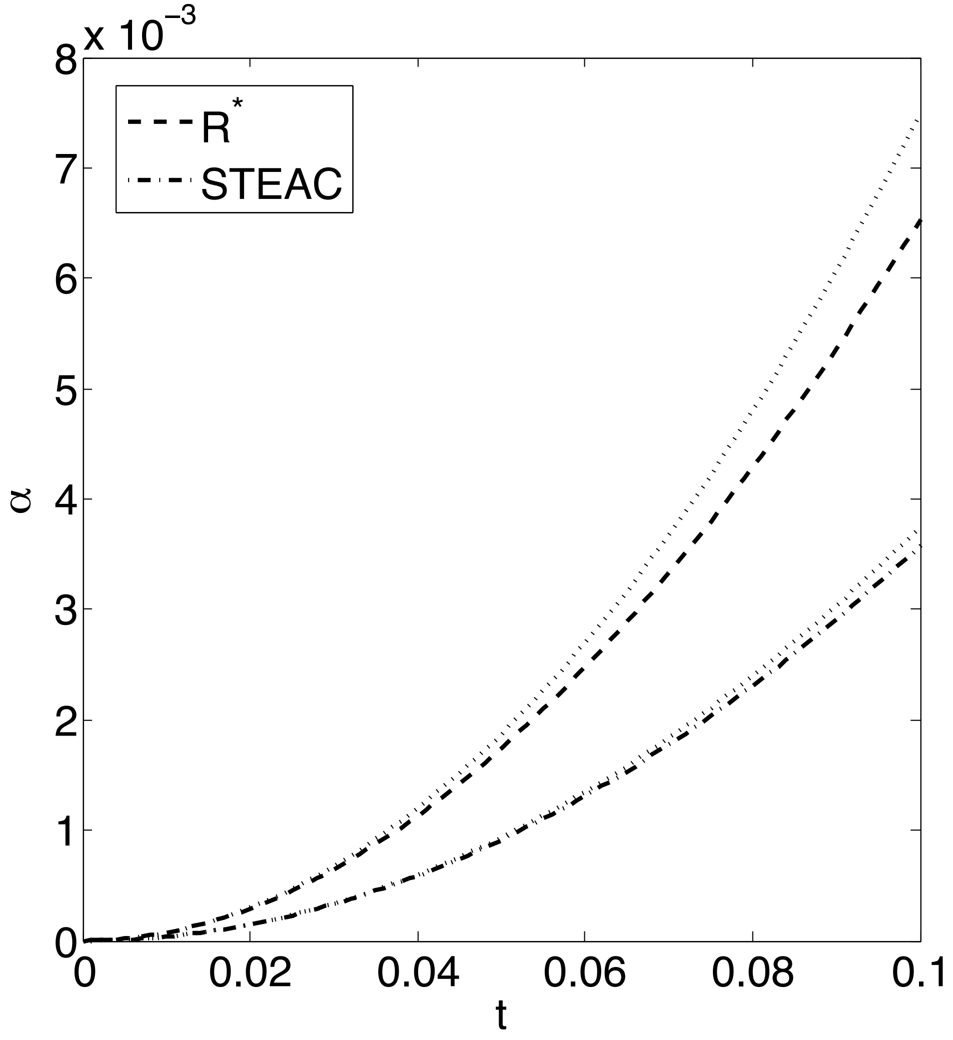


Fig. 2. Decay rates as $t \to 0$. The dotted lines indicate the respective predicted asymptotics. The decay rate for ML is not shown as it would be almost vertical.

the context of recently diverged populations, should not be necessary for our analysis. One should also consider larger numbers of taxa, varying population sizes, etc.

Simulation studies may provide further insight into these issues. However an analytical approach, such as the one we have used here, is valuable in that it allows the study of an entire class of models in one analysis. It can also provide useful, explicit predictions to guide the design of reconstruction procedures.

5. Supplementary Material

For lack of space, all proofs can be found in Ref. 18.

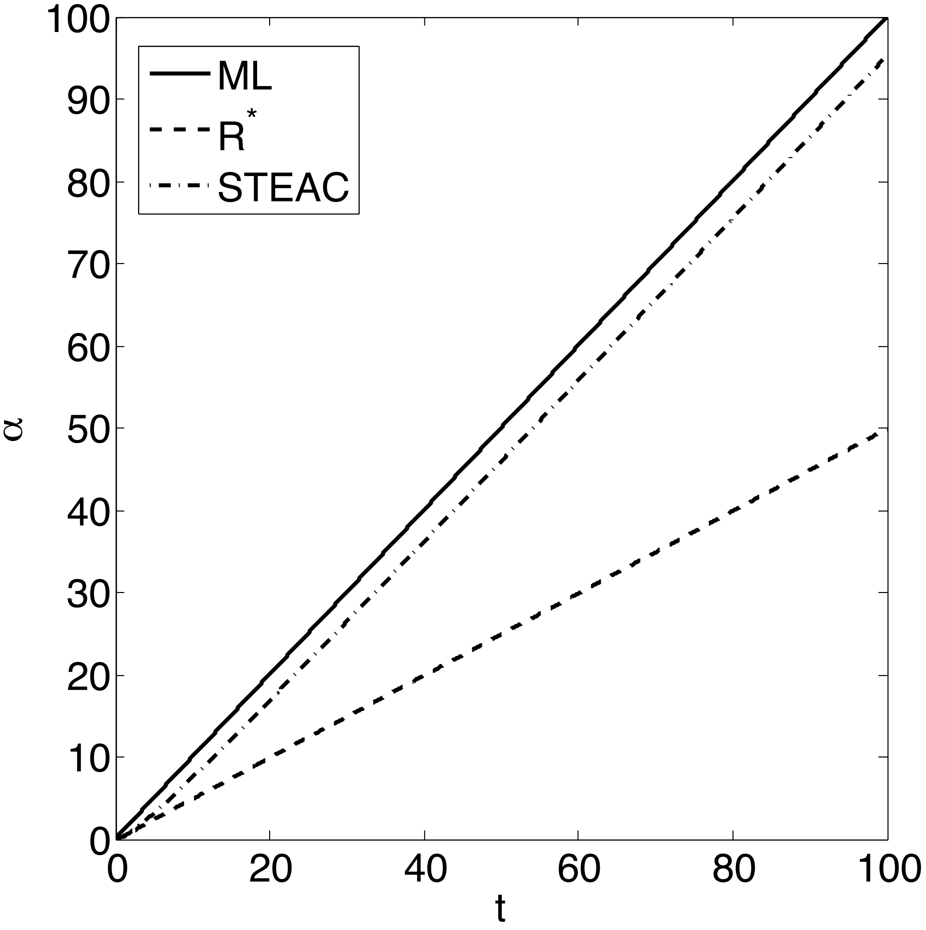


Fig. 3. Decay rates as $t \to +\infty$.

6. Acknowledgments

This work was supported by NSF grant DMS-1007144, NSF grant DMS-1149312 (CAREER), and an Alfred P. Sloan Research Fellowship. Part of this work was performed while the author was visiting the Institute for Pure and Applied Mathematics (IPAM) at UCLA.

References

1. W. P. Maddison, *Systematic Biology* **46**, 523 (1997).
2. J. H. Degnan and N. A. Rosenberg, *Trends in ecology and evolution* **24**, 332 (2009).
3. J. H. Degnan and N. A. Rosenberg, *PLoS Genetics* **2** (May 2006).
4. L. Liu, L. Yu, L. Kubatko, D. K. Pearl and S. V. Edwards, *Molecular Phylogenetics and Evolution* **53**, 320 (2009).
5. L. Liu, L. Yu, D. K. Pearl and S. V. Edwards, *Systematic Biology* **58**, 468 (2009).
6. J. H. Degnan, M. DeGiorgio, D. Bryant and N. A. Rosenberg, *Systematic Biology* **58**, 35 (2009).
7. E. Mossel and S. Roch, *IEEE/ACM Trans. Comput. Biology Bioinform.* **7**, 166 (2010).
8. L. Liu, L. Yu and D. Pearl, *Journal of Mathematical Biology* **60**, 95 (2010), 10.1007/s00285-009-0260-0.
9. A. D. Leaché and B. Rannala, *Systematic Biology* **60**, 126 (2011).
10. L. S. Kubatko, B. C. Carstens and L. L. Knowles, *Bioinformatics* **25**, 971 (2009).

11. L. Liu and D. K. Pearl, *Systematic Biology* **56**, 504 (2007).
12. D. Bryant, A classification of consensus methods for phylogenetics, in *Bioconsensus (Piscataway, NJ, 2000/2001)*, , DIMACS Ser. Discrete Math. Theoret. Comput. Sci. Vol. 61 (Amer. Math. Soc., Providence, RI, 2003) pp. 163–183.
13. B. Rannala and Z. Yang, *Genetics* **164**, 1645 (2003).
14. S. V. Edwards, L. Liu and D. K. Pearl, *Proceedings of the National Academy of Sciences* **104**, 5936 (2007).
15. C. Than and L. Nakhleh, *PLoS Comput Biol* **5**, p. e1000501 (09 2009).
16. G. Ewing, I. Ebersberger, H. Schmidt and A. von Haeseler, *BMC Evolutionary Biology* **8**, p. 118 (2008).
17. R. Durrett, *Probability: theory and examples*Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge Series in Statistical and Probabilistic Mathematics, fourth edn. (Cambridge University Press, Cambridge, 2010).
18. S. Roch, An analytical comparison of coalescent-based multilocus methods: The three-taxon case, Supplementary material available at http://arxiv.org/abs/1207.4074.
19. N. A. Rosenberg, *Theor. Popul. Biol.* **61**, 225 (March 2002).

POST-NGS: INTERPRETATION AND ANALYSIS OF NEXT GENERATION SEQUENCING DATA FOR BASIC AND TRANSLATIONAL SCIENCE

GURKAN BEBEK

Case Center for Proteomics and Bioinformatics,
Case Western Reserve University 10900 Euclid Ave.
Cleveland, OH 44106-4988, USA
Email: gurkan@case.edu

MEHMET KOYUTÜRK

Department of Electrical Engineering and Computer Science,
Case Western Reserve University 10900 Euclid Ave.
Cleveland, OH 44106-4988, USA
Email: koyuturk@eecs.case.edu

THOMAS LAFRAMBOISE

Department of Genetics,
Case Western Reserve University 10900 Euclid Ave.
Cleveland, OH 44106-4955, USA
Email: thomas.laframboise@case.edu

BENJAMIN J. RAPHAEL

Department of Computer Science,
Brown University, 115 Waterman St.
Providence, RI 20912, USA
Email: braphael@cs.brown.edu

MARK R. CHANCE

Case Center for Proteomics and Bioinformatics,
Case Western Reserve University 10900 Euclid Ave.
Cleveland, OH 44106-4988, USA
Email: mark.chance@case.edu

1. Introduction

Next generation sequencing has dramatically changed our view of what is achievable in genomics. In recent years, research has focused on using next generation sequencing data to characterize genomic content and many methods have been developed for de novo sequence assembly, identification of genomic variants, detection of splice variants etc. Now that the scientific community is equipped with efficient and reliable methods to characterize genomic content, it is natural to expect that the vast amount of information generated by these methods will be further analyzed to seek answers to fundamental biological and medical questions within the context of biological systems. Such questions range from the relationship between genotype and phenotype to regulatory mechanisms of development and principles of evolution.

Recently, large-scale projects such as the Cancer Genome Atlas (TCGA) Project or the 1000 Genomes Project have utilized these technologies extensively, driving remarkable conclusions. However, the methods that have been utilized therein have been specific to these applications and limited in number. Further downstream analyses of NGS data require development of new computational techniques to derive biological knowledge from this vast pool of information. Importantly, there is great need for new methods for integrating these large datasets within the current and emerging research paradigms. From basic science to clinical applications, the –omes that are identified can steer research efforts in transformative directions.

This session will provide a forum for methods and algorithms developed for analysis of finalized next-generation sequencing data. Motivated by the flourishing availability of genome sequences and related data, novel computational methods that interpret these data for research and clinical applications are included in this session. Developing innovative new methodologies and tools for analyzing post-NGS datasets will stimulate more basic and clinical investigation. The aim of this session is to raise awareness of these challenges in the biocomputing community and provide a forum for discussing and disseminating a broad range of computational methods that aim to construct this leap.

2. Session Summary

This session includes an invited talk, three reviewed papers contributed as oral presentations, two contributed papers as posters and a tutorial prepared by the session chairs. The studies presented in this session focus on the development of computational methods to utilize next generation sequencing data to further study diverse biological and translational science problems.

2.1. *Accepted Session Papers*

The following talks will be presented at the Post-NGS session:

- "ChIPModule: Systematic discovery of transcription factors and their cofactors from ChIP-seq data" by Jun Ding, Xiaohui Cai, Ying Wang, Haiyan Hu, and Xiaoman Li.
- "SHPlace: Fast phylogenetic placement using locality-sensitive hashing" by Daniel G. Brown and Jakub Truszkowski.
- "Detecting highly differentiated copy-number variants from pooled population sequencing" by Daniel R. Schrider, David J. Begun, and Matthew W. Hahn.

The following papers will be presented as posters at the symposium:

- "MetaSeq: Privacy preserving meta-analysis of sequencing-based association studies" by Angad Pal Singh, Samreen Zafer, and Itsik Pe'er.
- "Using BioBin to explore rare variant population stratification" by Carrie B. Moore, John R. Wallace, Alex T. Frase, Sarah A. Pendergrass, and Marylyn D. Ritchie

The breadth of research presented in the Post-NGS Session excites us, and we are hopeful that our session will help bring together researchers from various fields and lead to fruitful discussions.

3. Acknowledgments

We would like to thank all the anonymous reviewers who contributed with their invaluable time, and expertise to evaluate submitted papers, and the authors who submitted their work to the Post-NGS Session. This publication was made possible by the Clinical and Translational Science Collaborative of Cleveland, UL1TR000439 from the National Center for Advancing Translational Sciences (NCATS) component of the National Institutes of Health (NIH) and NIH roadmap for Medical Research. MRC, GB and MK would like to mention support from NIH Grant P30-CA043703. MK and MRC are supported in part by NIH grant R01LM011247. TL is supported by NIH Grant R01- CA131341 and American Cancer Society Grant RSG-12-159-01-DMC. MK and TL are supported in part by National Science Foundation (NSF) Award IIS-0916102. BJR is supported by the National Institutes of Health Grant R01-HG5690, a CAREER Award from NSF (CCF-1053753), and Burroughs Wellcome Career Award at the Scientific Interface.

LSHPlace: Fast phylogenetic placement using locality-sensitive hashing

DANIEL G. BROWN* and JAKUB TRUSZKOWSKI

David R. Cheriton School of Computer Science
University of Waterloo
Waterloo ON N2L 3G1 Canada
browndg,jmtruszk@uwaterloo.ca

We consider the problem of phylogenetic placement, in which large numbers of sequences (often next-generation sequencing reads) are placed onto an existing phylogenetic tree. We adapt our recent work on phylogenetic tree inference, which uses ancestral sequence reconstruction and locality-sensitive hashing, to this domain. With these ideas, new sequences can be placed onto trees with high fidelity in strikingly fast runtimes. Our results are two orders of magnitude faster than existing programs for this domain, and show a modest accuracy tradeoff. Our results offer the possibility of analyzing many more reads in a next-generation sequencing project than is currently possible.

Keywords: Phylogenetic placement; Nearest neighbour search;

1. Introduction

In the past few years, advances in sequencing technology have enabled the study of microbial communities from diverse environments, such as soil,[1] ocean,[2] and the human body.[3] Microbial ecologists are interested in the diversity of bacteria in a given environment, their evolutionary origins, and their metabolic relationships. They answer these questions by collecting sequence data from environmental samples and then comparing them to reference sequences from known microbial lineages.

Phylogenetics provides a natural framework for investigating the microbial diversity in these environments. Most microorganisms can be approximately located on the tree of life for bacteria, and then communities or environments can be characterized by the relative abundance of certain taxa.[4] Or, unusual sequences can be a focus for further investigation and directed sequencing.[5] The first step, however, is to locate each sequence on the tree.

While many phylogenetic algorithms have been developed over the years, the current flood of sequence data from next-generation sequencing presents new challenges for traditional methods. The massive amounts of data generated by NGS make traditional phylogenetic inference computationally prohibitive; indeed, in metagenomic contexts, a common first step is to cluster a data set, possibly consisting of millions of reads and instead analyze just the hundreds or thousands of cluster centres, which discards much valuable data.[4] Another problem is that environmental sequencing produces short sequence reads, instead of full gene sequences. This is partly due to inherent difficulties of assembling reads in the presence of sequences from many different species. For example, reads generated by Illumina have length $\approx$ 200 bp, which does not provide sufficiently strong phylogenetic signal for full phylogenetic inference. Maximum likelihood phylogenies from incomplete sequences tend to be biased towards grouping highly overlapping sequences together.[6]

These problems have recently motivated researchers to focus on placing individual environmental sequences into a fixed phylogeny, instead of performing full phylogenetic inference on the entire set of sequences. This has several advantages. The computational cost is greatly reduced, as the number of topologies that need to be considered is linear in the size of the tree for each sequence. By considering each query sequence separately, we also hope to avoid the biases associated with sequence overlaps. There are currently several programs performing this task, called *phylogenetic placement* .[6–8] Unfortunately, their speed is insufficient to place millions of reads, as we see in Section 4.3.

Recently, we have developed a fast and accurate phylogenetic reconstruction algorithm.[9] Our algorithm uses hash tables to identify closely related sequences without having to estimate all $\binom{n}{2}$ distances between sequences. Here, we adapt our technique to phylogenetic placement. More specifically, we develop the first algorithm that places sequences in a known reference phylogeny with running time sub-linear in the number of taxa in the reference tree. We show that our methods are both theoretically grounded and practically useful.

Our algorithm is based on several ideas. First, we use a hash table technique known as locality-sensitive hashing[10] to find a sequence in the tree that is close to the query sequence, in sublinear time. To make sure that such a sequence exists in the tree, we infer ancestral sequences at internal nodes of the reference tree. Finally, we use local search to find the optimal placement of the query sequence in a neighbourhood of the tentative placement discovered by locality-sensitive hashing. The running time of this procedure is determined by the number of hash tables needed to find a close enough sequence in the tree, which in turn depends on the lengths of the edges of the tree. Specifically, if p is an upper bound on the mutation probability on any edge, and $p < 1/2 - \sqrt{1/8}$, then we show that we can do the locality-sensitive hashing, which is the runtime-determining step, in $O(n^{\gamma(p)} \log^2 n)$ time for each sequence to place, where $\gamma(p)$ is always less than 1; the overall runtime is thus $O(mn^{\gamma(p)} \log^2 n)$ to place m new sequences onto a reference tree of n taxa. In practice, we choose to use a constant number of hash tables, reducing this phase to $O(\log n)$ time, though there is some runtime required for local search.

A novel algorithmic idea in this paper is to build the hash tables from slowly-evolving alignment columns. This reduces the probability of a mutation having occurred between two sequences in one of the hash table columns, which causes the algorithm to require fewer hash tables to guarantee a hash table collision.

We evaluate our algorithm on a number of synthetic and real data sets. The accuracy of our algorithm is lower than that of pplacer,[6] while its running time is around 2 orders of magnitude faster, making it a useful tool for handling large data sets. The current implementation of the local search phase of our method is distance based, and we expect that its accuracy could be substantially improved using maximum likelihood, at modest cost in running time.

2. Related work

Many tools determine the taxonomic origin of environmental sequences. These tools can be roughly divided into three categories: those based on phylogenetic modelling, those based on sequence composition, and those based on homology search. Gerlach[11] provides a recent survey of these methods.

Recently, researchers have developed several tools for placing environmental sequences onto a reference phylogeny. These methods generally take $O(n)$ time to insert a sequence in a tree of n taxa. MLTreeMap[8] was the first tool designed for this task. the Evolutionary Placement Algorithm[7] and pplacer[6] offer more efficient implementations of the same approach, which places a sequence optimally at each edge of the phylogeny, and then assigns the overall maximum likelihood placement as the answer for each individual sequence.

Some software pipelines for analyzing metagenomic data sets employ full phylogenetic reconstruction. These include TreePhyler[12] and CARMA.[13] While much progress has been made in fast phylogeny reconstruction in recent years,[9,14,15] reconstructing the full tree remains much slower than other classification methods.

Another approach is to build a classifier to discriminate between taxonomic groups at different levels of the Linnaean hierarchy. These classifiers do not attempt to model phylogenetic relationships between different taxa, but instead treat the problem as a supervised classification problem at each level of the hierarchy. The features used are usually derived from the k-mer composition of the sequence, which bypasses the need for aligning sequences. Many such classifiers have been developed, including PhyloPythia,[16] TACOA,[17] and PhyMM.[18] Taxy[19] uses mixture models and k-mers to estimate the relative abundance of different taxonomic groups in a set of sequences, without attempting to classify each sequence in detail. Taxonomic classifiers are often faster than phylogeny-based methods, but they do not offer the same explanatory power. Moreover, classifiers based on k-mers tend to behave badly on short sequences, as they lack sufficient information to distinguish different clades.

Yet another class of approaches uses BLAST[20] to determine evolutionary proximity of sequence reads to known species. Here, environmental sequences are expected to generate BLAST hits with the sequences they are closely related to. Several algorithms exist for mapping sets of BLAST hits to taxonomic classifications, including MEGAN[21] and SOrt-ITEMS.[22] Unfortunately, if the only close relatives to a sequence in a tree are internal nodes, this approach will fail; our method will avoid this problem due to our inference of internal sequences. The BLAST-based approach is also very fragile to short reads.

3. The algorithm

3.1. *Overview*

The input to the algorithm consists of three parts: the reference phylogeny T, on n taxa, which includes tree edge lengths; the multiple alignment A of the n sequences in the reference phylogeny; and a set of m query sequences, each aligned to that reference alignment. Our goal is to assign each query sequence to the edge where it joins the tree.

Our algorithm consists of the following steps. The first three are a preprocessing phase, and the resultant data structures could be stored for use, if a given tree and multiple alignment are going to be used to analyze many different sets of reads.

(1) Estimate evolutionary rates for each column of the reference alignment.
(2) Reconstruct ancestral sequences at each internal node of the reference phylogeny using maximum likelihood.
(3) Build a collection of hash tables, keyed on slowly-evolving columns of A. Add the keys for both leaf sequences and ancestral inferences into the hash tables.

(4) For each query sequence y:

   (a) Look for collisions between y and the hash tables. Choose the closest sequence x colliding with y.
   (b) Examine the neighbourhood of x in T. For each edge e in that neighbourhood, estimate the distance between e and y. Output the closest edge to y.

In what follows, we explain the details and the motivation behind each of these steps. For a more detailed discussion of the theoretical issues in this method, the reader is referred to our previous paper on phylogeny reconstruction.[9]

3.2. *Locality-sensitive hashing*

For a given query sequence y, we want to find sequences in T whose distance to y is small. We use a classical result by Indyk and Motwani[10] who solve a similar problem using a collection of randomized hash tables. We design hash tables so that the probability of y colliding with a similar sequence is high, and the probability of colliding with a distant sequence is low. We independently construct many such hash tables so that the probability y collides with a close sequence in at least one hash table is high. *Locality-sensitive hashing* has been applied to many problems in bioinformatics, such as motif finding.[23]

Specifically, Indyk and Motwani solve a related problem, the (r_1, r_2)-approximate Point Location in Equal Balls ((r_1, r_2)-PLEB):

Input: A set of n sequences P in $\{0,1\}^k$, a query sequence q, and radii $r_1 < r_2$

Output: Does there exist a sequence $p \in P$ within Hamming distance $r_1 k$ from q? If so, output "yes" and a sequence within $r_2 k$ of q. If there is no sequence in P within Hamming distance $r_2 k$ from q, output "no". Otherwise, output either "yes" or "no".

Indyk and Motwani's solution constructs n^{r_1/r_2} hash tables, each keyed on $c \log n$ randomly chosen sequence positions, where c depends only on r_2. Given y, a point within distance r_1 of it has probability at least n^{-r_1/r_2} of colliding with y in each hash table, while points further than r_2 from q have $O(1/n)$ probability of colliding. After inspecting a constant number of collisions with q, we can find, with constant probability, a point whose distance from q is at most r_2; if we boost by running $O(\log n)$ times independently, the success probability is $1 - n^{-\alpha}$, for any choice of α. Finding an (r_1, r_2)-approximate near neighbour for a query point q with high probability takes $O(n^{r_1/r_2} \log n)$ hash table lookups, each on a key of length $O(\log n)$ bits.

For large trees, parameter r_2 can be very close to the expected Hamming distance of unrelated sequences.[9] For binary characters, r_2 converges to $\frac{1}{2}$; similar constants can be computed for other mutation models, alphabets, and baseline letter frequencies.

Finding all neighbours within r_1 normalized Hamming distance of a sequence y thus takes $O(n^{2r_1+\epsilon} \log^2 n)$ time with high probability. The additive constant ϵ, which converges to 0 as n grows, accounts for error in distance estimates and that r_2 converges to $1/2$ as n grows. We use $O(n^{2r_1+\epsilon} \log n)$ hash tables, each of which requires $O(\log n)$ time to examine, and we take $O(\log^2 n)$ time examining the hash table hits.

3.3. *Ancestral states*

For locality-sensitive hashing to work, we have to ensure that our hash tables contain sequences sufficiently similar to the query sequence. If the query sequence branches out from an edge in the reference phylogeny that is not adjacent to a leaf, its distance from any leaf sequence may be too long for locality-sensitive hashing to work if we only index leaf sequences. We reconstruct ancestral sequences in the reference tree, using maximum likelihood, and add these reconstructions to the tables as well. The crucial question is how well the reconstructed ancestral sequences resemble the actual historical sequences: if a query is near an internal node in the true tree, and that node's sequence has been reconstructed well, it will likely result in a hash table hit, even with few hash tables.

There are several known upper bounds on the probability of incorrectly reconstructing an ancestral character from the leaf characters. This error probability depends on the edge lengths in the phylogeny. If most edges in T are very long, the number of errors in the reconstructed sequences will grow with the level of the internal node, and the reconstructed states at deep nodes of the tree will be junk, so adding them to the hash table will be pointless. On the other hand, if the edges are short enough, speciation outpaces mutation, and it is possible to reconstruct the ancestral state with guaranteed accuracy that does not depend on the size of the tree.

Evans *et al.*[24] have shown that for Cavender-Farris characters, if all the edge lengths correspond to mutation probabilities less than $\frac{1}{2} - \sqrt{\frac{1}{8}}$, then internal sequences will be reconstructed with accuracy at least a constant strictly above $\frac{1}{2}$. Similar results exist for more realistic evolutionary models, including the GTR model for DNA sequences.[25] Gascuel and Steel[26] established a similar result for trees generated from the birth-death process. This is an important complement to the result by Evans *et al.*, since birth-death trees will usually have a limited number of long edges. While the mathematical details of these results differ depending on the evolutionary model and branch length distribution, they all suggest the possibility of reasonably accurate ancestral sequence reconstruction for trees whose branches have moderate lengths.

The following is a useful bound by Steel.[27]

Theorem 3.1. *Let T be a phylogenetic tree where all mutation probabilities across edges are equal to $p_g < 1/8$. The probability p_{err} of incorrectly reconstructing the root state using Fitch parsimony is bounded by*

$$p_{err} < \frac{1}{2} - \frac{\sqrt{(1-4p_g)(1-8p_g)}}{2(1-2p_g)^2} < 1 - 4p_g$$

We have shown[9] that this bound also applies to trees with variable edge lengths if Felsenstein's maximum likelihood algorithm is used, rather than Fitch parsimony. Felsenstein's algorithm has optimal probability of correctly reconstructing ancestral states among all possible algorithms.

With the bound on p_{err}, we can determine the number of hash tables (and thus the runtime) required to find a node in the tree that is close to the true placement of the query sequence

y. Suppose sequences evolve according to the Cavender-Farris model. Let g_{err} be the distance corresponding to a mutation probability of p_{err}. If the evolutionary distance between y and the node that joins it to the tree is g_{query}, the effective evolutionary distance between the query and the nearest sequence in the tree is at most $g_{query} + g/2 + g_{err}$. This corresponds to a mutation probability of $\frac{1}{2} - \frac{1}{2}(1 - 2p_{query})(1 - 2p_g)^{1/2}(1 - 2p_{err})$. The number of hash tables required is thus bounded by

$$n^{1-(1-2p_{query})(1-2p_g)^{1/2}(1-2p_{err})}$$

Table 1 shows the running times of the hashing time for a single query, for different values of p_g, assuming for simplicity that $p_{query} \leq p_g$.

In practice, we do not know the upper bound on the distance of query sequences to the tree. The values in Table 1 should be understood only as illustration of the theoretical basis for our method. In the current version of the software, the number of hash tables per alignment region is fixed at 4, as this value gave good results for phylogenetic tree reconstruction, and maintains moderate memory use.

Table 1. Runtime of inserting a new taxon into a tree with n taxa, as a function of the maximum edge mutation probability

p_g	0.01	0.02	0.05	0.075	0.10
runtime	$n^{0.05}\log^2 n$	$n^{0.10}\log^2 n$	$n^{0.27}\log^2 n$	$n^{0.43}\log^2 n$	$n^{0.61}\log^2 n$

3.4. *Local search*

Our near-neighbour search procedure finds sequences that are in the vicinity of the correct placement for query sequence y, with high probability. However, it does not guarantee that the optimal placement is one of the edges adjacent to the node found. We estimate the distance between y and each edge $e = (x_1, x_2)$ near of the colliding sequence x as:

$$\hat{d}(y, e) = \frac{1}{2}(\hat{d}(x_1, y) + \hat{d}(x_2, y) - \hat{d}(x_1, x_2)),$$

estimating $\hat{d}(a, b)$ via the probabilistic model of evolution.

If x_1 and x_2 are reconstructed ancestral sequences and reconstruction errors at x_1 and x_2 are independent, then they will not bias the estimate $\hat{d}(y, e)$, as the additive terms in distance estimates associated with the reconstruction error will cancel out. We examine all edges within distance $\hat{d}(x, y)$ of node x. The edge e^* with smallest estimate is chosen as the placement of y, with pendant edge length $\hat{d}(x, e^*)$. If T has many very long and very short edges, this will prove slow; in practice this situation is quite rare. If we assume a minimum size for each edge, and a maximum size beyond which we consider $\hat{d}(x, y)$ too long to be meaningfully estimated, the neighborhood is of constant size.

For simplicity of implementation, the current implementation of the algorithm ignores possible dependencies between reconstruction errors. These dependencies could be ameliorated by using some of the techniques discussed in our previous paper.[9] We leave that as future work.

3.5. *Accommodating for different read locations*

In some cases, the sequences being placed are short reads, much smaller than the total length of the alignment, and their positions are distributed randomly across the alignment.[28] We need multiple sets of hash tables to make sure that each read is covered by at least one set of LSH tables. We solve this problem by using a sliding window approach. We construct groups of hash tables, each corresponding to short regions of the alignment, so that each query sequence maps to at least one set of hash tables. Specifically, if all reads have at least k' bases, we divide the alignment into blocks of length $k'/2$, and build a separate set of hash tables for each block. We then sort the reads by their starting position in the alignment and progressively process each block of hash tables, from the beginning to the end of the alignment. This ensures that we never have to store more than one set of hash tables in memory. While the sorting process adds a factor of $\log m$ to the running time per query sequence (where m is the number of query sequences), in practice this cost is negligible, and can be avoided with counting sort.

3.6. *Choosing slow-evolving sites*

So far, we have assumed all sites in the alignment evolve at the same relative rate. This is not true in practice, as genomes, proteins and RNAs contain conserved regions or individual sites. Rate heterogeneity across sites poses both statistical [29] and computational[30] challenges for phylogenetic inference. However, in our case, we can take advantage of rate heterogeneity across sites to improve the speed and accuracy of our algorithm. The running time of the LSH procedure depends on the effective mutation rate between the query sequence and the nearest sequence in the tree. By choosing hash table keys from slowly evolving columns, we reduce the number of hash tables needed to ensure a collision with the same probability, or, equivalently, we enable more distant sequences to be placed correctly using the same number of hash tables. This is particularly important since biologists are often interested in detecting previously unknown clades, many of whom are only distantly related to the known organisms.[5]

We identify slowly-evolving columns by estimating the maximum likelihood relative evolutionary rate for each column in the reference alignment, using the classical Newton-Raphson method. On the other hand, we also discard near-constant columns where the most common character appears in more than 95% sequences, as these are not informative.

An important question is how many slowest-evolving columns to choose. If too many columns are chosen, their average mutation rate will be relatively high. On the other hand, if we choose too few columns, this will lead to a huge variance in the Hamming distances on these columns, which will cause hashing to be less informative about the true evolutionary distances between sequences. We choose to randomly choose from the 200 slowest-evolving columns (not including the near-constant columns), or the bottom 50% slowest-evolving columns, for short alignments.

4. Experiments

4.1. *Data sets*

Our experiments used both simulated and real data. We generated synthetic short read data sets from a simulated 16S rRNA alignment on $n = 78132$ sequences from the FastTree paper.[14]

The length of the alignment was $m = 1287$ bases. The data were generated as follows. First, k taxa were chosen at random from the alignment, and the reference subtree induced by these taxa on the true tree was recorded. We then generated 100,000 simulated short reads. Each short read was generated by sampling with replacement from the $n - k$ sequences not included in the reference tree, and keeping m' contiguous positions, starting at a uniformly-chosen position. The read length m' was chosen from a Gaussian distribution with mean 200 and standard deviation 20.

For the real data set, we used a metagenomic 16S rRNA Illumina library from Alert, Nunavut, Canada.[31] The reads were located in the V3 variable region of the ribosomal RNA. The sequences were clustered at 97% identity, which resulted in 27848 sequences with a mean length of 152 bases, aligned to a reference alignment of 2759 sequences using Infernal.[32] For placement, we used a reference 16S tree from the Living Tree Project.[33] The diameter of the tree was 1.46 mutations per site, and the average distance between two nodes was 0.54.

4.2. *Accuracy*

Table 2 shows the results of our algorithm and pplacer on the synthetic data sets. Our algorithm was less accurate than pplacer, but placed reads within 3 edges from the correct location over 90% of the time. For larger trees, both algorithms tended to place sequences farther from their correct edges in terms of the topological distance (TD). However, the evolutionary distance (ED) between placements and correct locations tended to be lower for larger trees. Larger reference trees contained more short edges which were hard to distinguish for both algorithms. The magnitude of these effects was similar for LSHPlace and pplacer.

Table 2. Accuracy of LSHPlace and pplacer on three simulated data sets. LSHplace is less accurate, but still reasonably close for all data set sizes.

	data set								
	huge.1, 1000 taxa			huge.1, 5000 taxa			huge.1, 10000 taxa		
method	% correct	ED	TD	% correct	ED	TD	% correct	ED	TD
LSHPlace	51.2	0.054	1.12	48.6	0.033	1.17	46.1	0.028	1.34
Pplacer	79.0	0.011	0.30	74.0	0.007	0.39	69.1	0.006	0.51

The predictions of both programs disagreed much more on the real data set. Only 10% of reads were placed on the same edge by both Pplacer and LSHPlace. The average topological distance between a pplacer prediction and an LSHPlace prediction for the same sequence was 14 edges, with evolutionary distance 0.20 mutations per site. The medians of these distances were 10 and 0.15, respectively. We suspect that the accuracy of our algorithm might have been impacted by the presence of many near-constant sites in the alignment, which could have had an adverse effect on the accuracy of distance estimates. Obviously, future work is key to learning more about the accuracy difficulties we encounter with real data.

4.3. *Running times and scalability*

The running times for both programs are shown in Table 3. Each runtime corresponds to placing 100,000 reads into a tree of the given size. In all cases, LSHPlace is around 1.5 to 2 orders of magnitude faster than pplacer.

Table 3. The time to place 100000 taxa into a tree, as a function of the number of taxa in the tree

taxa	1000	5000	10000
our algorithm	7m	12m	17m
pplacer	3.8h	6.4m	18.8h

5. Conclusion

We have presented LSHplace, a new algorithm for phylogenetic placement. By using locality-sensitive hashing, and including inferred ancestral sequences in the hash tables, our algorithm allows us to approximately locate new sequences onto an existing phylogenetic tree extremely rapidly; a local-search procedure allows us to then find an optimal placement quickly for each new sequence. Our work can be used in a variety of domains, but we expect it will be especially useful in the context of metagenomic sampling, where millions of sequence reads are generated and analyzed at the same time to characterize environments. Experimental results, while preliminary, are encouraging, and show that our algorithm speeds up the process of placement by two orders of magnitude; we currently also take an accuracy penalty, but we expect this may be ameliorated by incorporating maximum likelihood inference into the program, which we are currently exploring. Future work will also explore the importance of alignment quality in the placement process.

6. Acknowledgement

We thank Andre Masella, Josh Neufeld and Michael Lynch for sharing data with us, and Erick Matsen for helpful conversations. This work is supported by the Government of Ontario, through an Early Researcher Award to DGB, and by the Natural Sciences and Engineering Research Council of Canada. JT is supported by a David R. Cheriton Scholarship at the University of Waterloo.

7. References

References

1. H. K. Allen, L. A. Moe, J. Rodbumrer, A. Gaarder and J. Handelsman, *ISME Journal* **3**, 243 (2009).
2. D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg *et al.*, *PLoS Biology* **5**, p. e77 (2007).
3. P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight and J. I. Gordon1, *Nature* **449**, 804 (2007).
4. J. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman *et al.*, *Nature Methods* **7**, 335 (May 2010).
5. M. Lynch, A. K. Bartram and J. D. Neufeld, *ISME Journal* (2012), to appear.

6. F. A. Matsen, R. B. Kodner and E. V. Armbrust, *BMC Bioinformatics* **11**, p. 538 (2010).
7. S. A. Berger, D. Krompass and A. Stamatakis, *Systematic Biology* **60**, 291 (2011).
8. M. Stark, S. A. Berger, A. Stamatakis and C. von Mering, *BMC Genomics* **11**, p. 461 (2010).
9. D. G. Brown and J. Truszkowski, Fast reconstruction of phylogenetic trees using locality-sensitive hashing, in *Proceedings of WABI*, (Ljubljana, Slovenia, 2012).
10. P. Indyk and R. Motwani, Approximate nearest neighbors: Towards removing the curse of dimensionality, in *Proceedings of STOC*, (New York, 1998).
11. W. Gerlach, Taxonomic classification of metagenomic sequences (2012), PhD thesis. Bielefeld University.
12. F. Schreiber, P. Gumrich, R. Daniel and P. Meinicke, *Bioinformatics* **26**, 960 (2010).
13. W. Gerlach, S. Jünemann, F. Tille, A. Goesmann and J. Stoye, *BMC Bioinformatics* **10**, p. 430 (2009).
14. M. N. Price, P. S. Dehal and A. P. Arkin, *Mol. Biol. Evol.* **26**, 1641 (2009).
15. D. G. Brown and J. Truszkowski, Towards a practical $O(n \log n)$ phylogeny algorithm, in *Proceedings of WABI*, (Saarbrücken, Germany, 2011).
16. A. C. McHardy, H. G. Martn, A. Tsirigos, P. Hugenholtz and I. Rigoutsos1, *Nature Methods* **4**, 63 (2007).
17. N. N. Diaz, L. Krause, A. Goesmann, K. Niehaus and T. W. Nattkemper, *BMC Bioinformatics* **10** (2009).
18. D. R. Kelley and S. L. Salzberg, *BMC Bioinformatics* **11**, p. 544 (2010).
19. P. Meinicke, K. P. Aßhauer and T. Lingner, *Bioinformatics* **27**, 1618 (2011).
20. S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang *et al.*, *Nucleic Acids Research* **25**, 3389 (1997).
21. D. H. Huson, S. Mitra, N. Weber, H.-J. Ruscheweyh and S. C. Schuster, *Genome Research* **21**, 1552 (2011).
22. M. M. Haque, T. S. Ghosh, D. Komanduri and S. S. Mande, *Bioinformatics* **25**, 1722 (2009).
23. J. Buhler and M. Tompa, *J. Comp. Biol.* **9**, 225 (2002).
24. W. Evans, C. Kenyon, Y. Peres and L. J. Schulman, *The Annals of Applied Probability* **10**, 410 (2000).
25. S. Roch, *Science* **327**, 1376 (2010).
26. O. Gascuel and M. Steel, *Mathematical Biosciences* **227**, 125 (2010).
27. M. A. Steel, Distributions on bicoloured evolutionary trees. (1989), PhD thesis. Massey University.
28. M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff *et al.*, *Science* **5891**, 956 (2008).
29. J. Felsenstein, *Inferring Phylogenies* (Sinauer, 2001).
30. A. Stamatakis, Phylogenetic models of rate heterogeneity: a high performance computing perspective, in *IPDPS*, (IEEE, 2006).
31. A. K. Bartram, M. D. J. Lynch, J. C. Stearns, G. Moreno-Hagelsieb, and J. D. Neufeld, *Applied and Environmental Microbiology* **77**, 3846 (2011).
32. E. P. Nawrocki, D. L. Kolbe and S. R. Eddy, *Bioinformatics* **25**, p. 1713 (2009).
33. P. Yarza, M. Richter, J. Peplies, J. Euzeby and R. A. others, *Systematic and Applied Microbiology* **31**, 241 (2008).

CHIPMODULE: SYSTEMATIC DISCOVERY OF TRANSCRIPTION FACTORS AND THEIR COFACTORS FROM CHIP-SEQ DATA

JUN DING
Department of EECS, University of Central Florida, 4000 central Florida Blvd
Orlando, FL 32816, USA
Email: jding@cs.ucf.edu

XIAOHUI CAI
Shanghai Center for Bioinformation Technology, 100 Qinzhou Rd, Bldg.1, Fl.12
Shanghai, 200235, China
Email: xhcai@scbit.org

YING WANG
Department of EECS, University of Central Florida, 4000 central Florida Blvd
Orlando, FL 32816, USA
Email: ying2010@knights.ucf.edu

HAIYAN HU
Department of EECS, University of Central Florida, 4000 central Florida Blvd
Orlando, FL 32816, USA
Email: haihu@cs.ucf.edu

XIAOMAN LI
Burnett School of Biomedical Science, University of Central Florida, 4000 central Florida Blvd
Orlando, FL 32816, USA
Email: xiaoman@mail.ucf.edu

We have developed a novel approach called ChIPModule to systematically discover transcription factors and their cofactors from ChIP-seq data. Given a ChIP-seq dataset and the binding patterns of a large number of transcription factors, ChIPModule can efficiently identify groups of transcription factors, whose binding sites significantly co-occur in the ChIP-seq peak regions. By testing ChIPModule on simulated data and experimental data, we have shown that ChIPModule identifies known cofactors of transcription factors, and predicts new cofactors that are supported by literature. ChIPModule provides a useful tool for studying gene transcriptional regulation.

1. Introduction

Systematic discovery of transcription factors (TFs) and their cofactors is important for studying gene transcriptional regulation. During gene transcriptional regulation, TFs and their cofactors bind short DNA segments to activate or repress the expression of genes nearby. In general, a TF can bind to a variety of similar DNA segments, called TF binding sites (TFBSs) of this TF. The common pattern of the TFBSs bound by a TF is termed a motif, often represented as a position weight matrix (PWM) or a consensus sequence. In eukaryotes, multiple TFs often bind their TFBSs in short DNA regions of several hundred base pairs long.[1-3] These short DNA regions are called cis-regulatory modules (CRMs).[1] CRMs are common in high eukaryotes.[3] For instance, more than 110,000 CRMs have been predicted in the human genome and are supported by various sources of functional evidence.[4,5] It is the interaction of multiple TFs and their TFBSs instead of individual TFs that determines the temporal spatial expression patterns of genes.[1,4,5] It is thus critical to identify and study TFs and their cofactors.

The chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) experiments provide an unprecedented opportunity for computational methods to study TFs and their cofactors.[6,7] In a typical ChIP-seq experiment, short DNA segments containing TFBSs of a TF are enriched by the chromatin immunoprecipitation (ChIP) using an antibody specific to the TF. These short DNA segments are then sequenced by the next generation sequencing technologies and mapped to a reference genome.[8-11] Finally, genomic regions in the reference genome enriched with the mapped DNA segments are identified as ChIP-seq peak regions.[12,13] These ChIP-seq peak regions likely contain TFBSs of the TF under consideration.[6,7,14] Compared with potential residing regions of the TFBSs of a TF for a gene,[4,5] which is often several hundred million base pair long, a ChIP-seq peak region is typically no longer than 1000 base pairs. Such short regions thus significantly increase the signal to noise ratio and dramatically help to improve the efficiency of computational identification of TFBSs of a TF and its cofactors.

Available computational methods have already started to provide useful prediction of TFs and their cofactors from ChIP-seq data.[12,14-21] The majority of these computational methods identify motifs of individual TFs at a time.[12,14-17,19-21] The underlying assumption of these methods is that motifs of individual cofactors of a TF are overrepresented in the ChIP-seq peak regions of this TF. However, given the fact that a TF has multiple cofactors and motifs of most cofactors only occur in a small portion of peak regions under a condition, motifs of individual cofactors may be often not overrepresented in the ChIP-seq peak regions of this TF.[18,22] Because TFs and their cofactors often regulate their target genes by binding to CRMs in eukaryotes, one recent study has considered motif co-occurrence of a TF and one of its cofactors.[18] However, a TF may bind regulatory regions together with more than one cofactor to regulate its target genes.[1,4,5]

Here we developed a computational method called ChIPModule to systematically identify TFs and cofactors from ChIP-seq data. ChIPModule considers the co-occurrence of TFBSs of any number of different TFs in ChIP-seq peak regions. In brief, starting from all known TF motifs in public databases,[23,24] ChIPModule scans the ChIP-seq peak regions with these motifs to define putative TFBSs of these TFs. ChIPModule then identifies frequently co-occurring TFBSs of a group of any number of TFs by frequent pattern mining methods.[25,26] Finally, ChIPModule assesses the statistical significance of each group of TFs with frequent co-occurring TFBSs by the Poisson clumping heuristic.[27] The significant groups of TFs are called interacting TF groups. The

TFs in the same interacting TF group with a given TF are designated as the cofactors of this TF. Tested on simulated and experimental data, ChIPModule has been shown to successfully predict known cofactors of TFs. It also predicts new cofactors that were supported by literature. Compared with other methods, ChIPModule shows superior performance in terms of dealing with large datasets and identifying known cofactors. We believe ChIPModule will be useful for future ChIP-seq data analysis and gene transcriptional regulation studies.

2. Materials and Methods

2.1. *Framework*

To systematically discover TFs and their cofactors from ChIP-seq data, ChIPModule utilizes the known TF motif information in the TRANSFAC database.[24] Instead of considering one or two TFs at a time, ChIPModule can consider any number of TFs simultaneously. Instead of assuming TFBSs of individual TFs are overrepresented in the ChIP-seq peak regions, ChIPModule assumes that TFBSs of a group of TFs (a TF and its cofactors) are overrepresented in the ChIP-seq peak regions. The framework of ChIPModule consists of the following three steps: prediction of putative TFBSs, identification of frequent co-occurring TF groups, and discovery of TFs and their cofactors. See Figure 1 for the flowchart of ChIPModule. The details are in the following sections.

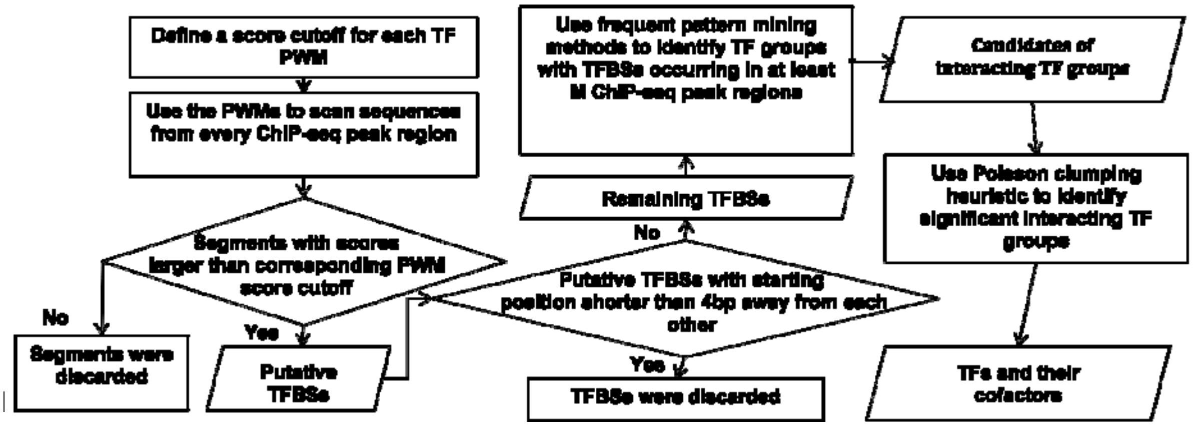

Define a score cutoff for each TF PWM
Use the PWMs to scan sequences from every ChIP-seq peak region
Segments with scores larger than corresponding PWM score cutoff
No
Yes
Segments were discarded
Putative TFBSs
Putative TFBSs with starting position shorter than 4bp away from each other
Yes
TFBSs were discarded
No
Remaining TFBSs
Use frequent pattern mining methods to identify TF groups with TFBSs occurring in at least M ChIP-seq peak regions
Candidates of interacting TF groups
Use Poisson clumping heuristic to identify significant interacting TF groups
TFs and their cofactors

Figure 1. The flowchart of ChIPModule to discover TFs and their cofactors.

2.2. *ChIP-seq Data and Vertebrate Motifs*

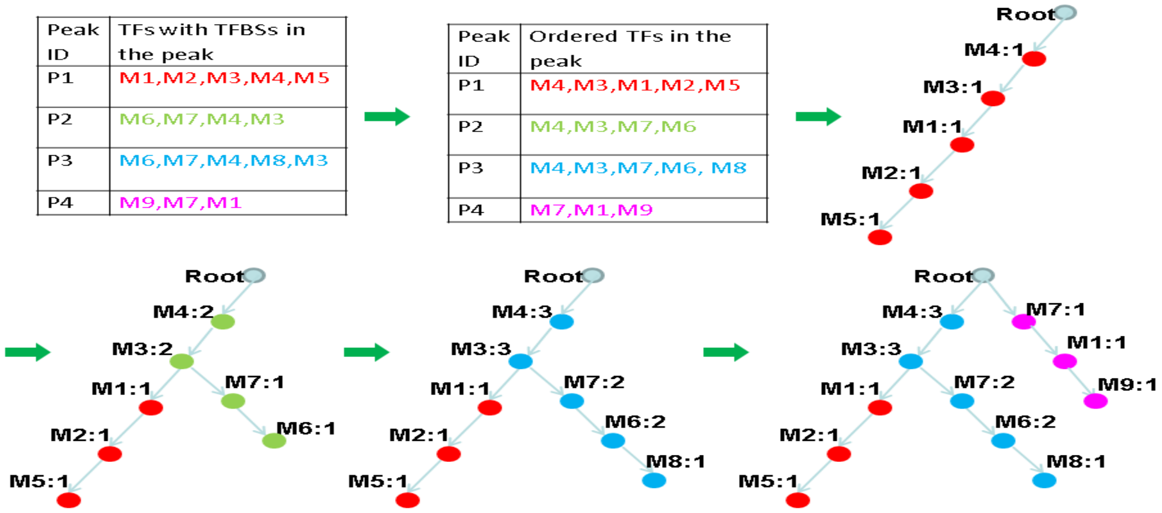
We tested ChIPModule on two ChIP-seq datasets and several simulated datasets. The two ChIP-seq datasets are corresponding to the two TFs ESR1 and E2F1, respectively. For ESR1, which is also called estrogen receptor alpha, the ChIP-seq peak regions defined at the p-value cutoff 0.001 were downloaded from the GSM365926 sample in the GEO database.[28] In total, we obtained 3257 peak regions, with the average length of 595 base pairs. For E2F1, the peak region at the p-value cutoff 0.001 were downloaded from the SYDH TFBS track at the UCSC genome browser.[29] We obtained 10196 peak regions, with the average length of 878 base pairs for E2F1. For each ChIP-seq peak region, we extended it equally on the two sides such that it is at least 800 base pairs long. This extension is to enhance the chance for cofactors to occur in peak regions, as TFBSs of certain cofactors may be not within the originally defined ChIP-seq peak regions. The known TF motifs used in the following study were obtained from the TRANSFAC 9.2 database,[24] where all 522 vertebrate PWMs were extracted. Pseudo counts were introduced to regularize each PWM, as in previous studies.[30,31]

2.3. *Identification of Putative TFBSs in ChIP-seq Peak Regions*

To identify putative TFBSs of a TF in ChIP-seq peak regions, we scan the non-repetitive sequences in each peak region and calculate the score of each segment in a peak region by using the above regularized PWM of this TF. A slide window is used to define segments. That is, given a TF motif of length k and a peak region of length L, we consider all L – k+1 distinct segments. We calculate the score of a segment by the following formula: $score(a\ segment\ x_1 x_2 \cdots x_k) = \sum_{i=1}^{k} \log\left(\frac{f(x_i,i)}{fb(x_i)}\right)$. Here $fb(x_i)$ is the average frequency of the nucleotide x_i in the human reference genome, $f(x_i, i)$ is the frequency of the nucleotide x_i at the i-th position of the motif PWM, and k is the width of the motif. If the score is larger than a predefined cutoff for this TF, this segment will be claimed as a putative TFBS of this TF. In this study, the predefined cutoff for each TF is defined as the 99.99% quartile of the score distribution of DNA segments of length k, when using the PWM of this TF to scan 100 kb long random sequences. The random sequences were generated by permuting input sequences from ChIP-seq peak regions. Note that motifs of certain TFs may have the tendency to occur together, merely due to the similarity of their PWMs. To deal with it, we sort the putative TFBSs by their start positions and discard the overlapped TFBSs with the lower score, when the start positions of two putative TFBSs are smaller than 4 base pairs. We use 4 base pairs here to remove overlapping TFBSs as in previous studies.[31,32]

2.4. *Identification of Groups of TFs with Frequently Co-occurring Motifs in Peak Regions*

We aim to identify groups of TFs whose putative TFBSs co-occur in more than a specified number of peak regions, say M peak regions. The rationale is that the chance that multiple TFs with their TFBSs co-occurring in a ChIP-seq region is much smaller than that of individual TFs. That is, if we observe a group of TFs with their TFBSs co-occurring in a large number of peak regions, it is likely their co-occurrence is not by chance and thus this group of TFs likely work together to regulate genes. To discover such a group of TFs, we use a tree to represent the above identified TFBSs and identify all groups of TFs with their TFBSs co-occurring in at least M peak regions (Figure 2). In brief, first, we count the number of peak regions containing TFBSs of each TF and sort these TFs according to the corresponding number, from the largest to the smallest. Second, we sort the TFBSs in each peak region, such that TFBSs of the TFs occurring in more peak regions rank at the beginning. Third, starting from the first peak region until the last peak region, we build a tree to store the TFs whose TFBSs occurring in a peak region (Figure 2). At the beginning, a tree with only a root node is built. Next, the nodes for TFs in the first peak region are added in order. Finally, nodes for TFs in other peak regions are added, if there is no branch in the current tree matching the order of TFs in the peak regions under consideration (Figure 2). With the built tree, we will identify all groups of TFs with TFBSs occurring in at least *M* peak regions. In brief, starting from the TF that occurs in at least M peak regions and occurs in the smallest number of peak regions, we will obtain all the branches in the built tree that contains this TF. For instance, when M=2, we will start from the TF M1 or M6 in Figure 2. Assume we will start from the TF M1. In this case, we obtain two branches, M4:3-M3:3-M1:1 and M7:1-M1:1. We will then construct a tree using the obtain branches for this specific TF, by assuming each branch represent motifs in a peak region. In this case, we will have a tree with the above two branches. It is clear that no group of TFs that includes TF M1 and occurs at least M times. Next, we will obtain all branches and construct a tree for the TF that occurs in the second smallest number of peak regions. Since we already consider the TF M1, this time TF M6 occurs in the second smallest number of peak regions. This time we have only one branch that containing M6, which is M4:3-M3:3-M7:2-M6:2. In this case, it is evident that the group of TFs (M4,M3,M7,M6) co-occur twice in the peak regions considered in Figure 2. We will keep considering a TF each time until we find the groups of TFs that co-occur at least M times for the TF occurring in the most peak regions.



Peak ID	TFs with TFBSs in the peak
P1	M1,M2,M3,M4,M5
P2	M6,M7,M4,M3
P3	M6,M7,M4,M8,M3
P4	M9,M7,M1

Peak ID	Ordered TFs in the peak
P1	M4,M3,M1,M2,M5
P2	M4,M3,M7,M6
P3	M4,M3,M7,M6, M8
P4	M7,M1,M9

Figure 2. The procedure to construct a tree to represent TF co-occurrence.

2.5. *Identification of TFs and Their Cofactors*

With groups of TFs identified above, we want to assess their statistical significance to obtain interacting TFs. As mentioned above, a TF in a group of interacting TFs is a cofactor of all other TFs in the same group and vice versa. We use the Poisson clumping heuristic[27] to compute the statistical significance of a group of TFs with the assumption that each TF bind a ChIP-seq peak region independently according to a Poisson process. In brief, assume there are N_1 ChIP-seq peak regions, and the average length of a peak region is L, the total number of known motifs is N_2, and

λ_k is the rate parameter of the Poisson process for the k-th TF. For a group of TFs composed of the $m_1, m_2, \ldots, m_n$-th TFs identified above, the probability that TFBSs of the n TFs occur in a peak region of length L is $P_1 = \prod_{i=1}^{n}(1 - e^{-L\lambda_{m_i}})$. The probability that this group of TFs with TFBSs co-occurring in at least K peak regions is $P_2 = 1 - \sum_{k=0}^{K-1} \mathrm{C}_{\mathrm{N}_1}^{\mathrm{k}} \mathrm{P}_1^{\mathrm{k}} (1 - \mathrm{P}_1)^{(\mathrm{N}_1 - \mathrm{k})}$. Since N_2 TFs can produce $\mathrm{C}_{\mathrm{N}_2}^{\mathrm{n}}$ different group of TFs by chance, we require $P_2 < 0.05/C_{N_2}^{n}$ to claim a group of TFs as a group interacting TFs. With the groups of interacting TFs, we then treat each TF and all other TFs from the same group of interacting TFs as TFs and cofactors. The genes closest to the peak regions containing TFBSs of a group of interacting TFs are defined as the target genes of this group of interacting TFs. Similarly, genes closest to the peak regions containing TFBSs of a TF are defined as the target genes of this TF.

3. Results

3.1. *ChIPModule Identified Implanted TFs and Their Cofactors in Simulated Data*

We tested ChIPModule on three simulated datasets with five different parameter setups (Table 1). In each simulated dataset, we generated 2000 to 8000 random sequences, with the length distribution of these sequences the same as those in the E2F1 ChIP-seq dataset. We then randomly inserted TFBSs of 20 groups of TFs, using known TF PWMs in the TRANSFAC database.[24] The number of TFs in a group varied from 2 to 13, the largest number of TFs in a TF group from a previous study.[5] For each group of TFs, we inserted their TFBSs in only 10% randomly chosen sequences. We then applied ChIPModule to these simulated datasets with M as the 10% of the number of sequences. Recall that M is the minimal number of sequences (peak regions) required to contain TFBSs of each TF in a TF group. From Table 1, it is clear that ChIPModule identified as many as 17 of the inserted TF groups, which represents a sensitivity of 85% (the percent of inserted TF groups predicted). We also calculated the specificity of ChIPModule by checking how many percent of predicted TF groups are similar to the inserted TF groups. Note that we could not require the predicted TF groups are exactly as the inserted TF groups, since different TFs may bind similar motifs. A group of predicted TFs is claimed to be similar to a group of inserted TFs, if for each TF in one group, there is one TF in the other group that share a similar motif with the TF under consideration. A pair of TFs shares similar motifs if the STAMP p-value of the similarity of the two motifs is less than 1E-5, as in previous studies.[33-35]

We also noticed that several inserted TF groups were not identified. We hypothesized that these TF groups were missed by ChIPModule because not all TFBSs of the TFs in these TF groups satisfied the required putative TFBS cutoff used in Section 2.3, or TFBSs of different TFs may overlap and some of them were thus discarded. If this hypothesis was true, ChIPModule could correctly identify even more inserted TF groups if one used a smaller M. We thus further tested two of the smaller datasets using a smaller M. From the last two rows in Table 1, it is clear that using a smaller M indeed improved the accuracy of ChIPModule. For instance, ChIPModule successfully predicted all inserted TF groups when we used M as 7.5% of the number of sequences. This demonstrates that the developed tool, ChIPModule, can systematically identify TFs and their cofactors in ChIP-seq datasets.

Table 1. Correctly Predicted TF groups by ChIPModule on simulated datasets.

#total sequences	**#sequences with inserted TFBSs of a group of TFs**	**M**	**#Correctly predicted TF groups**	**Sensitivity**	**specificity**
2000	200	200	17	85%	88.8%
4000	400	400	15	75%	93.6%
8000	800	800	12	60%	95.8%
2000	200	150	20	100%	79.9%
4000	400	350	16	80%	74.2%

3.2. *ChIPModule Identified TFs and Their Cofactors in Experimental Data*

We further tested ChIPModule on the two ChIP-seq datasets mentioned above. These two datasets were used because the two TFs, ESR1 and E2F1, are well studied. In addition, several cofactors are known for each TF. Similar to the simulated studies, we used *M* as 10% of the number of ChIP-seq peak regions we obtained for the two TFs, respectively, when applying ChIPModule to the two ChIP-seq datasets.

In total, we identified 1334 and 6428 groups of interacting TFs in the ESR1 dataset and the E2F1 dataset, respectively. The number of TFs in these interacting TF groups is from 2 to 5 for the ESR1 dataset (average 2.16), and from 2 to 7 for the E2F1 dataset (average 4.8). To see whether ChIPModule predicted these interacting TF groups by chance, we permuted the input sequences from the ChIP-seq peak regions in each dataset and applied ChIPModule to these random sequences generated by permutation for each TF. We found that ChIPModule predicted 0 and 87 interacting TF groups in the two permuted random datasets, respectively. The much lower number of predicted interacting TF groups demonstrates that ChIPModule has a low false positive prediction rate (87/6428=1.35%), which confirms a high specificity of ChIPModule and implies the functionality of the majority of the predicted interacting TF groups.

Table 2. Several identified cofactors and their literature support.

Dataset	Known cofactors	Supported new cofactors
ESR1	FOXA[36], OCT1[36], C/EBP[36], AP-1[36]	p300[37], VDR[38]
E2F1	SP1[39], MYC[40]	NF-kappaB,[41] YY1[42]

We next checked whether ChIPModule identified known cofactors of the two TFs. For the TF ESR1, we found a few known cofactors, such as FOXA, OCT1, C/EBP and AP-1 (Table 2).[36] For the TF E2F1, we also found several known cofactors, such as SP1[37] and MYC[38] (Table 2). The de novo discovery of the known cofactors of the two TFs supports the fact that ChIPModule can identify cofactors of TFs from ChIP-seq data.

Besides known cofactors, ChIPModule also identified new cofactors for the TF ESR1 (Table 2). ChIPModule predicted that P300 and VDR are also cofactors of ESR1, which are supported by literature.[37,38] For instance, ChIPModule identified a group of interacting TFs composed of three TFs, ESR1, VDR, and COUPTF. The TF VDR was reported to interact with ESR1.[38] It is also known that ESR1 is regulated by COUPTF, through both direct DNA binding competition and protein-protein interactions.[43] Moreover, it is suggested that COUPTF plays a master role in regulating the transactivation by VDR.[44] Based on these studies,[38,43,44] it is highly likely that TFs in the this predicted interacting TF group interact with each other, which supports the functionality of this group of TFs. We further investigated the function of the target genes of this group of TFs by the gene ontology (GO) enrichment analysis. The GO enrichment analysis is a common approach to test whether a group of gene significantly share functions based on their annotated GO terms.[45] We found that the target genes of this group of TFs significantly share a function, in utero embryonic development (GO:0001701, corrected p-value= 9.66E-05).[45] The sharing of functions by target genes suggests that these target genes are likely co-regulated, which further supports the functionality of this predicted interacting TF group.

ChIPModule identified new cofactors for the TF E2F1 as well (Table 2). Several of the predicted cofactors of E2F1 are supported by literature, such as NF-kappaB and YY1.[41,42] For instance, ChIPModule predicted a group of interacting TFs consisting of four TFs. These four TFs are YY1, E2F1, SP1, and BSAP. The TFs YY1 and SP1 are reported to be interacted with E2F1.[39,42] It is also known that YY1 interacts with the TFs SP1 and BSAP.[46,47] These studies suggest that the other three TFs in this group interact with E2F1 directly or indirectly, which supports the functionality of this group of interacting TFs. The GO enrichment analysis shown that the target genes of this group of TFs significantly share a function, positive regulation of transcription factor activity (GO: 0051091, corrected p-value=2.6E-4). Thus, the four TFs in this group of interacting TFs likely coordinately regulate their common target genes.

3.3. *A Large Number of Predicted Interacting TF Groups do not Contain the TFs Used for the ChIP-seq Experiments*

In the above analysis, we found that a large percentage of predicted interacting TF groups do not contain the TFs used for the ChIP-seq experiments. For instance, in the E2F1 ChIP-seq dataset, 4782 out of the 6248 predicted interacting TF groups do not contain the TF E2F1. We hypothesized that the exclusion of the corresponding TFs in our predictions is most likely due to the indirect binding of the corresponding TFs to the ChIP-seq peak regions through the interaction with cofactors. In other word, there are at least two types of ChIP-seq peak regions, one bound by the corresponding TF directly, the other bound by the cofactor of the corresponding TF that interact with the cofactors. To support this hypothesis, we examined the predicted interacting TF groups and found that this is the case for several interacting TF groups. We provided two of such supporting examples below.

Example 1. An interacting TF group composed of the TFs GATA1 and SP1 was found in the ESR1 dataset. A previous study has shown that GATA1 interacts with SP1 to regulate their target genes.[48] In addition, we found that the target genes of SP1 shared the function, synaptic vesicle (GO:000802, corrected p-value=9.0E-3). Meanwhile, the target genes of GATA1 shared a similar function, synaptic transmission (GO:0007268, corrected p-value=3.0E-2). Consistently, the target genes of this interacting TF group significantly shared the function, synaptic vesicle (GO:000802, corrected p-value =4.7E-3). The interaction of the two TFs and the consistency of the function of individual TFs and the TF group suggest that this group of interacting TFs is likely functional. In addition, the TF ESR1 was reported to interact with SP1 in breast cancer cells.[49] It is thus likely that ESR1 interacts with this group of interacting TFs, which directly bind the ChIP-seq peak regions.

Example 2. The interacting TF group with two TFs ETS and SP1 was identified from the E2F1 dataset. Although E2F1 was not included in this group, the two TFs in this group were found to interact with E2F1.[39,50] A previous study has shown that E2F1 specifically interacts with ETS-related TFs.[50] The TF SP1 has also been found to interact with E2F1.[39] Moreover, the ETS TF family cooperates with SP1 to activate the human Tenascin-C promoter.[51] In addition, the target genes of this TF group significantly shared a function, RNA splicing (GO:0008380, corrected p-value 5.29E-09). Therefore, E2F1 likely interacts with this group of TFs, which directly bind the ChIP-seq peak regions. These pieces of evidence support the above hypothesis that the corresponding TF indirectly bind the ChIP-seq regions through the interaction with its cofactors.

3.4. *Comparisons with Other Methods*

We attempted to compare ChIPModule with coMOTIF[18] and W-ChIPMotifs[17]. coMOTIF jointly considers two motifs in ChIP-seq peak regions, and W-ChIPMotifs is a web application tool for de novo motif discovery from ChIP-based high throughput data. Under default parameters, coMOTIF took more than a week to run on the ESR1 dataset (3257 peaks, each 595 base pair long on average). We could not make it work on the E2F1 dataset, which may be due to the much larger data size of this dataset (10196 peaks, each 878 base pair long on average). As to W-ChIPMotifs, we were unable to obtain a local version of this tool and the online version of this tool cannot accept more than 3000 sequences. On the contrary, ChIPModule took about 533 seconds on the ESR1 dataset and 1129 seconds to run on the E2F1 dataset on a desktop computer (Intel core 2 Duo CPU, 2.93 GHz, 4G RAM), which make it suitable for gene transcriptional regulation studies based on ChIP-seq experiments. We provide both the command line mode of the ChIPModule that can be run on the DOS, Linux, and OS environments and the GUI mode of the Windows version ChIPModule. Detailed information about ChIPModule is in the readme file on the download package at http://www.cs.ucf.edu/~xiaoman/ChIPModule/ChIPModule.html.

Because it is difficult to run W-ChIPMotifs and coMOTIF on the original datasets, we chose to compare ChIPModule with the two software tools on the top 100 peak regions of the ESR1 and E2F1 datasets. In the top 100 peak regions of ESR1, for the known cofactors FOXA, OCT1,C/EBP,AP-1,p300, and VDR mentioned above, ChIPModule identified two known co-factors VDR and p300, W-ChIPMotifs identified C/EBP, and coMOTIF did not identify any of the above co-factors. In the top 100 peak regions of E2F1, for the known aforementioned co-factors sp1, myc, NF-kappaB, and YY1, ChIPModule identified all four cofactors, W-ChIPMotifs identified sp1, and coMOTIF identified the TF combination E2F1 and sp1.

We also compared ChIPModule with the two tools on simulated data. We inserted TFBSs of 50 groups of TFs into 10 out of 100 random sequences. There are 43 TFs contained in the 50 groups. W-ChIPMotifs identified motifs of 10 out of 43 TFs. coMOTIF correctly predicted two TFs in 11 out of 50 inserted TF groups. ChIPModule discovered 45 out of 50 inserted TF groups. In addition, motifs of 39 out of the 43 inserted TFs have been included in these predictions.

4. Discussion

We developed a novel method, ChIPModule, to systematically discover TFs and their cofactors from ChIP-seq data. Tested on simulated datasets, ChIPModule identified the majority of all planted interacting TF groups. Applied to experimental datasets, ChIPModule identified known cofactors and predicted new cofactors, which were supported by literature. ChIPModule thus provides a useful method to study gene transcriptional regulation.

A main assumption in the ChIPModule is that multiple TFs instead of individual TFs regulate their target genes under a given condition. This assumption is supported by the GO enrichment analysis[45] of the target genes of the predicted interacting TF groups and those of individual TFs. We found that target genes of 119 out of 150 top groups of interacting TFs (79.33%) have smaller GO enrichment p-value than those of individual TFs in the same groups for the ESR1 dataset. Meanwhile, target genes of 149 out of 150 top groups of interacting TFs (99.9%) have smaller GO enrichment p-value than those of individual TFs in these groups for the E2F1 dataset. Moreover, the target genes of a group of interacting TFs often share functions while target genes of individual TFs may not share any function. For instance, for the interacting TF group composed of the TFs PAX4 and SP1 in the ESR1 dataset, we could find that its target genes significantly share the function, negative regulation of follicle-stimulating hormone secretion (GO:0046882, corrected p-value=8.18E-005). However, the target genes of PAX4 or SP1 share no similar function.

In the above study, we found that the predicted interacting TF groups often do not contain the corresponding TFs used for the ChIP-seq experiments. We provided concrete examples to support the hypothesis that the corresponding TFs could interact with their cofactors, while the cofactors directly bind the ChIP-seq peak regions. Note that alternative explanation exists. For instance, if we lower the p-value cutoffs used to define putative TFBSs in Section 2.3, or choose a smaller M in Section 2.4, we could find more predicted interacting TF groups containing the corresponding TFs. However, our experience with the two ChIP-seq datasets and other ChIP-seq datasets[5,52] suggests that the proposed hypothesis is likely the main reason for the exclusion of the corresponding TFs in the predicted interacting TF groups.

Several options in our developed software make ChIPModule a widely applicable tool for studying gene transcription regulation. First, besides using the TF PWMs in public databases,[23,24] users can use self-defined TF PWMs. Second, users can choose different p-value cutoffs to define putative TFBSs in ChIPModule. This is necessary, as one wants to use more stringent p-value cutoffs for large datasets while use looser p-value cutoffs for small datasets. Third, the discovered TFs and their cofactors by ChIPModule are organized in four different formats, which help users to study these interacting TF groups at different scales. We believe ChIPModule will be a useful tool for future gene transcriptional regulation studies.

5. Acknowledgement

This work was supported by a National Science Foundation, Chemical, Bioengineering, Environmental, and Transport Systems grant 1125679 (HH), a National Science Foundation, Faculty Career Development grant 1149955 (HH), and a National Science Foundation, Information & Intelligent Systems grant 1218275 (XL).

Reference

1. M. I. Arnone and E. H. Davidson, *Development (Cambridge, England)* **124** (10), 1851 (1997).
2. C. H. Yuh, H. Bolouri, and E. H. Davidson, *Science (New York, N.Y* **279** (5358), 1896 (1998).
3. L. Li, Q. Zhu, X. He et al., *Genome biology* **8** (6), R101 (2007).
4. M. Blanchette, A. R. Bataille, X. Chen et al., *Genome research* **16** (5), 656 (2006).
5. X. Cai, L. Hou, N. Su et al., *BMC genomics* **11**, 567 (2010).
6. D. S. Johnson, A. Mortazavi, R. M. Myers et al., *Science (New York, N.Y* **316** (5830), 1497 (2007).
7. G. Robertson, M. Hirst, M. Bainbridge et al., *Nature methods* **4** (8), 651 (2007).
8. J. Shendure, G. J. Porreca, N. B. Reppas et al., *Science* **309** (5741), 1728 (2005).
9. M. Margulies, M. Egholm, W. E. Altman et al., *Nature* **437** (7057), 376 (2005).
10. H. Li and N. Homer, *Brief Bioinform* (2010).
11. B. Langmead, C. Trapnell, M. Pop et al., *Genome Biol* **10** (3), R25 (2009).
12. H. Ji, H. Jiang, W. Ma et al., *Nature biotechnology* **26** (11), 1293 (2008).
13. Y. Zhang, T. Liu, C. A. Meyer et al., *Genome biology* **9** (9), R137 (2008).
14. A. Valouev, D. S. Johnson, A. Sundquist et al., *Nature methods* **5** (9), 829 (2008).
15. M. Hu, J. Yu, J. M. Taylor et al., *Nucleic acids research* **38** (7), 2154 (2010).
16. E. Mercier, A. Droit, L. Li et al., *PloS one* **6** (2), e16432 (2011).
17. V. X. Jin, J. Apostolos, N. S. Nagisetty et al., *Bioinformatics (Oxford, England)* **25** (23), 3191 (2009).
18. M. Xu, C. R. Weinberg, D. M. Umbach et al., *Bioinformatics (Oxford, England)* **27** (19), 2625 (2011).
19. M. Thomas-Chollier, E. Darbo, C. Herrmann et al., *Nature protocols* **7** (8), 1551 (2012).
20. I. V. Kulakovskiy, V. A. Boeva, A. V. Favorov et al., *Bioinformatics (Oxford, England)* **26** (20), 2622 (2010).

21. S. J. van Heeringen and G. J. Veenstra, *Bioinformatics (Oxford, England)* **27** (2), 270 (2011).
22. L. Li, *J Comput Biol* **16** (2), 317 (2009).
23. A. Sandelin, W. Alkema, P. Engstrom et al., *Nucleic acids research* **32** (Database issue), D91 (2004).
24. E. Wingender, P. Dietze, H. Karas et al., *Nucleic acids research* **24** (1), 238 (1996).
25. G. Grahne and J. Zhu, *IEEE transactions on knowledge and data engineering* **17**, 1347 (2005).
26. J. Han, J. Pei, and Y. Yin, in *ACM SIGMOD International Conference on Management of Data* (Dallas, USA, 2000).
27. D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*. (Springer-Verlag, 1989).
28. R. Edgar, M. Domrachev, and A. E. Lash, *Nucleic acids research* **30** (1), 207 (2002).
29. W. J. Kent, C. W. Sugnet, T. S. Furey et al., *Genome research* **12** (6), 996 (2002).
30. J. M. Claverie and S. Audic, *Comput Appl Biosci* **12** (5), 431 (1996).
31. J. Hu, H. Hu, and X. Li, *Nucleic acids research* **36** (13), 4488 (2008).
32. R. Sharan, I. Ovcharenko, A. Ben-Hur et al., *Bioinformatics (Oxford, England)* **19 Suppl 1**, i283 (2003).
33. F. Fauteux, M. Blanchette, and M. V. Stromvik, *Bioinformatics (Oxford, England)* **24** (20), 2303 (2008).
34. B. D. Reed, A. E. Charos, A. M. Szekely et al., *PLoS genetics* **4** (7), e1000133 (2008).
35. J. Ding, X. Li, and H. Hu, *Plant physiology* (2012).
36. J. S. Carroll, C. A. Meyer, J. Song et al., *Nature genetics* **38** (11), 1289 (2006).
37. B. D. Jeffy, J. K. Hockings, M. Q. Kemp et al., *Neoplasia (New York, N.Y* **7** (9), 873 (2005).
38. E. M. Colin, A. G. Uitterlinden, J. B. Meurs et al., *The Journal of clinical endocrinology and metabolism* **88** (8), 3777 (2003).
39. S. Y. Lin, A. R. Black, D. Kostic et al., *Molecular and cellular biology* **16** (4), 1668 (1996).
40. S. W. Hiebert, M. Lipp, and J. R. Nevins, *Proceedings of the National Academy of Sciences of the United States of America* **86** (10), 3594 (1989).
41. X. Palomer, D. Alvarez-Guardia, M. M. Davidson et al., *PloS one* **6** (5), e19724 (2011).
42. S. Schlisio, T. Halperin, M. Vidal et al., *The EMBO journal* **21** (21), 5775 (2002).
43. C. M. Klinge, B. F. Silver, M. D. Driscoll et al., *The Journal of biological chemistry* **272** (50), 31465 (1997).
44. A. J. Cooney, X. Leng, S. Y. Tsai et al., *The Journal of biological chemistry* **268** (6), 4152 (1993).
45. E. I. Boyle, S. Weng, J. Gollub et al., *Bioinformatics (Oxford, England)* **20** (18), 3710 (2004).
46. K. Calame and M. Atchison, *Genes & development* **21** (10), 1145 (2007).
47. E. Seto, B. Lewis, and T. Shenk, *Nature* **365** (6445), 462 (1993).
48. K. D. Fischer, A. Haese, and J. Nowock, *The Journal of biological chemistry* **268** (32), 23915 (1993).
49. K. Kim, R. Barhoumi, R. Burghardt et al., *Molecular endocrinology (Baltimore, Md* **19** (4), 843 (2005).
50. L. Hauck, R. G. Kaba, M. Lipp et al., *Molecular and cellular biology* **22** (7), 2147 (2002).
51. F. Shirasaki, H. A. Makhluf, C. LeRoy et al., *Oncogene* **18** (54), 7755 (1999).
52. Y. Wang, X. Li, and H. Hu, *Genomics* **98** (6), 445 (2011).

USING BIOBIN TO EXPLORE RARE VARIANT POPULATION STRATIFICATION[*]

CARRIE B. MOORE[†]
Center for Human Genetics Research, Vanderbilt University, 519 Light Hall
Nashville, TN 37232, USA
Email: carrie.c.buchanan@vanderbilt.edu

JOHN R. WALLACE[‡]
Center for Systems Genomics, Pennsylvania State University, 512 Wartik Laboratory
University Park, PA 16802, USA
Email: jrw32@psu.edu

ALEX T. FRASE
Center for Systems Genomics, Pennsylvania State University, 512 Wartik Laboratory
University Park, PA 16802, USA
Email: atf3@psu.edu

SARAH A. PENDERGRASS
Center for Systems Genomics, Pennsylvania State University, 512 Wartik Laboratory
University Park, PA 16802, USA
Email: sap29@psu.edu

MARYLYN D. RITCHIE
Center for Systems Genomics, Pennsylvania State University, 512 Wartik Laboratory
University Park, PA 16802, USA
Email: marylyn.ritchie@psu.edu

Rare variants (RVs) will likely explain additional heritability of many common complex diseases; however, the natural frequencies of rare variation across and between human populations are largely unknown. We have developed a powerful, flexible collapsing method called BioBin that utilizes prior biological knowledge using multiple publicly available database sources to direct analyses. Variants can be collapsed according to functional regions, evolutionary conserved regions, regulatory regions, genes, and/or pathways without the need for external files. We conducted an extensive comparison of rare variant burden differences (MAF < 0.03) between two ancestry groups from 1000 Genomes Project data, Yoruba (YRI) and European descent (CEU) individuals. We found that 56.86% of gene bins, 72.73% of intergenic bins, 69.45% of pathway bins, 32.36% of ORegAnno annotated bins, and 9.10% of evolutionary conserved regions (shared with primates) have statistically significant differences in RV burden. Ongoing efforts include examining additional regional characteristics using regulatory regions and protein binding domains. Our results show interesting variant differences between two ancestral populations and demonstrate that population stratification is a pervasive concern for sequence analyses.

[*] This project was funded by NIH grants LM010040, HL065962, CTSI: UL1 RR033184-01.

[†] This work is supported by F30AG041570 from the National Institute on Aging and Public Health Service award T32 GM07347 from the National Institute of General Medical Studies for the Medical-Scientist Training Program.

[‡] This work is supported by a grant with the Pennsylvania Department of Health using Tobacco CURE Funds.

1. Introduction and Background

In the field of human genetics research, there has been increasing interest in the role of rare variation in complex human disease. This is in many ways a response to changing technology, but more importantly a response to the inability to completely explain heritability in common complex diseases and recognition of the true multifactorial mechanisms of genetic inheritance. It is believed that rare variants (RVs) likely have a larger effect size (compared to genome-wide association study (GWAS) findings) and can act alone, in concert with other RVs, or together with common variants. There is increasing evidence to support a role for RVs to contribute to common, complex disease. Recent studies on obesity, autism, schizophrenia, hypertriglyceridemia, hearing loss, complex I deficiency, age-related macular degeneration, kabuki syndrome, and type-1 diabetes implicate RVs with moderate effect sizes.[1–6]

Because of the frequency of RVs and thus the necessary sample size to gain reasonable power, association signals for RVs in a simple SNP-phenotype association study are harder to detect. Methods can be used to group the RVs and test for group association with disease status. Grouping, also known as binning or burden testing, better accounts for genetic heterogeneity and the possibility for multiple RVs to act in concert, which would have otherwise been overlooked in GWAS. Collapsing methods are popular for many reasons: to reduce the degrees of freedom in the statistical test, easy application to case-control studies (not limited to family transmission filtering), applicability to whole-genome data, and an accessible way to enrich association signals by combining RVs (often otherwise undetectable). Several collapsing methods have been published in the past five years.[2,7–14]

Our BioBin approach meets a critical need for an improved binning algorithm through the advantage of prior biological knowledge and potential cumulative effects of biologically aggregated RVs. BioBin requires the Library of Knowledge Integration (LOKI), which contains diverse prior knowledge from multiple collections of biological data. BioBin can be used to apply multiple levels of burden collapsing/testing, including: regulatory regions, evolutionary conserved regions, genes, and/or pathways without a need for an external feature file. Users can define the boundaries of a feature based on a specific hypothesis of interest; for example, is there a difference in RV burden in regions with known transcription factor binding sites between two groups? The adaptable design of BioBin and incorporation of prior biological knowledge provides the user with a flexible binning system and the opportunity to test a range of hypotheses.

While BioBin was specifically developed to investigate RV burden in traditional genetic trait studies, this tool is useful for exploring the natural distribution of RVs in ancestral populations. Rapid population growth and weak purifying selection has allowed ancestral populations to accumulate low frequency variants, many of which are deleterious and potentially causal to human disease.[15,16] These RVs exhibit ancestral heterogeneity and can be completely unique to a single population. To demonstrate the magnitude of population stratification in RVs, Tennessen et al. identified more than 500,000 single nucleotide variants (SNVs) using 15,585 protein-coding genes from 2,440 individuals. Of these SNVs, 86% had a MAF < 0.5% and 82% were population specific (European American or African American).[16] Others have documented differences between ancestral populations using gene drug targets[15] and ENCODE data.[9,17] A thorough understanding of the distribution of RVs across populations will help uncover unknown demographic and evolutionary forces acting on the genome. Since RVs are likely essential in understanding the etiology of common complex traits, it is also critical to understand population stratification for the sake of sequence data analysis. The magnitude of population stratification (and consequential inflation of type I error) is not yet known and adequate methods to correct for stratification have not been developed.[18,19]

Herein we present the methodology of BioBin and the structure of LOKI that provides the prior knowledge for assignment of bins in BioBin. We have tested BioBin using data simulations specifying RVs and applied BioBin to European descent (CEU) and Yoruba (YRI) individuals from 1000 Genomes Project Phase I data. Our tests show BioBin is a flexible and effective method for biological knowledge directed binning of RV data and highlight the importance of investigating RV distribution differences across diverse populations.

2. Methods

2.1. *General framework*

The rare variant analysis occurs in two steps: first, BioBin generates bins based on user-defined parameters and information from LOKI; second, the user applies an appropriate statistical association test. To bin, the user can change options in the configuration file to select certain database sources, adjust feature types, and/or configure the minor allele frequency (MAF) binning threshold. The MAF binning threshold determines the allele frequency limit under which variants are binned. For example, if the threshold is 0.03, a locus with MAF 0.04 would not be included in a bin but a locus with a MAF of 0.029 would be included. The minor allele at a given locus is determined from the second most frequent allele in the control group. For a biallelic locus, this is always the rarer allele. For a triallelic locus, the MAF reported by BioBin is calculated from the second most frequent allele, but all rare alleles are binned. Common alleles (including loci with low frequency variants above the binning threshold) are not binned and are not considered in this analysis, but could be combined with RV bins in subsequent statistical analyses. An example of major and MAF inclusion/exclusion from a single group is shown in Table 1.

Table 1. Variant binning with a MAF binning threshold < 0.05

Major Allele (AF)	Minor Allele(s) (AF)	MAF	Variants Binned
C: 0.97	T: 0.03	0.03	T
T: 0.80	A: 0.16, G: 0.04	0.16	
G: 0.95	C: 0.03, T: 0.02	0.03	C, T

Although the major and minor alleles are designated by frequency in the control group, RVs in the case group also contribute to the variants binned. To simplify, “rareness” is calculated separately for cases and controls. If a variant is considered rare (allele frequency less than the MAF bin threshold) in either group, it will contribute to the bin. In this way, we are not only accumulating risk variants (higher frequency in cases than controls) but also potentially protective variants (lower frequency in cases than controls). This reduces the number of false positive bins and reduces the correlation between bin size and significance.

2.2. *Software*

2.2.1. *BioBin*

BioBin is a standalone command line application written in C++ that uses a prebuilt LOKI database. Source distributions are available for Mac and linux operating systems and require minimal prerequisites to compile. Included in the distribution are tools that allow the user to create and update the LOKI database by downloading information directly from source websites. The computational requirements for BioBin are quite modest; for example, during testing, a whole-genome analysis including 185 people took just over two hours using a single core on a cluster (Intel Xeon X5675 3.06 GHz processor). However, because the vast amount of data included in the analysis must be stored in memory, the requirements for memory usage can be high; the aforementioned whole-genome analysis required approximately 13 GB of memory to complete. Even with large datasets, BioBin can be run quickly without access to expensive and specialized computer hardware or a computing cluster. The number of rare variants is the primary driver of memory usage.

2.2.2. *Library of Knowledge Integration (LOKI) Database*

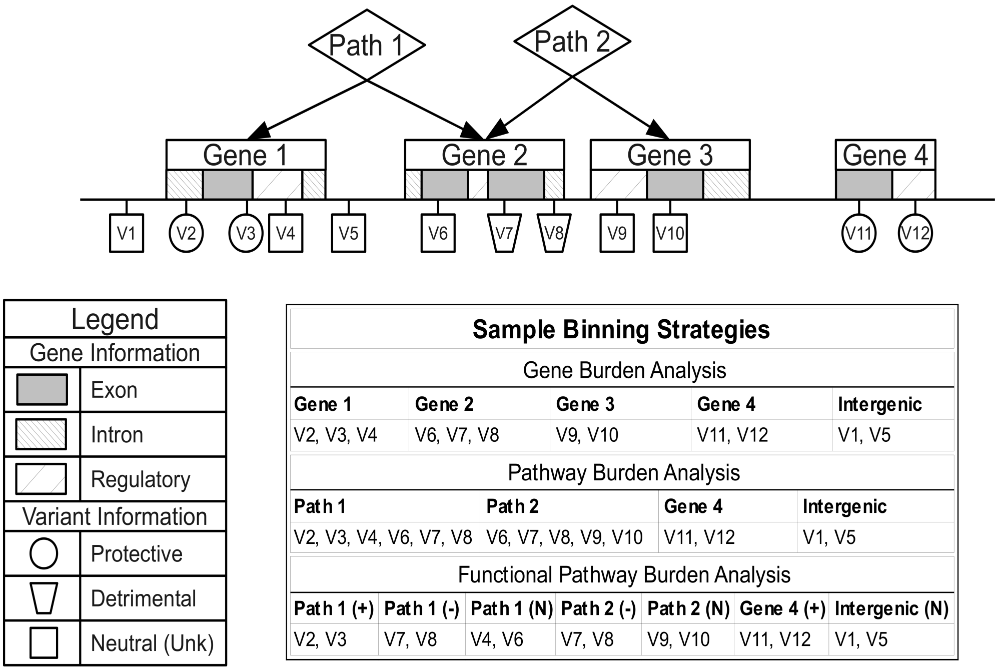
Harnessing prior biological knowledge is a powerful way to inform collapsing feature boundaries. BioBin relies on the Library of Knowledge Integration (LOKI) for database integration and boundary definitions. LOKI contains resources such as: the National Center for Biotechnology (NCBI) dbSNP and gene Entrez database information,[20] Kyoto Encyclopedia of Genes and Genomes (KEGG),[21] Reactome,[22] Gene Ontology (GO),[23] Protein families database (Pfam),[24] NetPath - signal transduction pathways,[25] Molecular INTeraction database (MINT),[26] Biological General Repository for Interaction Datasets (BioGrid),[27] Pharmacogenomics Knowledge Base (PharmGKB),[28] Open Regulatory Annotation Database (ORegAnno),[29] and information from UCSC Genome Browser about evolutionary conserved regions.[30]

LOKI is used as a means to provide a standardized interface and terminology to disparate sources each containing individual means of representing data. The three main concepts used in LOKI are *positions*, *regions* and *groups*. The term *position* refers to single nucleotide polymorphisms (SNPs), single nucleotide variants (SNVs) or RVs. The definition of *region* can be applied to a broader scope of biology. Any segment with a start and stop position can be defined as a region, including genes, copy number variants (CNVs), insertions and deletions, and evolutionary conserved regions (ECRs). *Sources* are databases (such as those listed above) that contain *groups* of interconnected information, thus organizing the data in some way.

LOKI is implemented in SQLite, a relational database management system, which does not require a dedicated database server. The user must download and run installer scripts (python) and allow for 10-12 GB of data from the various sources. The updater script will automatically process and combine this information into a single database file (~ 6.7 GB range). A system running LOKI should have at least 50 GB of disk storage available. LOKI runs locally wherever needed.

2.3. *Binning approach*

We chose NCBI dbSNP and NCBI Entrez Gene as our primary sources of position and regional information due the quality and reliability of the data, and clearly defined database schema. Intergenic regions are bins generated by BioBin to catch variants that do not fit into the user-defined feature types. For example, if one were testing RV burden differences between cases and controls across genes, all variants in genes would be collapsed into respective gene bins, and variants outside of gene boundaries would be binned corresponding to the intergenic regions. BioBin provides an option to generate intergenic bins of a user-specified size to catch intergenic variants. Figure 1 shows an example of RV binning strategies; different knowledge applied to the same variants produces alternate bins.



Legend	
Gene Information	
■	Exon
▨	Intron
◿	Regulatory
Variant Information	
○	Protective
▽	Detrimental
□	Neutral (Unk)

Sample Binning Strategies

Gene Burden Analysis

Gene 1	Gene 2	Gene 3	Gene 4	Intergenic
V2, V3, V4	V6, V7, V8	V9, V10	V11, V12	V1, V5

Pathway Burden Analysis

Path 1	Path 2	Gene 4	Intergenic
V2, V3, V4, V6, V7, V8	V6, V7, V8, V9, V10	V11, V12	V1, V5

Functional Pathway Burden Analysis

Path 1 (+)	Path 1 (-)	Path 1 (N)	Path 2 (-)	Path 2 (N)	Gene 4 (+)	Intergenic (N)
V2, V3	V7, V8	V4, V6	V7, V8	V9, V10	V11, V12	V1, V5

Figure 1. Binning strategies for three example burden analyses.

2.4. *Statistical analysis*

BioBin is a bioinformatics tool used to create new feature sets that can be analyzed in subsequent statistical analyses. We believe that statistical tests can and should be chosen according to the hypothesis being tested, the question of interest, or the type of data being tested. There are explicit situations that require the use of regression analysis (linear, logistic, polytomous), Fisher's exact test, permutation of unique statistical test, etc. For this reason, no specific statistical test is implemented into BioBin. Unless otherwise noted, the results presented herein were calculated using a Wilcoxon 2-sample rank sum test implemented in the R statistical package.[31] There was no need for adding covariates to the model and Wilcoxon provides simple implementation and interpretation. All individuals (CEU and YRI) are ranked according to number of variants they individually contribute to a bin (variants must be under binning threshold). Using a simple model, we assume the genotypes are independent and normally distributed.

2.5. *Data simulation strategy to assess type I error*

To test BioBin, genetic data was simulated using a forward time simulator, simuPOP.[32] We used a constant distribution for the selection coefficient and a mutation rate of 1.8 x 10^{-8} per nucleotide per generation. The population sizes were N_e = 8100, 8100, 7500, and 10000 with 5000 generations, 10 generations, and 370 generations respectively. A 10kb region and 50kb of genetic data were simulated using the standard parameters in the simuRareVariants.py script for simuPOP. This script simulates introduction and evolution of RVs and can allow complex fitness and selection modeling (http://simupop.sourceforge.net/cookbook/).

To generate a sample data set evaluating type I error, all individual's genotypes were generated by randomly choosing two haplotypes from a haplotype pool. This process was repeated for three different scenarios: 1) sample size of 1000 individuals (500 cases and 500 controls) on a 10kb region, 2) sample size of 4000 individuals (2000 cases and 2000 controls) on a 10kb region, 3) sample size of 4000 individuals (2000 cases and 2000 controls) on a 50kb region. Phenotypes were randomly assigned to each individual to test the null hypothesis of no association between variants and disease status. The type I error was calculated as the proportion of the 10,000 replicates with a p-value <= 0.05. In this case, an error rate above 5% would indicate a higher false-positive test and an error rate lower than 5% would indicate a conservative test.

2.6. *1000 Genomes Project data: CEU and YRI comparison*

In a recent resequencing study of 202 drug targets, Nelson et al. reported the abundance of rare variants to be approximately 1 every 17 bases and most often population specific.[15] To further investigate population stratification, we used 1000 Genomes Project data. The project was started in 2008 with the mission to provide deep characterization of variation in the human genome. As of October 2011, the sequencing project includes whole-genome sequence data for 1094 individuals, and aims to sequence 2,500 individuals by its completion.[33]

We conducted a pairwise comparison of RV burden differences between two ancestry groups (YRI and CEU) of the 1000 Genomes Project (October 2011 release ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/). The data includes 87 CEU samples and 88 YRI samples. We implemented a minimum bin size of two variants and set the binning threshold to 0.03. We performed the following feature specific analyses:

A. Gene and intergenic regions
B. Pathways
C. Regulatory regions
D. Evolutionary conserved regions (ECRs)

The NCBI Entrez source provided gene start and stop positions to form gene bin boundaries for the gene and intergenic region analyses (A). Intergenic bins (50kb) were generated to "catch" variants not collapsed into other source-informed bins; in this case, intergenic bins collapsed variants not binned into gene region bins. For the pathway-based analysis (B), pathway and group information came from many LOKI sources and collapsed variants from all genes/regions in a specific pathway together in a bin. The regulatory region analyses (C) bin boundaries used in this analysis were from ORegAnno, a database of regulatory region annotations. For the evolutionary conserved region analysis (D), boundaries were calculated from PhastCons score output downloaded from UCSC Genome Browser (http://genome.ucsc.edu/). There are three groups of ECRS available within the UCSC Genome Browser, the first group is derived from multiple alignments of 45 vertebrate genomes to the human genome, the second group is a set of placental mammals (32 placental Mammal genomes) aligned to the human genome, and the third group is a set of nine primates aligned with the human genome (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/). For each group, we calculated segments of the genome with 70% identity, a minimum length of 100bp, and allowed for 50bp gaps. These ECRs were clustered in bands according to the PhastCons output, which corresponded to an average of 13 ECRs per band. This was necessary since a single ECR is not large variable enough to generate a viable bin. In this paper, reported p-values have been corrected for multiple testing using a Bonferroni correction (number of generated bins in each analysis).

3. Results

3.1. *Type I error calculation*

It is important to investigate the level of type I error that might be present in any novel approach. Thus, using the script simuRareVariants.py from the simuPOP simulation algorithm, we simulated a 10kb genomic region with 31 RVs and 50kb genomic region with 154 RVs using the parameters described above in the methods section. Overall, 10,000 individuals were simulated, each with two haplotypes. We created populations by sampling the haplotypes, and generated 10,000 replicates of 1000 or 4000 individuals with balanced numbers of cases and controls. The threshold for significance was p≤ 0.05. We calculated the type I error rate as the number of replicates with Wilcoxon p-value less than or equal to 0.05 divided by the total number of replicates. The Wilcoxon 2-sample rank sum test seems to control the type I error in BioBin, but the false positive rate nominally increases as the sample size or bin size increases (see Table 2).

Table 2. Type I error calculation results.

Population Size	Simulated Region Size	Type I Error Rate
1000	10kb	0.0479
4000	10kb	0.0533
4000	50kb	0.0564

3.2. *1000 Genomes Project data: CEU and YRI comparison*

We tested BioBin using whole-genome ancestral data from 1000 Genomes Project using 87 CEU and 88 YRI individuals. There are considerably more variants in the YRI samples than in the CEU samples. Table 3 provides the total number of variants according to the Phase I generation of 1000 Genomes Project data for both populations. Of note, while there is only one more YRI individual compared to the number of CEU individuals, there is almost a 7 million variant difference between the two groups. Figure 2 shows a density function, which indicates the density of variants at each MAF; overall, there is a higher density of low frequency variants in YRI.

Table 1. 1000 Genomes Project Phase I data characteristics for CEU and YRI

Population	Number of Variants	Number of People
CEU	11,198,921	87
YRI	18,022,152	88

Using a MAF binning threshold of 0.03, we binned genes and intergenic regions, pathways, regulatory regions, and evolutionary conserved regions as described above in the methods. The top result from each feature in these four analyses (labeled A-D) is shown in Table 4.[34]

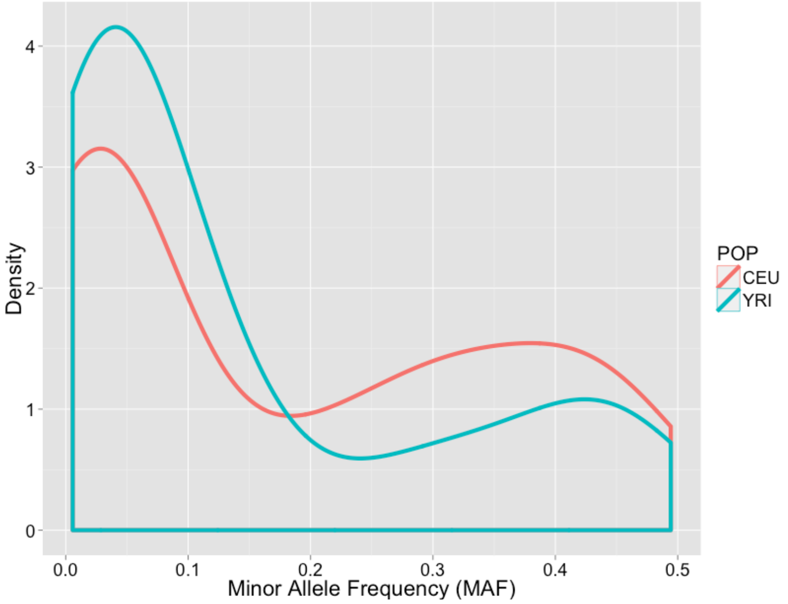


Figure 2. Minor allele frequency density distribution for CEU (red) and YRI (green)

Table 2. Top result from each feature across the four analyses (A-D)

	Feature	Top Bin	Adj. p-val	Annotation/ Location	Function
A	Genes	*CTXN2*	$7.18\text{x}10^{-29}$	Chr5:48483867-48495951	Cortexin 2-Integral to membranes
	Intergenic regions	chr15.638	$5.13\text{x}10^{-28}$	Chr15:31900000 -31950000	3' to OTUD7A, a protease that cleaves ubiquitin
B	Pathways	PF11057	$1.76\text{x}10^{-29}$	Cortexin protein family	Expressed in kidney and brain, involved in intra and extracellular signaling
C	ORegAnno	OREG0003872	$1.83\text{x}10^{-32}$	Chr5:142124712 -142125230	Transcription Factor Binding site, expressed in the heart
D	ECR-vertebrates	Chr5:33951654 -33951791	$3.24\text{x}10^{-33}$	SLC45A2	Melanocyte differentiation antigen. Substance transport for melanin biosynthesis.
	ECR-placental Mammals	Chr5:33951651 -33951791	$3.24\text{x}10^{-33}$	SLC45A2	
	ECR-primates	Chr15:48426444-48426724	$1.94\text{x}10^{-33}$	SLC24A5	Cation exchanger involved in pigmentation, melanosome ion transport

Next, we evaluated the prevalence of significant RV differences between CEU and YRI data. Using the Bonferroni corrected threshold of significance specific for each analysis, we calculated the proportion of bins that were significant. The results are shown in Figure 3.[34]

The height of each bar represents the total number of bins in each feature type; the dark blue indicates the proportion of significant bins. For example, 9.10% of the bins generated from ECR-primate multiple alignment comparison was significant after correction for multiple testing (which accounted for all tests performed in analysis D).

There are a surprising number of significant bins in each feature, but this can be explained by the difference in total number of variants between CEU and YRI. The total number of variants binned by BioBin using a MAF-binning threshold of 0.03 was 16,145,128 variants. Of these, 65.5% were private to YRI ancestral population.

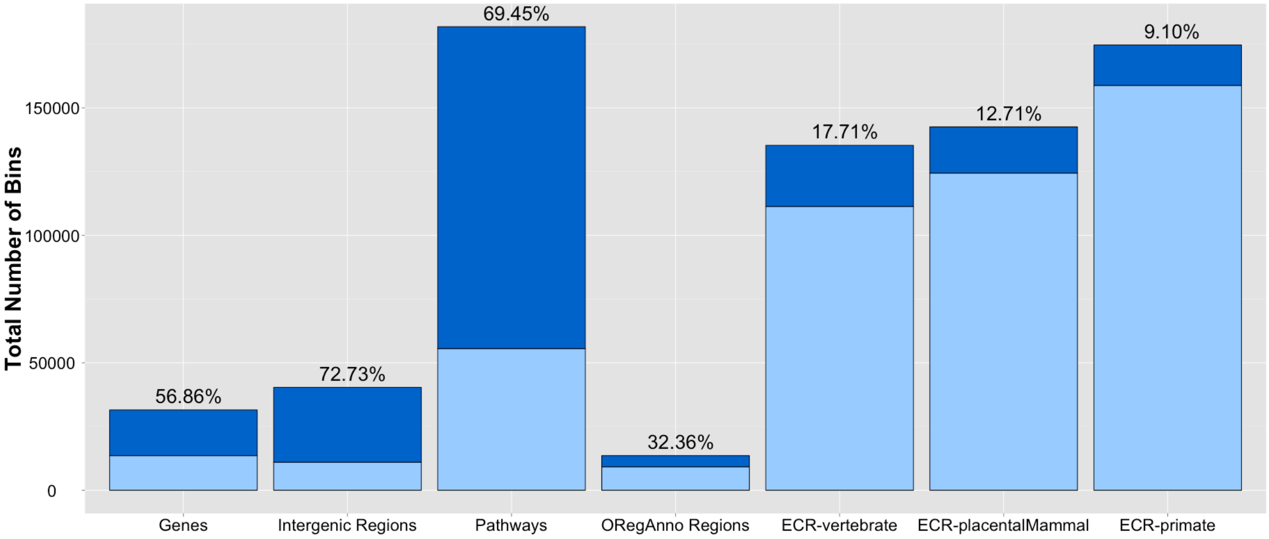


Figure 3. CEU-YRI pairwise comparison. Dark blue indicates the proportion of significant bins.

4. Discussion

4.1. *Type I error calculation*

As shown in Table 2, the Wilcoxon 2-sample rank sum test is slightly anticonservative in large population sizes and seems to worsen when more RVs are binned together. This is interesting since Li et al. reported that increasing the number of variants binned in a type I error simulation decreased the type I error rate using a collapsing approach and a Pearson χ^2 statistical test and others have reported conservative type I error rates using asymptotic statistical tests on relatively small sample sizes.[8,35]These methods were tested on simulated data with controlled RV allele frequencies and used different statistical tests, but highlights the importance and perhaps limitations of simulation testing. Although, the type I error seems to be well-controlled in this experiment, further investigation should be done to assess strictly how the RV allele frequency distribution affects type I error, calculate the type I error using additional sample population sizes and alternative statistical tests, and examine if the number of variants in a bin consistently inflate the false positive rate.

4.2. *1000 Genomes Project data: CEU and YRI comparison*

Using 1000 Genomes Project whole-genome data, we used BioBin to identify features (genes, intergenic regions, pathways, regulatory regions, and ECRs) with significant differences in rare RV burden between two ancestral populations. A population-genetics approach retains natural qualities of data (compared to simulated data) and incorporates case/control status according to ancestry group. Comparable approaches have been used by other groups.[9,17]

We compared multiple feature types between two ancestral populations from 1000 Genomes Project to highlight a known issue in genomic studies, population stratification. BioBin explored RV burden differences between CEU and YRI ancestral populations. In each RV burden test, there were a considerable number of statistically significant bins (after Bonferroni multiple testing correction). Table 4 shows the most significant bins for each feature type. The gene burden top result and the pathway burden top result corresponded to a Cortexin-2 gene and Cortexin pathway respectively. According to PFAM, this group of proteins is important for intracellular and extracellular signaling in the kidney and brain (http://pfam.sanger.ac.uk/family/PF11057). To our knowledge, Cortexin-2 has not been acknowledged in ancestry comparison studies. However, another protein in the Cortexin family was identified as a candidate gene for non-diabetic forms of end-stage renal disease in African Americans.[36] This is interesting since studies with admixed populations could contain a higher incidence of false positives due to RV population stratification and mixed ancestry.

We could not find biological interpretation for the significant intergenic RV burden differences on chromosome 15 or the transcription factor-binding site on chromosome 5. However, the ECR analyses highlighted *SLC45A2* and *SLC24A5*; both participate in pigmentation.

Mutation rates vary across the genome. They can vary according to specific sequence contexts, within regions on a chromosome, and between chromosomes.[37] While mutation rates are commonly studied between orthologous sequences, polymorphisms collapsed by regions within species can also provide interesting insight into evolutionary history and mutation. BioBin does not provide detailed sequence output to investigate mutation rate variation between CEU and YRI, but it does provide some information about higher rates of variation in regions (genes, intergenic regions, pathways, regulatory regions, and ECRs) and between chromosomes. The results in Figure 3 show an interesting trend between functional regions of the genome and variant tolerance. Approximately 57% of the gene bins had significant differences in RV burden, whereas approximately 73% of the intergenic region bins had significant differences in RV burden. There is some weak evidence that genes undergo adaptive evolution, which explains why regions in the genome with potential for highly deleterious mutations evolve lower mutation rates. There are two potential explanations: 1) additional level of repair of DNA damage in transcriptional active regions by transcription coupled repair (TCR), 2) approximately 3% of the genome is subject to negative selection, however it is estimated that functionally dense regions contain up to 20% sites under selection.[37,38] In this analysis, gene bins are inclusive of intronic regions, thus it would be interesting to break down the gene bins into intronic and exonic bins to see how the variant tolerance differs between coding and noncoding regions.

There are far fewer regulatory region bins, but there appears to be smaller proportion of significant differences between CEU and YRI compared to genes or intergenic regions. Again, perhaps mutations are less tolerated in these regions and we see overall less variability. ECRs have been long known to be conserved among species, and in this analysis they are also the features least likely to have variation between CEU and YRI. There is some debate about selection and functional significance in these regions. It is unknown what factors have the largest effect on mutation rates,[37] but it is possible that consistently low mutation rates in these sections have generated conserved regions throughout evolution.[38]

We found that over 65% of the variant loci in dataset were fixed in CEU individuals. This is not surprising since it is well known that individuals of African descent have more variation than individuals of other ancestral groups (see Table 3). This difference in rare variation is driving the high percentages seen in Figure 3. We should further investigate the effects of stratification in other ethnicities, and evaluate correction methods such as PCA and mixed models.[18,19]

5. Conclusion

There is a global health, scientific, and financial motivation for understanding the genetic etiology of common complex disease. It is imperative to consider genetic variants beyond common single nucleotide polymorphisms, as RVs may have larger phenotypic effects and can help us better comprehend the biology of a disease process. BioBin is a novel collapsing method that uses allele frequency data and biological information to bin RVs. It is unique because it is packaged with LOKI and is not coupled with any statistical method. Access to integrated biological knowledge (pathways, groups, interactions, ECRs, regulatory regions, etc.) is valuable to researchers that do not want to spend considerable effort to combine this knowledge manually. Freedom from implemented statistical methods provides users with the ability to apply association tests most appropriate for their data analysis. In general, for any given bin, statistical tests from other published collapsing methods can be applied to BioBin output. However, these other methods do not incorporate feature selection; therefore, the user must provide boundaries for each bin.

In this paper, we evaluated RV burden differences between CEU and YRI populations. Although population stratification is often considered in genomic analyses, to our knowledge, no previous studies have quantified the magnitude of RV burden differences across multiple features. From the ancestry comparison results, we learned RV burden differences among features showed a pattern consistent with current mutation rate theory but also highlighted the magnitude of RV stratification between CEU and YRI populations from 1000 Genomes Project data.

In summary, our results suggest that BioBin will be a useful tool to analyze sequence data. While no one can unequivocally guess the role RVs will play in uncovering hidden heritability for common complex disease, it seems that testing them in aggregate can provide valuable knowledge about the biology. Prerequisites for installation and running of BioBin and LOKI are documented in the manual, which is publicly available with the software and example statistical association scripts in R at https://ritchielab.psu.edu/ritchielab/software.

References

1. Johansen, C. T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet* **42**, 684–687 (2010).
2. Bhatia, G. *et al.* A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS.Comput Biol* **6**, e1000954 (2010).
3. Ionita-Laza, I., Buxbaum, J. D., Laird, N. M. & Lange, C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS.Genet* **7**, e1001289 (2011).
4. Haack, T. B. *et al.* Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. *Nat Genet* **42**, 1131–1134 (2010).
5. Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics* **42**, 790–793 (2010).
6. Raychaudhuri, S. *et al.* A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. *Nat Genet* **43**, 1232–1236 (2011).
7. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* **615**, 28–56 (2007).
8. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311–321 (2008).
9. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS.Genet* **5**, e1000384 (2009).
10. Han, F. & Pan, W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* **70**, 42–54 (2010).
11. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* **86**, 832–838 (2010).
12. Hoffmann, T. J., Marini, N. J. & Witte, J. S. Comprehensive approach to analyzing rare genetic variants. *PLoS.One.* **5**, e13584 (2010).
13. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
14. Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res* **21**, 1529–1542 (2011).
15. Nelson, M. R. *et al.* An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* **337**, 100–104 (2012).
16. Tennessen, J. A. *et al.* Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* **337**, 64–69 (2012).
17. Zhang, L., Pei, Y.-F., Li, J., Papasian, C. J. & Deng, H.-W. Efficient utilization of rare variants for detection of disease-related genomic regions. *PLoS ONE* **5**, e14288 (2010).
18. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459–463 (2010).
19. He, H. *et al.* Effect of population stratification analysis on false-positive rates for common and rare variants. *BMC Proceedings* **5**, S116 (2011).
20. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **39**, D38–D51 (2010).
21. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* **40**, D109–D114 (2011).
22. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* **39**, D691–D697 (2010).
23. Dimmer, E. C. *et al.* The UniProt-GO Annotation database in 2011. *Nucleic Acids Research* **40**, D565–D570 (2011).
24. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Research* **40**, D290–D301 (2011).
25. Kandasamy, K. *et al.* NetPath: a public resource of curated signal transduction pathways. *Genome Biol.* **11**, R3 (2010).
26. Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–861 (2012).
27. Stark, C. *et al.* The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.* **39**, D698–704 (2011).
28. McDonagh, E. M., Whirl-Carrillo, M., Garten, Y., Altman, R. B. & Klein, T. E. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark Med* **5**, 795–806 (2011).
29. Griffith, O. L. *et al.* ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Research* **36**, D107–D113 (2007).
30. Fujita, P. A. *et al.* The UCSC Genome Browser database: update 2011. *Nucl. Acids Res.* (2010).doi:10.1093/nar/gkq963
31. R Development Core Team *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing: Vienna, Austria, 2011).at <http://www.R-project.org>
32. Peng, B., Amos, C. I. & Kimmel, M. Forward-time simulations of human populations with complex diseases. *PLoS Genet.* **3**, e47 (2007).
33. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
34. Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer New York: 2009).at <http://had.co.nz/ggplot2/book>
35. Daye, Z. J., Li, H. & Wei, Z. A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic Acids Res.* **40**, e60 (2012).
36. Bostrom, M. *et al.* Candidate genes for non-diabetic ESRD in African Americans: a genome-wide association study using pooled DNA. *Human Genetics* **128**, 195–204 (2010).
37. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics* **12**, 756–766 (2011).
38. Ellegren, H., Smith, N. G. & Webster, M. T. Mutation rate variation in the mammalian genome. *Current Opinion in Genetics & Development* **13**, 562–568 (2003).

DETECTING HIGHLY DIFFERENTIATED COPY-NUMBER VARIANTS FROM POOLED POPULATION SEQUENCING

DANIEL R. SCHRIDER[*]

Department of Biology and School of Informatics and Computing, Indiana University, 1001 E Third St.
Bloomington, IN 47405, USA
Email: dschride@indiana.edu

DAVID J BEGUN[†]

Department of Evolution and Ecology, University of California, 3350A Storer Hall
Davis, CA 95616, USA
Email: djbegun@ucdavis.edu

MATTHEW W HAHN[‡]

Department of Biology and School of Informatics and Computing, Indiana University, 1001 E Third St.
Bloomington, IN 47405, USA
Email: mwh@indiana.edu

Copy-number variants (CNVs) represent a functionally and evolutionarily important class of variation. Here we take advantage of the use of pooled sequencing to detect CNVs with large differences in allele frequency between population samples. We present a method for detecting CNVs in pooled population samples using a combination of paired-end sequences and read-depth. Highly differentiated CNVs show large differences in the number of paired-end reads supporting individual alleles and large differences in read-depth between population samples. We complement this approach with one that uses a hidden Markov model to find larger regions differing in read-depth between samples. Using novel pooled sequence data from two populations of *Drosophila melanogaster* along a latitudinal cline, we demonstrate the utility of our method for identifying CNVs involved in local adaptation.

[*] D.R.S is supported by the Indiana University Genetics, Molecular and Cellular Sciences Training Grant T32-GM007757.

[†] D.J.B. is supported by National Institutes of Health grant GM084056.

[‡] M.W.H. is supported by National Science Foundation grant DBI-0845494.

1. Introduction

Technological advancements over the last two decades have given researchers the ability to efficiently and accurately search for genetic differences between individuals across entire genomes. While initial efforts identified millions of single nucleotide polymorphisms (SNPs) within human populations [1], it was soon discovered that any two individuals also differ in copy-number at many large genomic regions encompassing whole genes or parts of genes [2]. In recent years, many more such copy-number variants (CNVs) have been detected using microarray hybridization intensity or sequence read-depth [3-5], paired-end/mate-pair sequencing [6,7], or both types of evidence together [8,9]. These CNVs are ubiquitous across eukaryotes, with large numbers also present in chimpanzees [10], mice [11], *Arabidopsis* [4], fruit flies [12,13], yeast [14], and many other species. These polymorphisms have attracted a great deal of attention because their large size suggests that they could have a considerable functional impact [2]. This hypothesis has been borne out by the large number of CNVs found to cause or increase the risk of various diseases in humans, including autism [15], schizophrenia [16], Charcot–Marie–Tooth disease [17], Crohn's disease [18], and Parkinson's [19].

The availability of large population genomic data sets has also allowed for genome-wide tests of recent and ongoing adaptive natural selection on CNVs, especially in humans [20,21]. Methods that detect signatures of adaptive evolution—such as long haplotypes with reduced nucleotide diversity [22,23] or large differences in allele frequencies across selective environments [24]—have been used to provide evidence for selection on CNVs [3,5]. These and other studies have identified a large number of putatively adaptive CNVs (>100 listed in [25]), with fitness benefits ranging from improved digestion of starches [26] to reduced HIV susceptibility [27]. Thus, CNVs appear to be an important source of adaptive as well as deleterious mutations in humans, and likely in other organisms as well.

A cost-effective and powerful way to detect variants with large differences in allele frequencies among populations is to pool and sequence large numbers of individuals from each of several populations using next-generation sequencing technologies (e.g. refs. [28,29]). While this pooling approach does not provide information on individual haplotypes, it can be used to accurately estimate other important population-genetic parameters [30,31]. Such an approach is also very effective at detecting locally adapted alleles when interbreeding between the sampled populations is frequent, as allele frequencies at neutral loci (i.e. those not affected by spatially varying selection) are homogenized very rapidly and linkage disequilibrium between selected polymorphisms and nearby variants is quickly reduced [32]. This is an ideal scenario for identifying polymorphisms that facilitate adaptation to local environments because one allele is favored by natural selection in one population and another allele is favored in other populations.

Searching for CNVs with a greater-than-expected difference in allele frequencies among populations has been particularly effective at identifying candidate adaptive CNVs in humans [25]. However, previous uses of pooled population sequences to detect highly differentiated CNVs has been limited to procedures examining the tails of distributions of differences in read-depth between two populations [28]. In this paper, we describe a principled method leveraging pooled sequencing data from two geographically dispersed but interbreeding populations in order to detect differentiated CNVs with high resolution. This method combines information from paired-end reads with a hidden Markov model designed to detect large differences in read-depth. We demonstrate the utility of this method on data from *Drosophila melanogaster* sampled from opposite ends of a latitudinal cline along the East Coast of the United States. Our search identifies 140 CNVs with the most highly differentiated allele frequencies between the two ends of the cline. These data and functional enrichment analyses strongly suggest that many of these CNVs are experiencing spatially varying selection.

2. Method for detecting CNVs with differentiated allele frequencies

2.1. *Method overview*

Here we describe in detail a method that leverages pooled paired-end next-generation sequence data to detect CNVs that differ substantially in allele frequency between two populations; in the data presented below these populations are at opposite ends of an environmental cline. Briefly, this method begins by searching for read pairs suggestive of deletions or tandem duplications. These read pairs are then combined into single putative events using a simple clustering algorithm, and putative events that are supported by many more read pairs from one population than the other are retained. For each of these putative duplications (Figure 1A) or deletions (Figure 1B), read-depth across the entire CNV region is examined in both populations, and those with a significant difference in read-depth between the two ends are considered CNV candidates for spatially varying selection. Because this paired-end approach may not detect all CNVs, we also use a hidden Markov model (HMM) to detect CNVs with differentiated allele frequencies based on read-depth alone. Both of these approaches will fail to detect CNVs segregating at low frequencies, depending on the depth of coverage to which pooled samples were sequenced.

2.2. *Capturing and sequencing of* D. melanogaster *samples for test data*

Our method was tested on sequence data from two pooled DNA samples of *Drosophila melanogaster* from opposite ends of a latitudinal cline in temperature and seasonality along the East Coast of the United States. 36 flies were captured from Bowdoinham, Maine and 30 were captured from Homestead, Florida. DNA was extracted from these two populations and pooled into two samples that were each sequenced using the Illumina Genome Analyzer IIx using short-insert paired-end libraries (~244 bp for Maine and ~213 bp for Florida on average) with 72 bp reads on each end. 60,730,268 read pairs were sequenced from the Maine sample and 48,771,223 were sequenced from Florida. After mapping to the euchromatic portion of the genome, the Maine data had an average read-depth of 49.39X and Florida had an average depth of 38.10X.

2.3. *Detecting differentiated CNVs using paired-end sequencing*

2.3.1. *Mapping reads and detecting paired-ends indicative of CNVs*

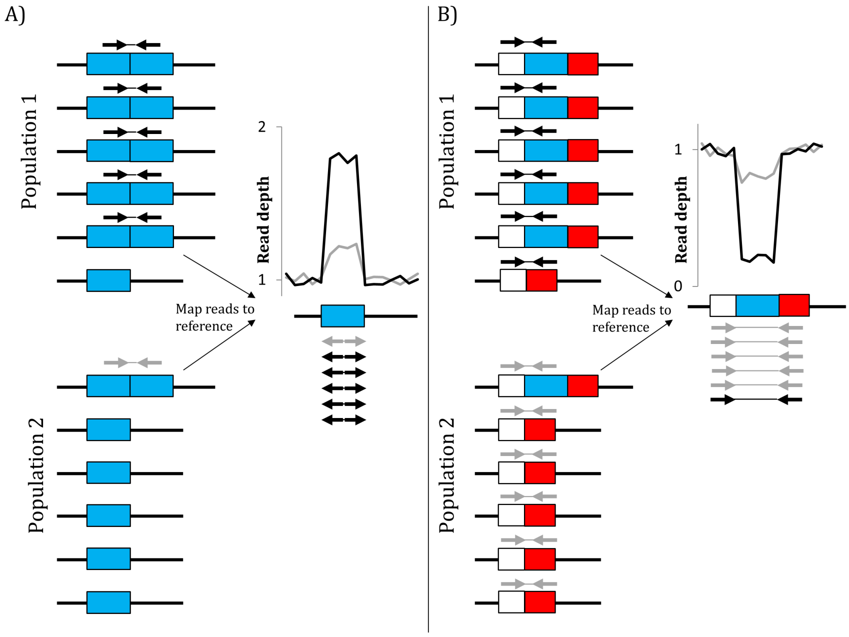


Figure 1: Detecting CNVs differing in allele frequency between two populations each sequenced from pooled DNA. A) Paired-ends spanning a tandem duplication are mapped to the reference genome in "everted" orientation. Thus, if one population has many more such reads than the other, then the duplication likely differs in allele frequency between populations (e.g., the excess of black versus grey reads in this example). Read-depth within the duplicated locus would also be much higher in the population in which the duplication appears at higher frequency (black line versus grey line). B) When paired-ends from a region with a deletion are mapped to the reference, those from chromosomes with the deletion allele that span the breakpoints of the deletion are mapped much further apart than expected. A deletion differing in allele frequency between two populations will exhibit far more of these read-pairs and lower read-depth (black) within the deletion in one population than the other (grey).

Our first method to detect CNVs under spatially varying selection leverages paired-end reads mapped in a manner suggestive of a duplication or deletion relative to the reference genome. In order to find such read pairs, we first map reads to the reference genome. In the case of our test data, we used BWA [33] to map reads to release 5 of the *D. melanogaster* genome assembly [34]. Whenever a read can be mapped equally well to multiple locations, and the mapping location(s) of the other read in the pair do not resolve this ambiguity, BWA randomly selects one of these locations to place the read. Other strategies, such as mapping the read to each location or discarding the read, are also compatible with the approach we describe here. In order to detect putative duplications, we then search for read pairs mapping to the same chromosome but in “everted” orientation, where the two reads are oriented away from one another rather than toward one another as expected. Such read pairs are expected in the case of head-to-tail tandem duplications, when read pairs crossing from the end of the first copy into the beginning of the second copy can only map to the first copy, to the left of the first read in the pair (Figure 1A; also illustrated in Figure 3A from ref. [35]). While this approach will only successfully identify tandem duplications in head-to-tail orientation, the vast majority of duplication polymorphisms in *D.*

melanogaster meet these criteria [12,13]. For other species in which an appreciable number of duplication polymorphisms are non-tandem—such as humans [36]—this method would need to be extended to detect signatures of other types of duplications. Paired-end reads indicative of deletions relative to the reference are simply those that map in proper orientation but further apart than expected given the distribution of insert sizes [6]. We consider the most extreme 1% of read-pairs with respect to mapping distance from one another as potentially indicative of deletions. Discordantly-mapped paired-ends can also be used to detect inversions [6]: indeed, using an approach analogous to that described here we find that at least two known large inversions, *In(2L)t* and *In(3L)P*, are differentiated along the East Coast of the United States in *Drosophila* (data not shown).

2.3.2. *Clustering discordant read-pairs*

Sequenced read-pairs signifying copy-number variants (referred to as discordant read-pairs or discordant inserts) were clustered into distinct putative polymorphisms using a simple greedy clustering algorithm. Briefly, each discordant insert is initially assigned to its own cluster, and then pairs of clusters are examined and merged into a single cluster if any pair of inserts, one from each cluster, meets the following two criteria: 1) The coordinates of the reads must differ by no more than the largest expected insert-size (350 bp for our data); 2) For deletions, the inferred insert-sizes of the two inserts must differ by no more than the typical difference between insert-sizes from the sequenced library (200 bp for our data). If these criteria are met, then it is possible that the two inserts support the same mutation. This process is repeated until no more clusters can be merged. All clusters containing only a single discordant read-pair were ignored. When clustering putative deletions with our *Drosophila* data, there were a large number of clusters containing only a single insert. This is because we sequenced a large number of inserts, resulting in hundreds of thousands of inserts in the upper 1% tail with respect to insert-size that are likely not the result of deletions. Because a normal insert misclassified as discordant may therefore appear near a deletion supported by true discordant inserts, when examining any pair of deletion-supporting insert clusters $<C_1,C_2>$ for overlap, we merged C_1 and C_2 only if 75% of all possible pairs of inserts (one insert from C_1 and one insert from C_2) met the merging criteria. Finally, after clustering was completed, any inserts within a cluster not appearing to support the putative mutation (according to the merging criteria) were removed from the cluster. Clustering is performed separately for each population.

2.3.3. *Detecting highly differentiated CNVs*

Once putative CNVs have been identified by clustering discordant paired-ends, we next search for polymorphisms differing in allele frequency between the two populations. This is done by first determining whether each polymorphism (i.e. each cluster of discordant inserts) present in one sample is present in the other using the merging criteria described in the previous section, and then comparing the numbers of distinct inserts supporting the event in each population; this number is zero when the event is not detected in a population. To reduce the number of false positive CNVs, we then removed all deletions supported by fewer than four paired-end reads. For each CNV, we calculate the difference between the number of inserts (after correcting for the total number of read pairs mapped to the reference genome from each sample) supporting the mutation in each cline-end, and retain all CNVs with a large difference (i.e., in either of the 5% tails of the empirical distribution of insert-number differences for deletions or 2.5% for deletions, where more stringency is required). For these events we then counted in each sample the total number of read pairs mapped in proper orientation and within the expected distance range, correcting for differences in average read-depth across the euchromatic genome.

The ratio of these corrected read-depths is then used as an independent mode of confirmation of highly differentiated CNVs detected by paired-ends. This was done in the *D. melanogaster* data by selecting roughly 10,000 random genomic regions of various lengths and calculating the read-depth ratios between the populations for these regions, and then calculating the 5% tails for each length. A CNV was considered confirmed as differentiated by read-depth if the depth ratio was biased in the same direction as the numbers of paired-ends in the two samples, and more extreme than 95% of read-depth ratios within random genomic regions of approximately the same length.

2.4. *Detecting differentiated CNVs from read-depth using a hidden Markov model*

The method described above requires both discordant paired-ends and read-depth to show evidence of differentiation in allele frequency between cline-ends. Because the number of discordant-paired ends may be small due to chance, some events with biologically significant differences in allele frequency may be missed. Dispersed duplications will also be missed because we only examine paired-ends supporting tandem duplications. We therefore complement the above approach by using a hidden Markov model (HMM) to detect differentiated CNVs based on read-depth alone. The observations for this HMM are, for small adjacent windows, the ratios of the (corrected) number of mapped paired-ends from one sample within the window over the (corrected) number of mapped paired-ends from the other sample in the same window. When testing this approach on our pooled *D. melanogaster* data, we binned these ratios to produce discrete observation symbols, though continuous distributions can also be used. The HMM has three states: no difference in read-depth between pooled samples, higher read-depth in sample 1, and higher read-depth in sample 2. The initial, transition, and emission probability matrices, which model the density and length of differentiated CNVs, and the distribution of read-depth ratios appearing in each of the hidden states, can be trained in a supervised manner using results from the paired-end based method described above. Once the HMM has been trained, the genome can be segmented into regions that have higher copy-number in sample 1, higher copy-number in sample 2, or similar copy-number in both samples, using the Viterbi algorithm.

2.5. *Estimating allele frequencies*

The methods outlined above detect strongly differentiated polymorphisms not by using allele frequency estimates directly, but by searching for differences in the numbers of mapped inserts between the two populations. However, for many applications researchers may wish to estimate actual allele frequencies at putatively differentiated CNVs. For duplications relative to the reference genome, this is given by:

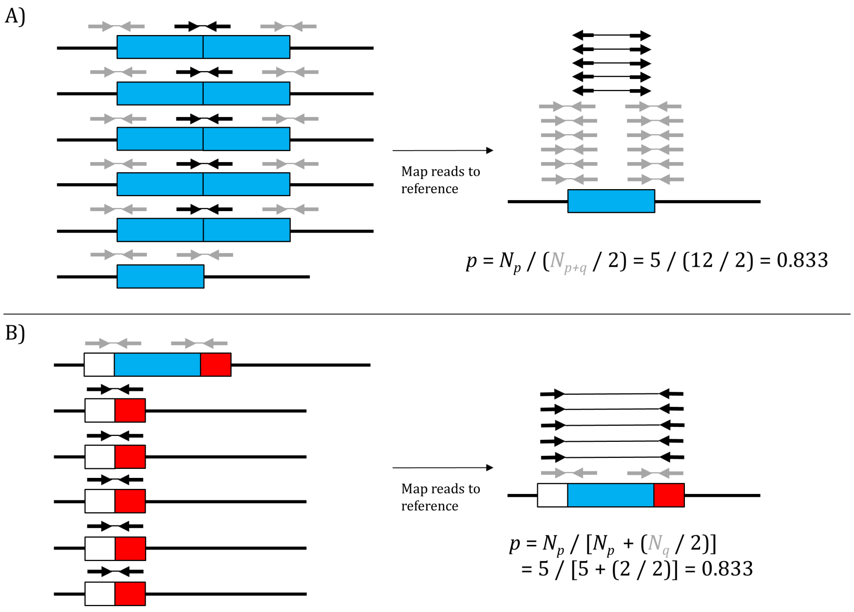

A)
Map reads to reference
$p = N_p / (N_{p+q} / 2) = 5 / (12 / 2) = 0.833$
B)
Map reads to reference
$p = N_p / [N_p + (N_q / 2)]$
$= 5 / [5 + (2 / 2)] = 0.833$

Figure 2: Estimating allele frequencies of CNVs from pooled DNA sequence data. A) For duplications, read pairs having one read flanking and one read within the duplicated locus (grey) are contributed from all chromosomes, while read-pairs mapped to the reference in "everted" orientation (black) are only sequenced from chromosomes with the duplication. B) For deletions, read pairs having one read flanking the deleted region and one read within the region (grey) are contributed only from chromosomes lacking the deletion, while read pairs spanning the entire deleted locus (black) are only derived from chromosomes with the deletion. Allele frequencies are estimated as shown in the examples using Equations 1 and 2, respectively.

$$p = N_p / (N_{p+q} / 2), \tag{1}$$

where N_p is the number of “everted” inserts supporting the duplications, and N_{p+q} is the total number of inserts mapping across either of the breakpoints and not supporting the duplication (as both chromosomes with and without the duplication will have this sequence; Figure 2A). For deletions relative to the reference genome,

$$p = N_p / [N_p + (N_q / 2)], \tag{2}$$

where N_p is the number of inserts mapping across the deletion breakpoints and supporting the deletion, and N_q is the total number of inserts mapping across either of the breakpoints (Figure 2B). In order to measure population differentiation, F_{ST} [37] can then be calculated using allele frequency estimates from the two samples.

Because these allele frequency estimates may be calculated from a small number of inserts they can be quite noisy, and it may be preferable to estimate allele frequencies in the two populations from read-depth across the entire CNV. This can be done by modeling expected read-depth across the genome as a response to variables such as GC composition [13,38] and density of

SNPs and indels [13]. We do not describe this approach in detail here, but instead refer readers to Appendix B of ref. [13] for a more detailed discussion of modeling expected read depths in order to detect copy-number variation. Once this is done, the allele frequency of biallelic CNVs (those with only one copy polymorphic for presence/absence across individuals) is simply the percent excess or deficit of reads compared to the expectation in a given sample.

3. Assessing the utility of the method using *D. melanogaster* data from a latitudinal cline

3.1. *Differentiated CNVs called from paired-ends and confirmed by read-depth*

We assessed the utility of the method described above by searching for highly differentiated CNVs in *Drosophila melanogaster* along the East Coast of the United States. We pooled and sequenced DNA samples from Maine and Florida as described in Section 2.2, and then mapped paired-ends to the reference genome and clustered paired-ends indicative of deletions or tandem duplications (Sections 2.3.1 and 2.3.2), referred to as discordant paired-ends. We identified 9,428 putative deletions and 1,829 duplications relative to the reference genome. While the number of duplications is similar to those discovered in a recent examination of 37 whole-genome sequences of fruit flies captured from Raleigh, NC (2,588 duplications in [13]), we find substantially more deletions (only 3,336 in [13]), perhaps due to a large number of false positives in our set. 1,114 deletions and 129 duplications showed a large difference in the number of discordant paired-ends supporting the event in the two samples; we also calculated average read-depth across these putative CNVs in each sample. 102 deletions and 29 duplications were confirmed as differentiated CNVs by read-depth (Section 2.2.3), far more than the 5% expected by chance given our confirmation cutoffs (P=1.95x10^{-10} for deletions; P<2.2x10^{-16} for duplications; χ^2 tests). In order to further assess the accuracy of this approach, we asked how many of these CNVs were found in a recent examination of 37 whole-genome sequences from Raleigh, North Carolina [13]. Because CNVs in these Raleigh genomes were detected from read-depth alone, an approach that has lower sensitivity to detect smaller variants, we examined only events >500 base pairs in length. We find that 31 of 52 (59.6%) of our differentiated CNVs >500 bp in length are also found in the Raleigh genomes (based on mutual 50% overlap with events in [13], or with paired-end sequences collected from two of these genomes indicative of CNVs). This is likely a substantial underestimate of our accuracy because some highly differentiated CNVs may have very low frequency in Raleigh, and suggests that most false CNV calls are removed by our two complementary tests for differentiated allele frequencies. Furthermore, the high correlation between the difference in number of paired-ends supporting a CNV in each sample and the ratio of read-depths between the two samples (ρ=0.786; P<2.2x10^{-16}) suggests that both of these independent measures are estimates of allele frequency differentiation, and that we are accurately detecting differentiated CNVs. The average lengths of these deletions and duplications relative to the reference genome were 1,985 and 5,506 bp, respectively. Larger duplications than deletions have been observed previously in polymorphism data [13].

Because this method is designed to use pooled data, with the goal of finding polymorphisms differentiated across pooled samples, its performance cannot be compared to previous CNV-detection methods (e.g., [7-9,39-41]) in a straightforward manner. Indeed, to our knowledge the only existing method for detecting structural variants from pooled data is designed to detect transposable element insertion polymorphisms [42]. Thus, while many existing methods for detecting CNVs could be extended to the problem we address here, a comparison of the effectiveness of these methods with ours is beyond the scope of this paper.

3.2. *Differentiated CNVs detected from read-depth using an HMM*

Because the discordant paired-end approach may not detect all differentiated CNVs (Section 2.4), we supplemented this search using a hidden Markov model (HMM) examining only read-depth. Briefly, this HMM was used to segment the genome into three hidden states: differentiated CNVs with higher copy-number (and substantially higher read-depth) in Maine (State 1), regions with no differentiated CNVs (approximately equal read-depth; State 2), and differentiated CNVs with higher copy-number in Florida (State 3). It should be noted that this approach can only identify regions that differ in copy-number between the populations, and which population has higher copy-number—it does not determine whether the CNV is a duplication or deletion relative to the reference genome. It also does not detect regions with CNVs that do not differ in frequency between the two pools, unlike the paired-end method.

Observations for the HMM were ratios of read-depths of the two samples (Florida:Maine) in 100 bp windows, binned into one of the following categories of ratios: [0, 0.67), [0.67, 0.8), [0.8, 1), [1, 1.25), [1.25, 1.5), [1.5, ∞). We estimated the initial probabilities (the probability of the first genomic window lying within a differentiated CNV or not), transition probabilities modeling the average length of differentiated CNVs and the average distance between them, and emission probabilities (modeling the distribution of Florida:Maine read-depth ratios both within and outside of differentiated CNVs) from the properties of differentiated CNVs detected via the discordant paired-end method, yielding the following vectors/matrices (with minor manual adjustment based on prior expectations):

Initial probabilities, $\Pi = [\pi_3 \ = \ 0.005 \quad \pi_3 \ = \ 0.99 \quad \pi_3 \ = \ 0.005]$

Transition probabilities, $$\Phi = \begin{bmatrix} \varphi_{1,1} = 0.90025 & \varphi_{1,2} = 0.0995 & \varphi_{1,3} = 0.00025 \\ \varphi_{2,1} = 0.90005 & \varphi_{2,2} = 0.9999 & \varphi_{2,3} = 0.00005 \\ \varphi_{3,1} = 0.00025 & \varphi_{3,2} = 0.0995 & \varphi_{3,3} = 0.90025 \end{bmatrix}$$

Emission probabilities, $\Theta =$
$$\begin{bmatrix} \theta_{1,<0.67} = 0.025 & \theta_{1,<0.8} = 0.06 & \theta_{1,<1.0} = 0.11 & \theta_{1,<1.25} = 0.15 & \theta_{1,<1.5} = 0.36 & \theta_{1,<\infty} = 0.30 \\ \theta_{2,<0.67} = 0.15 & \theta_{2,<0.8} = 0.12 & \theta_{2,<1.0} = 0.21 & \theta_{2,<1.25} = 0.27 & \theta_{2,<1.5} = 0.13 & \theta_{2,<\infty} = 0.12 \\ \theta_{3,<0.67} = 0.28 & \theta_{3,<0.8} = 0.10 & \theta_{3,<1.0} = 0.30 & \theta_{3,<1.25} = 0.12 & \theta_{3,<1.5} = 0.06 & \theta_{3,<\infty} = 0.14 \end{bmatrix}$$

In order to infer the most likely sequence of hidden states across the genome, we ran the Viterbi and traceback algorithms on windowed Florida:Maine read-depth ratios, finding 11 highly differentiated CNVs, or stretches of genomic sequence assigned to either State 1 (elevated copy-number in Maine) or State 3 (elevated in Florida), with an average length of 5,776 bp. We found that two of these calls were also detected using paired-ends, implying that the remaining nine are either highly differentiated CNVs that the paired-end approach failed to detect or false positives. Of these nine CNVs, two were identified in previous analyses of highly differentiated genomic regions (containing *Cyp12d*, and *Ace*, respectively; Turner et al. 2008), and another two were detected in the Raleigh genomes [13]. Thus, these events may be true positives, underscoring the complementarity of the HMM approach to the paired-end approach discussed above.

3.3. *Evidence of natural selection acting on differentiated CNVs*

Although we have several lines of evidence that the vast majority of the putatively differentiated CNVs described in the sections above are true variants, additional evidence is required to show that these CNVs are not evolving neutrally with respect to the environmental cline. We estimated F_{ST} and found that many of these CNVs have high estimates (30 have $F_{ST} > 0.2$), but we cannot assess the significance of this given that the neutral expectation for our data is unknown (synonymous SNPs may be linked to nearby selected polymorphisms), and our F_{ST} estimates may have considerable variance. Thus, the best way to evaluate the impact of natural selection on these CNVs may be to search for enrichment of certain genes and functional categories. We noticed that several CNVs contain complete or partial cytochrome P450 genes, including *Cyp28d2*, *Cyp12d1-p*, *Cyp12d1-d*, *Cyp6a17*, *Cyp6a22*, *Cyp6a23*, *Cyp12c1*, *Cyp313a4*, and *Cyp12a4*. This is more than are expected by chance (P=0.0032; permutation test of the 140 most differentiated CNVs; P<0.0001 when testing for an excess of CNVs containing at least one *Cyp* gene—both of these tests control for spatial clustering of related genes). Members of this superfamily are often involved in insecticide resistance [43], and overexpression of *Cyp12d1* [44] and *Cyp12a4* [45] have been shown to increase insecticide resistance; this selective pressure may therefore be the cause of the extensive differentiation seen at these genes. In addition, *Cyp6a17* has been shown to affect temperature preference [46], implying that insecticide resistance may not be the only geographically dependent fitness effect conferred by cytochrome P450s in *Drosophila*. *Ace* (acetylcholinesterase), another gene involved in insecticide resistance [47] and previously identified as lying partially within a CNV differentiated along this cline [48], was also found in our analysis.

We searched for overrepresented Gene Ontology (GO) terms associated with genes lying within differentiated CNVs, using the hypergeometric distribution as our null hypothesis for each term. It is important to note that GO enrichment analyses conducted on CNVs or other large regions can be biased away from the null distribution by the clustering of functionally related genes [48,49]. We therefore allowed each term to be counted at most once per CNV before calculating significance. In order to correct for multiple testing, we calculated the false discovery rate (FDR) following ref. [50]. We identified a large number of terms with FDR <0.1, including 67 biological process terms. In addition to terms involved with response to pesticides (e.g., response to organophosphorus, response to carbamate), this set of enriched terms included several related to neuronal development and activity, including regulation of short-term neuronal synaptic plasticity, synaptic transmission, and synaptic target attraction. The enrichment of these terms is not driven by the overrepresented insecticide resistance genes discussed above, suggesting that CNVs confer distinct spatially dependent fitness benefits related to nervous system development and insecticide tolerance.

The overrepresentation of the functional categories listed above lends further confidence to our assertion that a substantial fraction of the CNVs detected by the method described here are under spatially varying selection. Although the results of this type of enrichment analysis should not be taken as proof of the action of natural selection [49], they do support our assertion that natural selection is driving differentiation of CNVs in *D. melanogaster* along the East Coast, thereby demonstrating the utility of our method for identifying CNVs under spatially varying selection.

4. Discussion

The method presented here accurately detects differentiated copy-number variants from pooled DNA sequence data, and we show that many of the CNVs identified likely reside in regions experiencing spatially varying selection. Because of the high level of gene flow between the two *Drosophila* samples examined here, differentiation at neutral variants is short-lived, and regions with polymorphisms differing in allele frequency between the two samples are quite small, often on the order of 5 kb or less [48]. Thus, while it is difficult to be certain that any given CNV identified by the approach described here is indeed responsible for allele frequency differentiation, it is likely that many of the CNVs identified by our method are indeed beneficial mutations. We believe this approach has the potential to identify CNVs under spatially varying selection in other species and environmental gradients, and significantly improve our understanding of the contribution of copy-number variation to adaptive evolution.

References

1. R. Sachidanandam, et al., *Nature* **409**, 928-933 (2001).
2. J. Sebat, et al., *Science* **305**, 525-528 (2004).
3. D. F. Conrad, et al., *Nature* **464**, 704-712 (2010).
4. S. Ossowski, et al., *Genome Res* **18**, 2024-2033 (2008).
5. R. Redon, et al., *Nature* **444**, 444-454 (2006).
6. E. Tuzun, et al., *Nat Genet* **37**, 727-732 (2005).
7. F. Hormozdiari, et al., *Genome Res* **21**, 2203-2212 (2011).
8. P. Medvedev, et al., *Genome Res* **20**, 1613-1622 (2010).
9. R. E. Handsaker, et al., *Nat Genet* **43**, 269-U126 (2011).
10. G. H. Perry, et al., *Genome Res* **18**, 1698-1710 (2008).
11. T. A. Graubert, et al., *PLoS Genet* **3**, e3 (2007).
12. J. J. Emerson, et al., *Science* **320**, 1629-1631 (2008).
13. C. H. Langley, et al., *Genetics*, doi: 10.1534/genetics.1112.142018 (2012).
14. L. Carreto, et al., *BMC Genomics* **9**, 524 (2008).
15. S. Girirajan, et al., *PLoS Genet* **7**, e1002334 (2011).
16. D. Moreno-De-Luca, et al., *Am J Hum Genet* **87**, 618-630 (2010).
17. J. R. Lupski, *Nat Genet* **39**, S43-S47 (2007).
18. S. A. McCarroll, et al., *Nat Genet* **40**, 1107-1112 (2008).
19. A. B. Singleton, et al., *Science* **302**, 841 (2003).
20. D. Altshuler, et al., *Nature* **467**, 1061-1073 (2010).
21. D. M. Altshuler, et al., *Nature* **467**, 52-58 (2010).

22. P. C. Sabeti, et al., *Nature* **419**, 832-837 (2002).
23. B. F. Voight, et al., *PLoS Biol* **4**, e72 (2006).
24. L. B. Barreiro, et al., *Nat Genet* **40**, 340-345 (2008).
25. R. C. Iskow, O. Gokcumen, C. Lee, *Trends Genet* **28**, 245-257 (2012).
26. G. H. Perry, et al., *Nat Genet* **39**, 1256-1260 (2007).
27. E. Gonzalez, et al., *Science* **307**, 1434-1440 (2005).
28. B. Kolaczkowski, et al., *Genetics* **187**, 245-260 (2011).
29. C. Cheng, et al., *Genetics* **190**, 1417-1432 (2012).
30. A. Futschik, C. Schlotterer, *Genetics* **186**, 207-218 (2010).
31. Y. Zhu, et al., *PLoS ONE* **7**, e41901 (2012).
32. M. Slatkin, *Genetics* **99**, 323-335 (1981).
33. H. Li, R. Durbin, *Bioinformatics* **25**, 1754-1760 (2009).
34. M. D. Adams, et al., *Science* **287**, 2185-2195 (2000).
35. G. M. Cooper, et al., *Nat Genet* **40**, 1199-1203 (2008).
36. D. R. Schrider, M. W. Hahn, *Mol Biol Evol* **27**, 103-111 (2010).
37. S. Wright, *Genetics* **28**, 114-138 (1943).
38. C. Alkan, et al., *Nat Genet* **41**, 1061-1067 (2009).
39. A. Abyzov, et al., *Genome Res* **21**, 974-984 (2011).
40. S. Lee, et al., *Nat Methods* **6**, 473-474 (2009).
41. S. Sindi, et al., *Bioinformatics* **25**, I222-I230 (2009).
42. R. Kofler, A. J. Betancourt, C. Schlotterer, *PLoS Genet* **8**, e1002487 (2012).
43. H. Ranson, et al., *Science* **298**, 179-181 (2002).
44. P. J. Daborn, et al., *Insect Biochem Mol Biol* **37**, 512-519 (2007).
45. M. R. Bogwitz, et al., *Proc Natl Acad Sci U S A* **102**, 12807-12812 (2005).
46. J. Kang, J. Kim, K.-W. Choi, *PLoS ONE* **6**, e29800 (2011).
47. P. Menozzi, et al., *BMC Evol Biol* **4**, 4 (2004).
48. T. L. Turner, et al., *Genetics* **179**, 455-473 (2008).
49. P. Pavlidis, et al., *Mol Biol Evol*, doi: 10.1093/molbev/mss1136 (2012).
50. J. D. Storey, *Journal of the Royal Statistical Society Series B-Statistical Methodology* **64**, 479-498 (2002).

METASEQ: PRIVACY PRESERVING META-ANALYSIS OF SEQUENCING-BASED ASSOCIATION STUDIES*

ANGAD PAL SINGH, SAMREEN ZAFER AND ITSIK PE'ER†

Department of Computer Science, Columbia University
New York, New York 10027-7003, USA
Email: {aps2157, sz2317, ip2169†} @columbia.edu

Human genetics recently transitioned from GWAS to studies based on NGS data. For GWAS, small effects dictated large sample sizes, typically made possible through meta-analysis by exchanging summary statistics across consortia. NGS studies groupwise-test for association of multiple potentially-causal alleles along each gene. They are subject to similar power constraints and therefore likely to resort to meta-analysis as well. The problem arises when considering privacy of the genetic information during the data-exchange process. Many scoring schemes for NGS association rely on the frequency of each variant thus requiring the exchange of identity of the sequenced variant. As such variants are often rare, potentially revealing the identity of their carriers and jeopardizing privacy. We have thus developed MetaSeq, a protocol for meta-analysis of genome-wide sequencing data by multiple collaborating parties, scoring association for rare variants pooled per gene across all parties. We tackle the challenge of tallying frequency counts of rare, sequenced alleles, for meta-analysis of sequencing data without disclosing the allele identity and counts, thereby protecting sample identity. This apparent paradoxical exchange of information is achieved through cryptographic means. The key idea is that parties encrypt identity of genes and variants. When they transfer information about frequency counts in cases and controls, the exchanged data does not convey the identity of a mutation and therefore does not expose carrier identity. The exchange relies on a 3rd party, trusted to follow the protocol although not trusted to learn about the raw data. We show applicability of this method to publicly available exome-sequencing data from multiple studies, simulating phenotypic information for powerful meta-analysis. The MetaSeq software is publicly available as open source.

1. Introduction

Human genetics has recently undergone a transition from genomewide association studies (GWAS) based on genotyping common polymorphisms[1-4] to studies based on next generation sequencing (NGS) data[5-7], that ascertains common and rare variants across individuals[8]. For GWAS, low effect sizes of most of the causal common alleles on common diseases and quantitative traits dictated large sample sizes to achieve statistical power[9]. In many studies, such sizes were made possible by consortia of multiple collaborating groups, each contributing hundreds or thousands of samples, together amassing tens or hundreds of thousands of genotyped samples to detect minute effects on various phenotypes[10]. Computational methods for meta-analysis of such collated GWAS datasets have been instrumental in facilitating their joint analysis[11].

* This work is supported by NSF grants CCF-0829882, CCF-0845677 and NIH grant U54-CA121852

† Author to which all correspondence should be addressed.

NGS studies met initial success using only a handful of samples for sequencing exomes[12,13] or whole genomes[14,15] to detect novel, fully-penetrant alleles that disrupt genes and cause disease. Yet, detecting disease genes with rare alleles of partial penetrance, that explain only a small fraction of the cases, is more challenging. First, the limited power to detect such alleles on their own motivates testing for association of multiple alleles along the gene [16]. Indeed, multiple methods for groupwise testing of alleles have been developed to optimize power of detecting such multiply disrupted genes[17-22]. Second, the tautological problem with rare variants is their low frequency. Large numbers of samples are still required in order to observe such alleles and detect their significant association. Fortunately, the cost of NGS keeps dropping, and the throughput keeps increasing. Sequencing exomes now require reagent-cost and labor resources comparable to early GWAS, with genomes likely to soon follow. This paper is motivated by the assumption that these power constraints along with throughput opportunities will lead to large-scale disease sequencing studies[23] that would be more rapidly, and therefore more competitively executed by groups operating in parallel, but jointly meta-analyzing their data.

Privacy had been a thorny issue in genetics research[24-26]. The irreversible labeling of individuals if their genetic information is known requires broad consent by study participants in order for researchers to have the ethical right and legal permit to expose their genotype data or even to share it with peers and collaborators[27,28]. This, along with some investigators' sense of ownership of their data and cohorts typically makes data-access in human genetics (unlike other fields[29,30]) restricted, at least initially, often to the investigator. In GWAS, large consortia had preserved such access restriction, as meta-analysis required only exchange of summary statistics across collaborating groups and institutional barriers, rather than sharing explicit genotype data[31]. Such summary statistics typically include essentially allele frequencies (and their confidence levels) per marker. Although formally individuals and their relatives can be identified as members of a cohort just based on these summary statistics[32], this identification requires expert computation, and may be underpowered, depending on study parameters such as number of SNPs, sample size and allele frequencies[33].

Meta-analysis of sequencing data poses unique challenges in terms of subject privacy. Specifically, such data includes hundreds of thousands of rare alleles per genome[34], among them de novo mutations[35], one or two of which can uniquely identify an individual among the entire world population. Even exome sequences typically include thousands of alleles that are currently novel[13]. Even assuming future expansion of variant databases, a typical human exome will have thousands of very rare (frequency $< 10^{-4}$) alleles, typically singletons within a cohort of size in the low thousands. Such alleles, alone or in concert, readily provide unambiguous identification of carrier of the sample. The classical summary statistic for meta-analysis, which is the list of allele frequencies in a sample, therefore provides clear indication of membership for each and every sample in the cohort if applied genomewide with the exception of monozygotic twins, simply by virtue of including the singleton alleles carried by this sample. A similar rationale would decide or rule out membership in an exome-sequenced cohort based on presence of rare mutations. Yet, allele frequencies in cases and controls across the entire set of analyzed samples are a key ingredient in multiple methods for association to rare alleles[18,19,21]. Exchange of allele frequencies between consortium members in order to tally alleles across datasets is instrumental for meta-analysis of sequencing data, posing an apparent conflict with ethical requirements to protect against identification of samples.

This paper tackles the challenge of facilitating the tally of frequency counts of rare, sequenced alleles between consortium members, thus enabling meta-analysis of sequencing data while not disclosing the allele identity and counts, therefore providing considerable protection of sample identity. This apparent paradoxical exchange of information is achieved through cryptographic means. The key idea is that parties hide the identity of the variants. When they transfer information about frequency counts in cases and controls, it does not convey the identity of a mutation, therefore not exposing the identity of the carriers. The parties do use an identical encryption key, thus identical variants will be encrypted identically. One could therefore sum up the counts for identical variants, without knowing the identity of the alleles whose counts are being tallied.

2. Methods

2.1. *Notation*

We hereby describe MetaSeq, a privacy preserving protocol for meta-analysis of sequencing data coming from C collaborators such that:

- Each collaborator c has data on a set $S[c]$ of samples.
- Such data includes a set $V_m[c]$ of positions along each gene g_m among the $M \sim 20{,}000$ genes $g_1, g_2 \ldots g_M$. $V_m[c]$ specifies all positions where variant (no-reference) calls had been made for at least one sequenced individual $i \in S[c]$.
- The data further includes for each individual $i \in S[c]$, and each variant position $v \in V_m[c]$ the actual genotype of i at v: heterozygote or non-reference homozygote, denoted by $h_m[c](v,i)$, represented in a standard vcf format[36]. We define $H_m[c]$ to be full matrix of genotype values, across all rows $v \in V_m[c]$, and columns $i \in S[c]$. Effectively, $H_m[c]$ is a matrix of values 0,1, or 2 for each position and individual.
- For each individual $i \in S[c]$, the data also includes the affection status or the phenotype value of i, denoted by $p(i) \in \{1,0\}$ for cases and controls, respectively. We denote $P[c]$ as the list of phenotype values $p(i)$ for each $i \in S[c]$.

We assume $V_m[c]$ is listed as genomic coordinates: chromosome and position along the chromosome. For each such position x, we define the coordinate, $\varphi_m(x)$, which is its offset from the start of the chromosome. We naturally extend $\varphi_m(.)$ to operate on sets of positions. In practice we assume $\varphi_m(x)$ is a 32-bit integer.

We define the set of all variable positions along the chromosome for gene, $g_{m,}$ and the total set of individuals respectively, as follows:

$$V_m = \bigcup_c V_m[c] \tag{1}$$

$$S = \bigcup_c S[c] \tag{2}$$

We further define the full listing P of phenotype values for all individuals across all cohorts and the full set G of genes, $g_1 .. g_M \mid M \sim 20000$. H_m is defined as the genotype matrix across all cohorts, with columns for all $i \in S$, and rows for all $v \in V_m$. $H_m[c]$ is the minor of H_m induced on $V_m[c] \times S[c]$. The data for gene m is $D_m = \{V_m, H_m\}$, and the entire genetic dataset is given as:

$$D = \bigcup_m D_m \tag{3}$$

2.1.1. *Association score*

Let $F(D_m = (V_m, H_m), P)$ be the scoring function used for testing association of g_m. We assume F has certain properties that are shared by standard methods for testing association[18]. Specifically, F remains fixed when swapping rows (variants) of H_m along with V_m , if we assume all variants considered by the test are similarly likely to be causal (this assumption can be relaxed). Also, the set of scores for all genes by definition remains fixed when swapping genes g_m.

$$F(D,P) = \{F(D_m,P)\}_{m=1}^{M} \tag{4}$$

The goal of the protocol is to encrypt the data using a secret key k, such that gene labels and variant labels are swapped (or permuted). Specifically, we define key-dependent permutations g_k and ρ_k on gene labels and potential coordinates (32-bit integers), respectively. The permuted data for each gene is denoted by the following equations:

$$\rho_k(D_m) = (\rho_k(V_m), \rho_k(H_m)) \tag{5}$$

$$\rho_k(V_m) = \rho_k(\varphi_m(v)) \,|v \in V_m \tag{6}$$

where, $\rho_k(V_m)$ is the set of permuted coordinates and $\rho_k(H_m)$ is the matrix of genotype calls with permuted rows, i.e., with values $h_m[\rho_k(\varphi_m(v)), i]$ for all $v \in V_m, i \in S$. We observe that the score is unchanged by this transformation: $F(\rho_k(D_m), P) = F(D_m, P)$. Yet, if one were to observe only a minor of $\rho_k(D_m)$, corresponding to a subset of individuals and the corresponding subsets of variants that they carry, one does not obtain any information on the individuals not in this subset, nor on the variants not carried by these individuals. Specifically, for each cohort c, the relevant subset of the data, $D_m[c] = (V_m[c], H_m[c])$, when encrypted into $\rho_k(D_m)[c]$, does not provide information regarding any other cohort $c' \neq c$, nor on any variants not in $V_m[c]$. In this sense, the encryption is privacy preserving. Finally, if gene labels are permuted, then receiver of the permuted data $D_{g(k)} = \{D_{g_m(k)} \mid m \in 1..M\}$ cannot learn anything about the identity of any gene.

We have developed a 5-step protocol for meta-analysis of genomewide sequencing data, computing association scores for pooled rare variants. The protocol is presented here in simplified form, with the following leniencies:

1. We discuss only two-way meta-analysis, where two investigators (collaborators), Alice and Bob (or c_1 and c_2), each have their own sequenced association cohorts.
2. We consider case-control association testing.
3. We present the calculation of a simple variable allele-frequency threshold score[21].
4. Alice and Bob rely on the assistance of a semi-trusted third party, Trent, to help compute the score.

The protocol preserves privacy of the subjects in the following respects:

1. The only information Alice and Bob learn about each other's cohort is the scores of top-associated genes.
2. Trent does not have direct or practical information that could expose the identity of the subject in Alice and Bob's cohorts. Specifically, Trent does not learn which genes harbor which mutations in each cohort, and given an exome of an individual, cannot determine whether that individual is a member of any of the cohorts. Even upon publication of the research results by Alice and Bob, the information that Trent learns, is limited.

2.1.2. *Protocol*

The protocol proceeds as follows:

1. Key Exchange:
   Alice and Bob choose a shared secret key k, that can serve as an encryption key
2. Annotation and Encryption:
   Alice and Bob each encrypt their data as follows:
   a. Variants are annotated for the genes they belong to and variant classification, e.g. *known* or *nonsense*, needed for scoring. Such classification is kept unencrypted.
   b. Alice and Bob generate a secret permutation $g(k)$ over the set of genes $g_1...g_M$, creating permuted gene identifiers, $g_1(k)...g_M(k)$.
   c. They further secretly permute the set of variants V_m, creating $V_m(k)$.
3. Data Transfer:
   Alice and Bob send Trent their encrypted gene names $g_1(k)...g_M(k)$ and variant positions, $V_m(k)$ along with the (unencrypted) (frequency) counts $f_{V_m(k)[c_i]}$.
4. Merging and association testing:
   Trent computes, for each (permuted) gene $g_m(k)$ a total count for each (encrypted) variant,

$V_m(k)$ by summing the two counts $f_{V_m(k)[c1]}$ and $f_{V_m(k)[c2]}$ if both Alice and Bob report the variant in $g_m(k)$ or collapsing the association score for the variants otherwise.

5. Decrypt results:
   Trent sends Alice and Bob the (top) association scores assigned to specific (encrypted) gene names, that they are able to decrypt.

Note that many rare-variant association tests focus on particular type of variants, e.g. non-synonymous, or loss-of-function variants. Such information is lost upon encryption, and Trent will thus be unable to restrict analysis to a particular class of variants. A convenient workaround is to communicate a set of per-variant weights by both Alice and Bob. Weights depend on classification of variant type that is agreed upon in advance, i.e. Alice and Bob decide on a weight function $W{:}T \rightarrow [0,1]$ on the domain of all variant types $T = \{$*missense, synonymous coding* …..$\}$. Each variant v is assigned type $t(v) \in T$ and therefore a real-valued weight $W(t(v)) \in [0,1]$, is communicated to Trent in clear text. We make note of the fact that since both gene names and variant positions are encrypted, for a sufficiently large class of variant types it becomes difficult for Trent to make any concrete inferences on variant identities using this information.

2.1.3. *Implementation: MetaSeq*

We implemented this protocol as MetaSeq, an open source PERL package. A step-by-step illustration of the protocol as is in the MetaSeq code is given in Figure 1. We assume that the collaborators have their data stored on a server that is remotely accessible using the server name. We also require tools for annotation and encryption of the data on the server. MetaSeq works on variant call files (*.vcf format) that include genotypes and phenotypes for each collaborating party, and is available as open source at https://github.com/angadps/Rare-Variant-Association.

We provide implementation details regarding specific steps of MetaSeq:

Step 1: Registration & key exchange
MetaSeq guides the collaborating parties through the key exchange procedure using the PERL encryption modules Crypt::DES[37], Crypt::CFB[38] and Crypt::CBC[39], and allows an arbitrary number of collaborators, instead of just the pair of Alice and Bob. In detail, the collaborators register with Trent using their server names. Communication between the servers is via the use of sockets. A specific port is designated on the servers for all data exchange and communication between the servers. Trent signals the key generation process after registration. All collaborators contribute a seed towards the generation of the key, of which Trent has no information about or contributes in any way towards the generation of either. We use the MD5 algorithm to generate a 32-bit key.

Step 2: Annotation & encryption
Each party then encrypts the data, which are first annotated by the vcfCodingSnps tool[40] on a per gene basis. The purpose of annotation is two-fold. Firstly, it helps us prepare the genotype and phenotype files separately for every gene as required by the association test. Secondly, it helps us

in restricting the analysis to certain class of variants, or in assigning different weights to different classes. For that purpose, additional input to MetaSeq is a file of weights that needs to be agreed upon in advance. Variant data is encrypted per the protocol, and communicated as numeric 32-bit dumps – sufficient to uniquely index positions along any chromosome. At the same time we would like to point out that we have tested MetaSeq to work with gene level annotations only, although the idea could be extended to any general definitions of region for annotations as long as it is consistent across studies.

Step 4: Merging and association testing

MetaSeq is implemented with the Variance Threshold (VT) test[21] of association, but can in principle include other tests as well. The encrypted files received by Trent from Alice and Bob are first merged by their (encrypted) gene names. This prepares the data from all collaborators for the pooled association test.

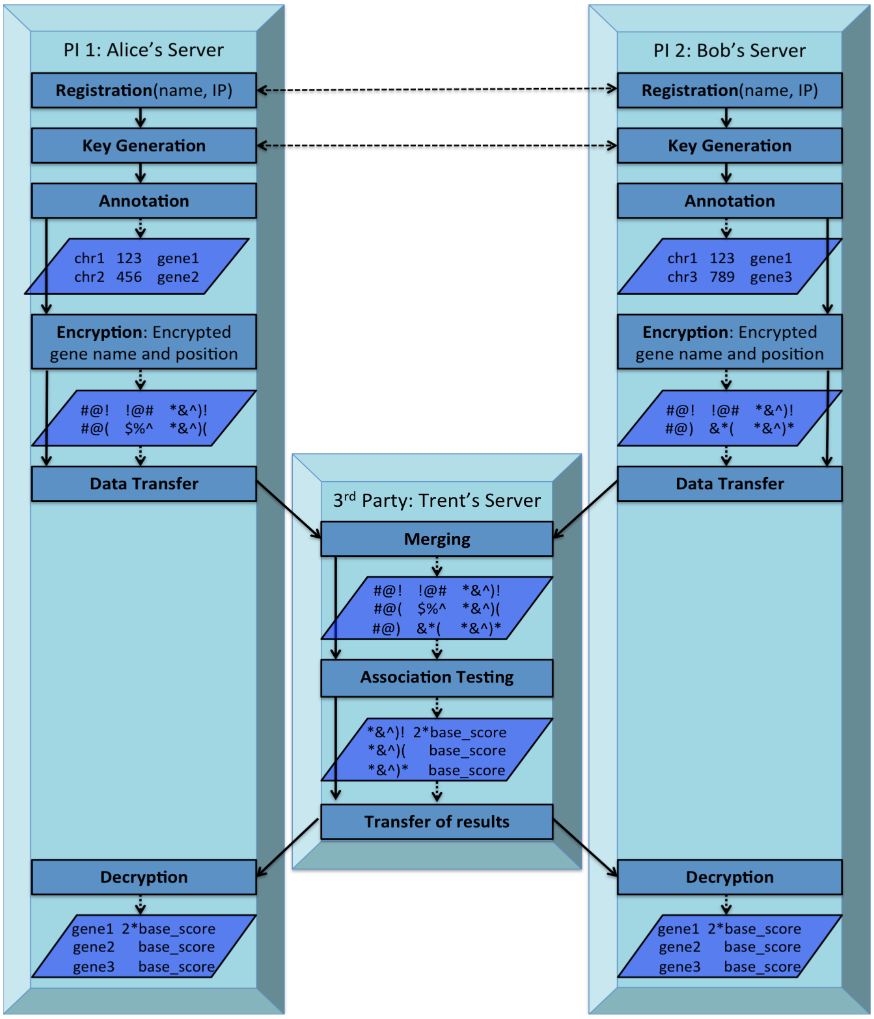


Figure 1 Flow diagram of MetaSeq: Two Investigators, Alice and Bob compute per-gene scores on their pooled data without revealing the data to one another nor to a third party, Trent, who computes association scores "blindfolded". The figure describes a simple scenario using three genes, one of which, including a single variant in it, is common to Alice and Bob. The gene name and variant position for this is encrypted to the same text, thus being merged together by the 3rd party before association testing. This gene scores higher compared to other genes, as shown in the results decrypted by individual collaborators on their servers. Note that the generation of the key to be used for encryption is coordinated between Alice and Bob, excluding Trent in the process. Also note that while the figure does not point out phenotype information explicitly, the association testing step of the protocol receives the frequency data segregated for case and control cohorts, respectively.

2.2. *Simulation Testing*

We used simulation to evaluate the power of meta-analysis assuming different numbers of causal variants in a single gene. Power here is defined as the fraction of successful association tests. Specifically, for each such number, we simulated 100 datasets of 50 cases and 50 controls collected by each of C=10 collaborators. We tested association by each single-collaborator vs. pooled across collaborators in a privacy-preserving manner. We tallied the fraction of successful association tests, but note that reporting a success requires more care in this study than usual. In detail, a conservative definition of success is when the true gene is the unique top-scoring gene (for either single-collaborator or pooled testing modes). A more lenient definition allows other top-scoring genes to tie with the true gene (again, for both modes). Finally, without privacy-preserving data analysis, one can consider independent PIs running the association test, and then decide about the associated gene based on the individual results of all of them, by taking a majority vote. We report power based on each of these 5 modes of analysis. We repeated this for $1,2,2^2, \ldots 2^{10}$ causal variants for the causal gene, in addition to 1000 neutral variants for each gene. We simulated the case and control sequencing data using an implementation of the Wright-Fisher model[41], that allows setting particular numbers of causal and neutral variants. The Wright-Fisher Model gives the probability density function *f*(*p*), of the probability of encountering a mutation, *p* as follows:

$$f(p) = c \ * \ p^{bs-1} * \ (1-p)^{bn-1} * \ e^{s(1-p)} \qquad (7)$$

Here, *f*(*p*) is the probability function of the mutation-probability *p*, b_s is the scaled mutation rate of disease mutations, b_n is the scaled back-mutation rate, *s* is the scaled selection rate and *c* is the constant that normalizes the integral of *f*(*p*) to 1.

3. Results

3.1. *Power of pooled-collaborators vs. single-PI testing*

We report results from all the variants of the power tests stated above. Plots for the same are shown in Figure 2. Throughout the range of parameters, pooled tests are better powered compared to single-PI tests. This advantage is most pronounced when there are only few causal variants along the truly causal gene. At the extreme, 1-8 causal variants in a gene, we observe decently powered pooled test (up to 55% power for the conservative test) compared to a severely (<5%) underpowered single-PI test, an improvement of up to 50 percentage points or 10 – 30 times with the pooled tests. Naturally, lenient reporting of success enjoys higher power, but would potentially require following up multiple promising genes, rather than only one.

We note that the number of causal, case-only variants is a natural parameter here – the rare-allele analog of the size of effect to be detected. Power is further influenced by nuisance parameters, such as the span of a gene in basepairs (hence, the number of neutral variants along it, here normalized to be 1,000), and the genetic length of a gene in centimorgans (hence, the effective number of independent variants along it). This explains some of the genes being hard to

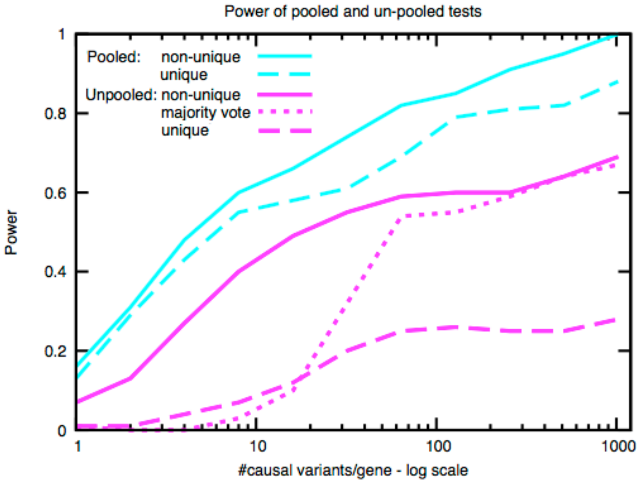


Figure 2 Log-scaled power plot for pooled and single-PI tests. Results are plotted for all three definitions of success (unique and non-unique causal gene for pooled and single-PI tests, majority vote for single-PI tests only). In the unique and non-unique gene plots for the single-PI tests, the final success rate is calculated by averaging the number of successes across all PIs per dataset. In the majority vote plot, a majority vote of the number of successes is taken per single-PI per dataset. The different nature of success here explains the region in the figure where the plot for single-PI unique gene tests is higher than the majority vote plot.

find as associated, even with many rare case-only variants simulated. Potential false positives or false negatives in the context of meta-analysis alone are expected to be minimal (otherwise, the same concerns may apply as in the case of single cohort tests). Since variant frequencies are collapsed across all cohorts and for all variants in a gene, such loss of data, which is the primary input for the protocol is not expected. Also, encryption is performed in a loss-less manner i.e., no genes or variant ids are expected to be lost in the due course of execution of the protocol.

3.2. *MetaSeq requirements of computing resources*

We state the time and space requirements for MetaSeq in Table 1. The tests were run on a Sun Grid Engine controlled cluster with sufficient number of compute cores and maximum 8GB of RAM given to a single test at any time. We state the time and space that was required for a single run of

Table 1 Space and time requirements of MetaSeq. The benchmark runs included 10 collaborators, with each one contributing 100 samples to the pooled analysis including 1000 neutral variants per gene. Runs include all ~20,000 genes along the genome. Steps performed by the collaborating parties ("Alice & Bob", though in this benchmark also 8 other collaborators) are evaluated for resources required per party. Also time taken for the association test mentioned is with a parallelism of 20. A total of 20 CPU hours were effectively needed for the association testing, although total memory required is less than 1MB. Note that decryption takes negligible time as opposed to encryption since the parties only need to decrypt the list of top-scoring gene names.

Step	Performed by	Elapsed time [min]	CPU time [min]	Memory [MB]
1.1 Register	Alice & Bob	Nil	Nil	Nil
1.2 Generate key		Nil	Nil	Nil
2.1 Annotate		18	180	15
2.2 Encrypt		27	270	12
3 Transfer data		1	10	12
4.1 Merge	Trent	80	80	105
4.2 Test association		60	1200	1
5.1 Transfer results		Nil	Nil	Nil
5.2 Decrypt	Alice & Bob	Nil	Nil	Nil

MetaSeq with 1000 neutral variants per gene, broken down by small steps of the protocol. Some of the steps, i.e., registration, key generation, transfer of results, and decryption are insignificant both in terms of time and space. Yet, these steps are reported here for completion. In total, MetaSeq can be completed in 3.5 hours of elapsed time using less than 30 hours of CPU resources, using at its peak 150MB of space in total. Network footprint is even smaller, as transmitted files are archived and zipped. The most intensive parts in terms of computing resources are the annotation and encryption stages that need to I/O information in 200,000 files (one per gene per collaborator). The most CPU is used during association testing, for permuting the data 100,000 times to assess significance. We parallelize this stage over 20 cores.

4. Discussion

We developed MetaSeq, a protocol that relies on a trusted third-party to compute the association scores over the intersection of the variant set. We implemented the protocol in PERL and have made it available as an open source package. Our protocol is designed to be robust in securing private genetic information, while at the same time making only minimal assumptions about compliance of the parties to the protocol.

In securing private genetic information, we try to preserve privacy against participating collaborators knowing individual-level identifying information, such as private mutations. This is achieved by computing an association score, not by one of the parties, but rather by a designated third-party, who also needs to stay in the dark and not learn the identity of the study participants and their private mutations. The third-party, after collating data and performing the desired computations, is assumed to follow protocol, and not to share variant information with any of the collaborators. The third-party is considered to be “trusted” in this regard. At the same time we need to secure information from the third-party as well. We achieve this by this party only working with encrypted data, never having access to the secret key that was used to encrypt all the genetic information. Hence, while the third-party has access to all the data, it is still meaningless to that party, since the data is in encrypted form and the encryption key is not available to it.

We make the assumption that no collaborator conspires with the third-party to share the key, as that would violate the desired privacy requirements. Another potential breach that can arise is when more than one collaborator plan to collate their datasets so as to draw inferences regarding the data from the remaining collaborators. However, such estimations can only be effectively made only if all but one of the collaborators get together and conspire against the remaining one. Even then, the coalition would, at best, learn limited information about the cohort of the conspired-upon collaborator, e.g., presence of variants that they already have in their cohort. The coalition will not learn the identity of private variants.

Another way that collaborators can violate protocol to learn the alleles is to send monomorphic data to the third-party for their own dataset. In this way they are sure that any identified carrier alleles are coming only from the datasets of other collaborators. This is possible only if all but one of the collaborators is sending monomorphic data, and we assume the parties follow protocol. At the same time it is assumed that there may be (approximately) a minimum of 5 collaborators in any run of the protocol. Under this assumption it is difficult for a single collaborator to learn the datasets of any other single collaborator by employing such mechanisms.

Finally, a collaborator may try to estimate datasets by computing a prior distribution of the results obtained from the final computation of scores, which is OK, and then use their own dataset to obtain a better posterior distribution. However, they only have a chance to learn about variants that are shared, rather than private to a cohort, and only within the top-scoring genes. A theoretical analysis of the privacy guarantees of the protocol may resemble the one by Sankararaman[42] to some extent although we are now working in the MAF < 0.5 range. A complete analysis however remains out of scope for this paper and will be considered for future work.

Privacy preserving protocols of this sort have been investigated in the cryptography literature as secure multiparty computation[43]. Over the last decade, protocols have been proposed for joint computation of the intersection of two or more subsets[44] that can be employed to compute the intersection of the variant set. More generally, theoretical results guarantee the ability to simulate any privacy-preserving protocol that uses a third trusted party without the need of such a party[45]. Similar to meta-analysis techniques in GWAS, the application of similar techniques for NGS studies is expected to reveal the role of many rare variants in Mendelian diseases.

Acknowledgements

We are grateful to Rajan Banerjee for his work in building some of the supporting cryptographic modules in MetaSeq.
We are also thankful to the NSF and the NIH for their funding provided via the NSF grants CCF-0829882, CCF-0845677 and the NIH grant U54-CA121852.

References

1. Burton, P.R., Clayton, D.G. *et al. Nature* **447**, 661-78 (2007).
2. Easton, D.F. *et al. Nature* **447**, 1087-93 (2007).
3. Hirschhorn, J.N. & Daly, M.J. *Nat Rev Genet* **6**, 95-108 (2005).
4. Sladek, R. *et al. Nature* **445**, 881-5 (2007).
5. Harismendy, O. *et al. Genome Biol* **10**, R32 (2009).
6. Schuster, S.C. *Nat Methods* **5**, 16-8 (2008).
7. Shendure, J. *Genome Biol* **12**, 408 (2011).
8. Shen, Y. *et al. Genome Res* **20**, 273-80 (2010).
9. Spencer, C.C., Su, Z., Donnelly, P. & Marchini, J. *PLoS Genet* **5**, e1000477 (2009).
10. Speliotes, E.K. *et al. Nat Genet* **42**, 937-48 (2010).
11. Evangelou, E., Maraganore, D.M. & Ioannidis, J.P. *PLoS One* **2**, e196 (2007).
12. Bilguvar, K. *et al. Nature* **467**, 207-10 (2010).
13. Ng, S.B. *et al. Nature* **461**, 272-6 (2009).
14. Lupski, J.R. *et al. N Engl J Med* **362**, 1181-91 (2010).
15. Roach, J.C. *et al. Science* **328**, 636-9 (2010).
16. Cohen, J.C. *et al. Science* **305**, 869-72 (2004).
17. Ionita-Laza, I. & Ottman, R. *Genetics* **189**, 1061-8 (2011).

18. Li, B. & Leal, S.M. *Am J Hum Genet* **83**, 311-21 (2008).
19. Madsen, B.E. & Browning, S.R. *PLoS Genet* **5**, e1000384 (2009).
20. Neale, B.M. *et al. PLoS Genet* **7**, e1001322 (2011).
21. Price, A.L. *et al. Am J Hum Genet* **86**, 832-8 (2010).
22. Zeggini, E. *et al. Nat Genet* **40**, 638-45 (2008).
23. KA, W. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program Available at: http://www.genome.gov/sequencingcosts/. Accessed Apr 17, 2012. (2012).
24. Fuller, B.P. *et al. Science* **285**, 1359-61 (1999).
25. Lin, Z., Owen, A.B. & Altman, R.B. *Science* **305**, 183 (2004).
26. Regalado, A. *Wall St J (East Ed)* **R10**(2002).
27. Caulfield, T. *et al. PLoS Biol* **6**, e73 (2008).
28. Kaye, J., Heeney, C., Hawkins, N., de Vries, J. & Boddington, P. *Nat Rev Genet* **10**, 331-5 (2009).
29. Berman, H.M. *et al. Nucleic Acids Res* **28**, 235-42 (2000).
30. Edgar, R., Domrachev, M. & Lash, A.E. *Nucleic Acids Res* **30**, 207-10 (2002).
31. de Bakker, P.I. *et al. Hum Mol Genet* **17**, R122-8 (2008).
32. Homer, N. *et al. PLoS Genet* **4**, e1000167 (2008).
33. Jacobs, K.B. *et al. Nat Genet* **41**, 1253-7 (2009).
34. Durbin RM, A.G., Altshuler DL *et al. Nature* **467**, 1061-73 (2010).
35. Au, K.S. *et al. Genet Med* **9**, 88-100 (2007).
36. Danecek, P. *et al. Bioinformatics* **27**, 2156-8 (2011).
37. http://search.cpan.org/~dparis/Crypt-DES-2.05/
38. http://search.cpan.org/~kjh/Crypt-CFB-0.02/
39. http://search.cpan.org/~lds/Crypt-CBC-2.30/
40. http://www.sph.umich.edu/csg/liyanmin/vcfCodingSnps/index.shtml
41. Cheung, Y.H., Wang, G., Leal, S.M. & Wang, S. *Genet Epidemiol* (2012).
42. Sankararaman, S., Obozinski, G., Jordan, M.I. & Halperin, E. *Nat Genet* **41**, 965-7 (2009).
43. Yao, A. Protocols for secure computation. in *23rd FOCS* 160-164 (1982).
44. Song, L.K.a.D. *School of Computer Science, Carnegie Mellon University* **CMU-CS-04-182**(2004).
45. S.S.M. Chow, J.L., and L. Subramanian. *in Proc. NDSS* (2009).

TEXT AND DATA MINING FOR BIOMEDICAL DISCOVERY

GRACIELA GONZALEZ

Department of Biomedical Informatics,
Arizona State University
Scottsdale, AZ 85259, USA
Email: ggonzalez@asu.edu

KEVIN BRETONNEL COHEN

Computational Bioscience Program
U. Colorado School of Medicine
Aurora, CO
Email: kevin.cohen@gmail.com

MARICEL G. KANN

Department of Biological Science
University of Maryland, Baltimore County
Baltimore, MD, 21250, USA.
Email: wspc@wspc.com

CASEY S. GREENE

Department of Genetics
Geisel School of Medicine at Dartmouth
Hanover, NH 03755, USA
Email: Casey.S.Greene@dartmouth.edu

ROBERT LEAMAN

National Center for Biotechnology
Information
Bethesda, MD 20894, USA
Email: robert.leaman@nih.gov

UDO HAHN

Friedricdh Schiller University Jena
Language and Information Engineering Lab
Jena, Germany
Email: Udo.Hahn@uni-jena.de

NIGAM SHAH

Stanford Center for
Biomedical Informatics Research
Stanford, CA 94305, USA
Email: nigam@stanford.edu

JIEPING YE
Computer Science and Engineering,
Arizona State University, Tempe, AZ 85287
Email: jieping.ye@asu.edu

The biggest challenge for text and data mining is to truly impact the biomedical discovery process, enabling scientists to generate novel hypothesis to address the most crucial questions. Among a number of worthy submissions, we have selected six papers that exemplify advances in text and data mining methods that have a demonstrated impact on a wide range of applications. Work presented in this session includes data mining techniques applied to the discovery of 3-way genetic interactions and to the analysis of genetic data in the context of electronic medical records (EMRs), as well as an integrative approach that combines data from genetic (SNP) and transcriptomic (microarray) sources for clinical prediction. Text mining advances include a classification method to determine whether a published article contains pharmacological experiments relevant to drug-drug interactions,

a fine-grained text mining approach for detecting the catalytic sites in proteins in the biomedical literature, and a method for automatically extending a taxonomy of health-related terms to integrate consumer-friendly synonyms for medical terminologies.

1. Introduction

The explosion of genomic data available to researchers from countless gene expression and sequencing experiments, coupled with the abundance of knowledge in the published literature and curated databases, fuels the need for novel and transformative methods for knowledge extraction, visualization, and analysis that take advantage of all of these sources to elicit new and meaningful hypotheses. The biggest challenge for text and data mining is to truly impact the biomedical discovery process, enabling scientists to generate novel hypothesis to address the most crucial questions. Formulation of a flexible and general approach for integrating heterogeneous data and knowledge sources for discovery is elusive and highly dependent upon the specific underlying scientific question. The true impact of text and data mining is only realized if it goes beyond a focus on the methods for extraction and storage, and into the true impact they can have on enabling understanding of the molecular underpinnings of biological processes.

This session seeks to bring together researchers with a strong text or data mining background who are collaborating with bench scientists for the deployment of integrative approaches in translational bioinformatics. It serves as a unique forum to discuss novel approaches to text and data mining methods that respond to specific scientific questions, enabling predictions that integrate a variety of data sources and can potentially impact scientific discovery.

In order to find the optimal way to integrate relevant information that will help translational and clinical researchers pinpoint novel findings, a thorough understanding of the decision process through which active researchers vet the discoveries proposed by automated systems is required. However, very little is presently known about how scientists actually interpret this information. Cohen and Hersh argue that the "major challenge of biomedical text mining over the next 5-10 years is to make these systems useful to biomedical researchers" [1]. Langley notes the tendency for such tools to be developed for the use of professional data-miners rather than active researchers, and argues for the development of discovery systems with a greater degree of user interactivity [2].

There have been increased efforts to develop such systems and approaches, but there is no single place to present them. This session attracted cross-discipline collaborators with focused applications of discovery and prediction methods. Given the ever increasing deluge of data and knowledge that overwhelms bench scientists around the world; interest in such systems will only increase over time. Some examples of topics of interest to this session include novel approaches that integrate empirical data with knowledge extracted from the literature, curated databases or ontologies to perform discovery-related tasks such as:

- Gene prioritization
- Binding site prediction
- Gene/protein function prediction,
- Prediction of associations (protein-protein, gene-drug, gene-disease, drug-drug)
- Pathway generation or validation

for translational applications such as pharmacogenomics, genome-phenome validation, or detection, diagnostic and prognosis of disease.

2. Challenges

Improving text and data mining methods for any task requires careful consideration and evaluation. The biomedical domain presents specific challenges given the diversity, complexity and volume of the information being mined. This section presents a brief overview of the fundamental challenges faced by researchers in these areas.

2.1. *Text Mining*

Although in general there are challenges such as summarization and question answering, for the type of applications focus of this session, two text mining tasks seem to be specifically relevant: named entity recognition and association extraction.

Named entity recognition (NER) is the problem of finding references to entities (mentions) such as genes, proteins, diseases, drugs, adverse reactions, or organisms in natural language text, and tagging them with their location and type. NER is also referred to as “entity tagging”. This is a basic building block for all other extraction tasks. While there has been significant progress into named entity recognition in the biomedical domain, research has been primarily focused on genes and proteins. Attempts to recognize other entities of interest have concentrated on dictionary matching or statistical approaches. Machine-learning based systems overcome this limitation to a certain extent, given it is possible to retrain such systems to recognize different entity classes. Retraining requires, however, considerable effort in annotation to create a suitable corpus for training the engine, as well as some feature analysis.

Tagging specific entities is of interest as a fundamental step towards the true goal: extracting true associations between terms, such as genes and diseases. Information extraction (IE) from the biomedical literature is usually developed around the extraction of such relationships of interest from text. A typical architecture is composed of special-purpose programs that perform a pipeline of processing modules, including sentence splitters, tokenizers, named entity recognizers, shallow or deep syntactic parsers, and finally extraction based on a collection of patterns. Such systems are usually file-based, so large amounts of processed data can be passed from one module to the next. Relational databases would play a limited role at the end of the extraction pipeline to store the extracted relationships.

This session includes specialized examples of entity recognition and association extraction, showing the trend towards finer granularity in the type of information needed for meaningful applications of text mining for biomedical discovery, requiring a tighter collaboration between the text mining community and domain experts.

2.2. *Data Mining*

In 2006, a paper in the International Journal of Information Technology & Decision Making explored “10 Challenging Problems in Data Mining Research” [3], based on the replies of 14 experts (organizers of the most prestigious Data Mining conferences). It is interesting to note that not only “Data mining for biological and environmental problems” is listed specifically as one of these challenging problems, but that 8 out of the 9 other challenges apply specifically to biomedical data, namely:

- Scaling up for high dimensional data and high speed data streams
- Mining sequence data and time series data
- Mining complex knowledge from complex data
- Data mining in a network setting
- Distributed data mining and mining multi-agent data
- Data Mining process-related problems
- Security, privacy and data integrity
- Dealing with non-static, unbalanced and cost-sensitive data

Of particular interest to this session is the recognized need, when mining complex data, ”for integrating data mining and knowledge inference” and to “to incorporate background knowledge into data mining”. Included in this session are works that address precisely these aspects.

3. Overview of Contributions

Holzinger et al present an integrative approach, ATHENA, used to combine data from genetic (SNP) and transcriptomic (microarray) sources to predict a clinically important feature (HDL-C level). The combined data are capable of predicting HDL-C levels better than either of the individual data sources. Methods capable of connecting measurements of the genome to measurements of transcript and protein abundance for prediction of a clinically relevant phenotype are expected to play a key role in precision medicine.

Hu et al explores the application of the statistical epistasis networks (SEN) approach as filters in the discovery of 3-way genetic interactions. Genetic epistasis is considered an important factor that is related to the etiology of complex diseases. Exhaustive search for high-order interaction is unrealistic due to the large data volume. The authors show that SEN can significantly reduce the number of candidates that need to be considered in a high-order interaction model with improved accuracy.

Kolchinski et al describe a document triage task (binary text classification) for biomedical (Pubmed) articles to determine whether the article contains pharmacological experiments relevant to drug-drug interactions. This joint work between a BioNLP lab (Rocha’s) and a lab doing research in pharmacokinetics (Li’s) exemplifies the type of collaborations likely to result in fruitful advances in biomedical discoveries. The approaches used are variations of approaches known to perform well on similar tasks. The sort of dimensionality reduction and feature transforms performed in the paper are not used as often in BioNLP as they probably should be.

Verspoor et al discuss a fine-grained text mining approach for detecting the catalytic sites in proteins in the biomedical literature. The authors create a silver standard corpus, apply a machine learning technique, and achieve reasonable results. The work has application in computational prediction of the functional significance of protein sites as well as in curation workflows for databases that capture this information.

Bush et al describe workflows for the analysis of genetic data in the context of electronic medical records (EMRs). Using EMR data in conjunction with genetic data is an important step in the study of both genetic and environmental factors related to complex human diseases, but analyses combining these data pose substantial privacy concerns. This contribution discusses such concerns, as well as a system that has been developed to allow such analyses via a web server while maintaining appropriate privacy for individuals participating in the study.

Seedorff et al seek to extend a taxonomy of health-related terms, the Mayo Consumer Health Vocabulary (MCV), that helps customers understand the terminology used by healthcare professionals. The authors argue for the importance of integrating synonyms for medical terminologies as well as both genetic risk factors and non-genetic risk factors for diseases into MCV, and present a method for automatically extending it using text mining. The successful extension of MCV can then form a basis to build consumer- oriented products and sophisticated search and information retrieval standards for patient-facing applications.

References

1. Cohen AM, Hersh WR, A survey of current work in biomedical text mining, Briefings in Bioinformatics, 2005, 6: 57-71

2. Langley, P., Lessons for the computational discovery of scientific knowledge. In Proceedings of First International Workshop on Data Mining Lessons Learned, 2002, p. 9-12.

3. Yang, Q., Xindong Wu, 10 Challenging Problems in Data Mining Research, International Journal of Information Technology & Decision Making, Vol. 5, No. 4 (2006) 597–604.

ENABLING HIGH-THROUGHPUT GENOTYPE-PHENOTYPE ASSOCIATIONS IN THE EPIDEMIOLOGIC ARCHITECTURE FOR GENES LINKED TO ENVIRONMENT (EAGLE) PROJECT AS PART OF THE POPULATION ARCHITECTURE USING GENOMICS AND EPIDEMIOLOGY (PAGE) STUDY

WILLIAM S. BUSH[*]

Department of Biomedical Informatics, Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: william.s.bush@vanderbilt.edu

JONATHAN BOSTON*

Center for Human Genetics Research, Vanderbilt University, 1207 17th Avenue, Suite 300
Nashville, TN 37232, USA
Email: boston@chgr.mc.vanderbilt.edu

SARAH A. PENDERGRASS

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 503 Wartik Lab
University Park, PA 16802, USA
Email: sap29@psu.edu

LOGAN DUMITRESCU, ROBERT GOODLOE

Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: logan.dumitrescu@chgr.mc.vanderbilt.edu; robert.j.goodloe@vanderbilt.edu

KRISTIN BROWN-GENTRY, SARAH WILSON, BOB MCCLELLAN, JR

Center for Human Genetics Research, Vanderbilt University, 1207 17th Avenue, Suite 300
Nashville, TN 37232, USA
Email: kristin.brown@chgr.mc.vanderbilt.edu; sarah.wilson@chgr.mc.vanderbilt.edu; bob.mcclellan@chgr.mc.vanderbilt.edu

ERIC TORSTENSON

Center for Human Genetics Research, Vanderbilt University, 2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: eric.torstenson@chgr.mc.vanderbilt.edu

[*] Contributed equally to the work

MELISSA A. BASFORD

Office of Research, Office of Personalized Medicine, Vanderbilt University, 2525 West End Avenue
Nashville, TN 37203, USA
Email: melissa.basford@vanderbilt.edu

KYLEE L. SPENCER

Biology and Environmental Science, Heidelberg University, Bareis Hall 131, 310 East Market Street
Tiffin, OH 44883, USA
Email: kspencer@heidelberg.edu

MARYLYN D. RITCHIE

Center for System Genomics, Department of Biochemistry and Molecular Biology, , Pennsylvania State University,
512 Wartik Lab
University Park, PA 16802, USA
Email: marylyn.ritchie@psu.edu

DANA C. CRAWFORD

Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University,
2215 Garland Avenue, 519 Light Hall
Nashville, TN 37232, USA
Email: crawford@chgr.mc.vanderbilt.edu

Genetic association studies have rapidly become a major tool for identifying the genetic basis of common human diseases. The advent of cost-effective genotyping coupled with large collections of samples linked to clinical outcomes and quantitative traits now make it possible to systematically characterize genotype-phenotype relationships in diverse populations and extensive datasets. To capitalize on these advancements, the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) project, as part of the collaborative Population Architecture using Genomics and Epidemiology (PAGE) study, accesses two collections: the National Health and Nutrition Examination Surveys (NHANES) and BioVU, Vanderbilt University's biorepository linked to de-identified electronic medical records. We describe herein the workflows for accessing and using the epidemiologic (NHANES) and clinical (BioVU) collections, where each workflow has been customized to reflect the content and data access limitations of each respective source. We also describe the process by which these data are generated, standardized, and shared for meta-analysis among the PAGE study sites. As a specific example of the use of BioVU, we describe the data mining efforts to define cases and controls for genetic association studies of common cancers in PAGE. Collectively, the efforts described here are a generalized outline for many of the successful approaches that can be used in the era of high-throughput genotype-phenotype associations for moving biomedical discovery forward to new frontiers of data generation and analysis.

1. Introduction

In a typical genome-wide association study (GWAS), a single or limited number of traits or diseases are tested for association with common single nucleotide polymorphisms (SNPs) assayed regardless of presumed function across the human genome. Since 2005, GWAS has been successful in confirming already known and identifying novel genotype-phenotype associations relevant to the biomedical community. GWAS is now a mainstay discovery approach in human genetics.

With hundreds to thousands of genotype-phenotype associations now catalogued across the human genome(1,2), there is great interest in expanding the characterization of these associations beyond the initial population or phenotype studied. Indeed, the systematic characterization and fine-mapping of known GWAS-identified variants from European-descent populations has begun in earnest(3-10). In addition, large scale methods to identify pleiotropy, such as phenome-wide association studies (PheWAS) (11,12), are increasing in frequency. To propel research in these two avenues, the National Human Genome Research Institute founded the Population Architecture using Genomics and Epidemiology (PAGE) study in 2008. PAGE is a collection of large, diverse epidemiologic and clinical collections with DNA samples linked to hundreds of disease outcomes, quantitative traits, and exposures(13)(Figure 1). A major activity of the PAGE study is the systematic characterization of GWAS-identified genotype-phenotype relationships across populations and phenotypes. The Epidemiologic Architecture for Genes Linked to Environment (EAGLE) project, one of PAGE's four study sites, accesses the National Health and Nutrition Examination Surveys (NHANES) and the Vanderbilt University biorepository linked to de-identified electronic medical records (BioVU)(14) to pursue PAGE study goals.

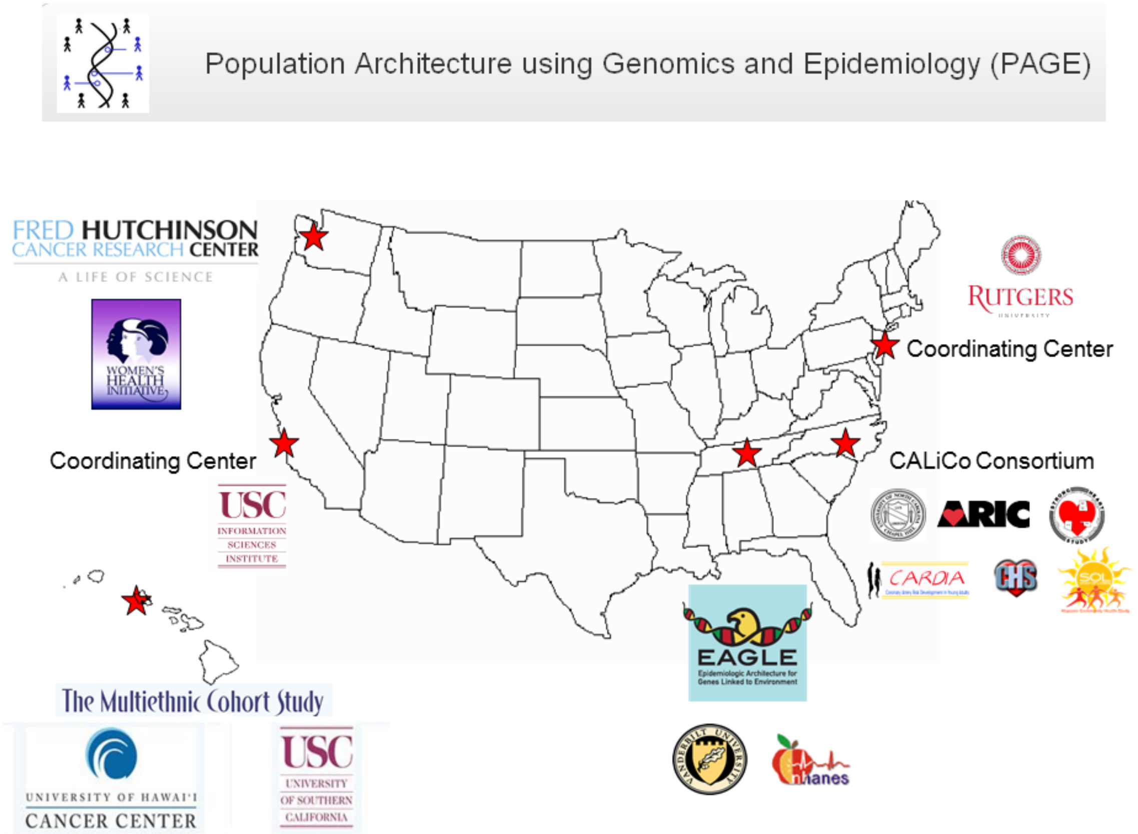
EAGLE participates in collaborative PAGE studies for disease and traits related to cardiovascular, metabolic, and cancer phenotypes among many others. To enable characterization of genotype-phenotype relationships in EAGLE and PAGE, EAGLE has developed high-throughput workflows customized to test GWAS-identified variants for all outcomes and traits in multiple populations available in both EAGLE collections. The development of a systematic workflow was and continues to be necessary to harmonize EAGLE analyses with analyses from other PAGE study sites and to facilitate meta-analysis across multiple studies. We describe herein each EAGLE collection, including characteristics of each data collection that impact both the workflow design for effective data analysis as well as data sharing, all crucial elements for collaborative high-throughput human genetic association studies for biomedical discovery.

2. Methods

1.1. *Study populations*

EAGLE currently accesses two diverse study populations as part of the PAGE study: the National Health and Nutrition Examination Surveys (NHANES) and BioVU, the Vanderbilt University biorepository linked to de-identified electronic medical records (EMRs). NHANES is a population-based survey conducted by the National Center for Health Statistics at the Centers for Disease Control and Prevention(15). NHANES ascertains Americans regardless of health status at the time of the survey. For each study participant, data on demographics, health, and lifestyle are collected. A physical exam is conducted by a CDC physician or health professional, and laboratory measures are assayed from blood and urine. DNA samples were collected on consenting participants for the Third NHANES (NHANES III) conducted between 1991 and 1994 (n=7,159), NHANES 1999-2000 (n=3,570), NHANES 2001-2002 (n=4,269), and NHANES 2007-2008 (n=4,615). A total of 19,613 DNA samples are available for research representing self-reported non-Hispanic whites (n=8,858), non-Hispanic blacks (n=4,325), Mexican Americans (n=4,768), and other race/ethnicities (n=1,662).

Figure 1. The Population Architecture using Genomics and Epidemiology (PAGE) study. The PAGE study, funded in 2008, consists of a coordinating center (Rutgers University and Information Sciences Institute at the University of Southern California) and four study sites: the Causal Variants Across the Life Course (CALiCo) consortium accessing the Atherosclerosis Risk in Communities (ARIC), Coronary Artery Risk in Young Adults (CARDIA), Cardiovascular Heart Study (CHS), Strong Heart Cohort and Family Studies (SHS/SHFS), and Study of Latinos (SOL); Epidemiologic Architecture for Genes Linked to Environment (EAGLE) accessing the National Health and Nutrition Examination Surveys (NHANES) and Vanderbilt University's biorepository linked to de-identified medical records (BioVU); the Multiethnic Cohort (MEC); and the Women's Health Initiative (WHI).


Population Architecture using Genomics and Epidemiology (PAGE)
FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE
WOMEN'S HEALTH INITIATIVE
Coordinating Center
USC INFORMATION SCIENCES INSTITUTE
RUTGERS UNIVERSITY
Coordinating Center
CALiCo Consortium
ARIC
CARDIA
CHS
SOL
EAGLE
Epidemiologic Architecture for Genes Linked to Environment
The Multiethnic Cohort Study
UNIVERSITY OF HAWAI'I CANCER CENTER
USC UNIVERSITY OF SOUTHERN CALIFORNIA
VANDERBILT UNIVERSITY
nhanes

In contrast to NHANES, BioVU is a clinic-based collection of patients visiting the outpatient clinics affiliated with Vanderbilt University in Nashville, Tennessee(14). DNA is extracted from discarded blood collected for routine outpatient clinic use and linked to a de-identified version of the electronic medical record known as the Synthetic Derivative (SD). The SD is updated routinely and contains outpatient as well as inpatient clinical structured and unstructured data including billing codes, procedure codes, labs, tumor registry entries, demographic data, vital signs, and text-based clinical notes. Because of extensive de-identification procedures, BioVU is considered non-Human Subjects research(16). As of June 2012, BioVU contained 143,993 DNA samples, 57% of which are from females and 10% from African Americans.

2.2. Genotyping

The majority of EAGLE's genotypic data are a result of *de novo* targeted genotyping. Briefly, SNPs were selected in 2008 to mid-2010 representing index genetic variants from GWAS of common diseases and traits such as HDL-C, LDL-C, triglycerides, total cholesterol, markers of inflammation, bone mineral density/osteoporosis, electrocardiographic traits, body mass index, complete blood count traits, type 2 diabetes and eight major cancers. SNPs were then genotyped using a variety of assays/platforms including TaqMan, TaqMan OpenArray, Illumina BeadXpress, and Sequenom. To date, EAGLE has submitted greater than 5.1 million genotypes to the CDC Genetic NHANES database, and these data are available for secondary analyses via NCHS/CDC.

2.3. Statistical analyses

In EAGLE (single site) and PAGE (multi-site) studies, genotype-phenotype association analyses are conducted as defined by the following "tiers"(13):

- Tier 1: High-throughput unadjusted linear or logistic regressions assuming an additive genetic model. For categorical phenotypes, binning was used to create new variables of the form "A versus not A" for each category, and logistic regression was used to model the new binary variable. All continuous phenotypes were natural log transformed, following a y to log (y+1) transformation of the response variable with +1 added to all continuous measurements before transformation to prevent variables recorded as zero from being omitted from analysis. All analyses are stratified by race/ethnicity. Statistical analyses are performed by each PAGE study site independently. The phenotypic and exposure variables are not harmonized across PAGE study sites.
- Tier 2: Low throughput unadjusted linear and/or logistic regressions performed for select genotypes and phenotypes of interest in a single PAGE study site. The genetic modeling and levels of stratification are dependent on a specific hypothesis or study question. The study subjects are carefully phenotyped and multiple covariates (also well-defined) are considered in the models.
- Tier 3: Low throughput unadjusted linear and/or logistic regressions performed for select genotypes and phenotypes of interest across PAGE study sites where the genetic modeling and levels of stratification are dependent on the hypothesis or study question. The study subjects are carefully phenotyped like Tier 2 analyses; however for Tier 3, phenotypes and exposures are harmonized across multiple PAGE study sites. Statistical analyses are performed by each PAGE study site independently, and aggregate results are shared across study sites for meta-analysis by the lead author(s).

All PAGE study results, regardless of Tier, must be available in aggregate form for the PAGE Coordinating Center browser(13) and possible dbGaP(17) deposition. To facilitate the uniform submission of PAGE study aggregate data by study site, the PAGE Coordinating Center created three "Results Template" files consisting of the phenotype file, the SNP file, and the Association file (version 8). The phenotype file currently consists of 32 column headers such as phenotype label, PAGE study site, phenotype units, information on transformation and analysis tier, type of variable (binary versus quantitative), types of covariates included in the models, race/ethnicity, gender, sample size, and descriptive statistics of the phenotype used in the analysis. The SNP files currently consists of 19 column headers such SNP ID (rs number), PAGE study site, race/ethnicity, gender, alleles and counts (including coded allele designation), genotypes and counts, Hardy Weinberg p-values, genotype call rates, and strand information. The Association file currently consists of 53 column headers such as SNP ID, phenotype, PAGE study site, race/ethnicity, gender, genetic effect size of association and standard errors and/or confidence intervals, modeling label (defined by lead of the analysis plan), p-values, sample sizes, alleles (included allele and frequency of coded allele), genotype counts by affection status, median values and quartiles of quantitative traits by genotype, and genetic model.

In EAGLE, all NHANES genotype-phenotype associations are performed using SAS v9.2 and SUDAAN v10.0(SAS Institute, Cary, NC) using the Analytic Data Research by Email (ANDRE) portal of the CDC Research Data Center (RDC) in Hyattsville, MD (further described below). EAGLE analyses accessing BioVU data are performed using a variety of software packages including PLINKv1.07(18), SASv9.3, and Rv2.14.1(19). The EAGLE workflows described here are supported by multiple scripts written in several computer languages such Ruby with Ruby on Rails framework and Javascript with Backbone framework.

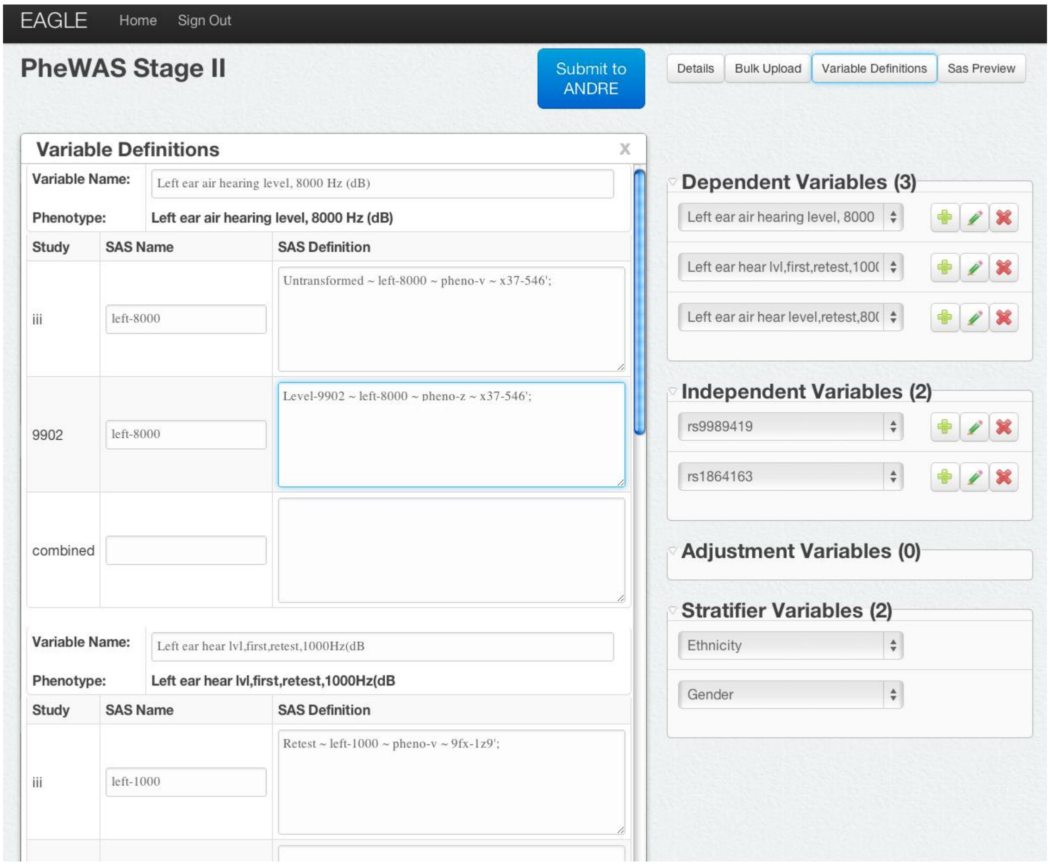
3. Workflow

3.1. The epidemiologic collection (NHANES)

Like many epidemiologic collections, NHANES consists of thousands of DNA samples linked to thousands of variables and, in the case of EAGLE, hundreds of genetic variants. To automate the high-throughput genotype-phenotype associations such as the PheWAS approach, the workflow for this and many epidemiologic collections must accommodate the fact that sample size, phenotypic/exposure variable list, and genetic variant content can vary substantially across the years of survey. Also, the workflow must acknowledge and work with various data access models that can range from open access to highly restricted access to individual level data within and across collaborating studies. Finally, the workflow must anticipate high volumes of structured data that will require accessible archival or storage for specialized searches.

Specifically for NHANES, EAGLE accesses up to 19,613 DNA samples that have anywhere from one to 1,100 genetic variants and approximately 3,500 phenotypic/demographic variables available for analysis. Due to concerns related to confidentiality even for aggregate data(20), genetic data are considered restricted variables by CDC and therefore cannot be linked to phenotypic variables and accessed outside of the CDC RDC firewall. To facilitate analyses such as genotype-phenotype association studies for research groups outside of CDC, the RDC created Analytic Data Research by Email (ANDRE). ANDRE is the remote server for CDC that accepts and runs analyses generated in Statistical Analysis System (SAS) or Survey Data Analysis (SUDAAN). ANDRE is an e-mail exchange that serves as an interface for processing code. Only analyses or SAS commands that result in aggregate data are allowed, and specific SAS commands and macros are explicitly forbidden. SAS output resulting from analyses sent to ANDRE by outside investigators are further inspected to ensure that counts fewer than five are redacted or suppressed from the output before the output is returned to outside investigators for consumption. And, ANDRE e-mail exchange is limited to outgoing files <20MB in size, which includes both the log and output files. The time elapsed between submitting code to ANDRE and receiving the output files from ANDRE via e-mail is typically less than 30 minutes, but this can range from two minutes to several hours.

Figure 2. EAGLE project web-based Experiment Designer. We developed a web-based Experiment Designer to assist EAGLE analysts in generating standard SAS code for high-throughput genotype-phenotype tests of association. The SAS designer allows each EAGLE analyst to create experiments by selecting pre-defined variables approved for study by CDC by NHANES dataset. EAGLE analysts can also specify dependent variables, independent variables, and stratification variables (gender and race/ethnicity) for linear or logistic regression modeling. The SAS Generator takes the experiment created with the Experiment Designer and generates the appropriate SAS code for submission to ANDRE.



PheWAS Stage II

Variable Definitions

Variable Name: Left ear air hearing level, 8000 Hz (dB)

Phenotype: Left ear air hearing level, 8000 Hz (dB)

Study	SAS Name	SAS Definition
iii	left-8000	Untransformed ~ left-8000 ~ pheno-v ~ x37-546';
9902	left-8000	Level-9902 ~ left-8000 ~ pheno-z ~ x37-546';
combined		

Variable Name: Left ear hear lvl,first,retest,1000Hz(dB

Phenotype: Left ear hear lvl,first,retest,1000Hz(dB

Study	SAS Name	SAS Definition
iii	left-1000	Retest ~ left-1000 ~ pheno-v ~ 9fx-1z9';

Dependent Variables (3)

- Left ear air hearing level, 8000
- Left ear hear lvl,first,retest,100(
- Left ear air hear level,retest,80(

Independent Variables (2)

- rs9989419
- rs1864163

Adjustment Variables (0)

Stratifier Variables (2)

- Ethnicity
- Gender

The restrictions posed by the RDC present several challenges for high-throughput genotype-phenotype associations in EAGLE and for data sharing with the PAGE study sites. To work within the restrictions and to minimize analyst workload, we created a web-based "Experiment Designer" and "SAS Generator". With the Experiment Designer (Figure 2), analysts create and edit the variables for an experiment that will be sent to ANDRE. Analysts can then select dependent and independent variables along with any adjustments and stratifications. The Experiment Designer allows analysts to focus on the data and desired results instead of the SAS code itself. The Experiment Designer also ensures uniform SAS coding of the genetic model (and coded allele), an important feature for large datasets accessed by three analysts at any one time. The SAS Generator then takes the experiment created with the Experiment Designer and generates the appropriate SAS code for submission to ANDRE. Each experiment can be queued and sent to ANDRE when output from the previous experiment is received by the analyst via e-mail. Thus, the SAS Generator ensures that there are no gaps between sending SAS code and receiving output from ANDRE. The SAS Generator ultimately saves the analyst time from constantly checking e-mail for receipt of ANDRE output. To date, EAGLE analyses for EAGLE and PAGE study analysis plans have generated >400 experiments resulting in >20,000,000 SAS output files each with approximately 50 lines of unstructured SAS data output.

Most tests of association performed in NHANES result in tens of thousands of SAS output files from ANDRE. With so many output files and lines of data per output file, a second major challenge is translating the output into a condensed, accessible, and readily available format. For each set of output we have developed the "Parser" software to do the following: 1) parse the file headers to classify the files (e.g. Linear Regression, SNP Frequency, etc), and 2) process the text of each SAS output file and extract the appropriate data values. The Parser can be utilized only when necessary, allowing EAGLE analysts to store the SAS output files and then process them in real-time, as needed. This also allows EAGLE analysts to view any single output file and also view the parsed results.

Once the SAS output file results are parsed, the data are compiled into the PAGE Coordinating Center Results Template file format. To automate this process, we created the "Template Generator" step. In this step, an experiment's SAS output files are parsed and combined into a template for submission to the PAGE Coordinating Center and to PAGE collaborators for meta-analysis or for visualization using Synthesis-View(21), PheWAS-View(22), or other software. Automation of this step results in analysis results required for meta-analysis or dbGaP submission.

The full epidemiologic workflow for EAGLE, from SAS code generation to Results Template file generation for data dissemination, is given in Figure 3. The code is open source and will be available on the EAGLE website (https://eagle.mc.vanderbilt.edu/).

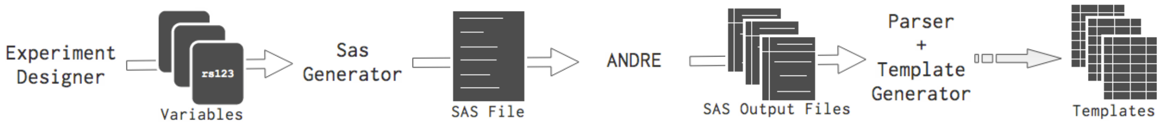


Figure 3. EAGLE project epidemiologic collection workflow. The epidemiologic collection workflow begins with the Experiment Designer, designed as a web-interface and accessed by EAGLE analysts. The analyst can easily use the Experiment Designer to create standardized SAS code based on parameters set by the analysts. The resultant ANDRE-friendly code is automatically generated. Once the code has been submitted, ANDRE will send censored output files back to the EAGLE analysts. These resultant files are first crudely parsed and stored in a database in preparation for "real-time" parsing by analysts. Finally, analysts use the "Template Generator" to create standard PAGE Results Template files for sharing data across PAGE study sites for meta-analysis.

3.2. The clinical collection (BioVU)

The epidemiologic collection of NHANES described above is an extensive and rich source of phenotypic and genotypic data for genetic association studies of quantitative traits; however, because of the wide age range and lack of health information for specific diseases, the collection is underpowered for many diseases, including common diseases such as cardiovascular disease, type 2 diabetes, and various cancers. To supplement EAGLE sample sizes for clinical outcomes in diverse populations, a clinical collection at Vanderbilt University known as BioVU was accessed.

Additional cancer cases and controls were first identified in BioVU using billing (ICD-9) codes. Specific cancers such as melanoma could be defined with high positive predictive values whereas others such as endometrial cancer could not. Therefore, to increase the positive predictive value of all EAGLE case/control definitions, data from the tumor registry were utilized. These data include primary site designations and histology information collected for clinical reporting purposes for the North America Association of Central Cancer Registries. A combination of the tumor registry data, along with ICD-9 billing codes, procedure codes, vital signs, and free text clinical notes, were used to identify cases for eight cancers among all patients aged 18 or greater in the SD with DNA samples using the following algorithms:

- Breast cancer: Three or more mentions of ICD-9 primary code 174 (malignant neoplasm of the female breast) and all sub-codes (denoted "*" here and throughout) on separate clinic visits OR a tumor registry entry for breast cancer AND female
- Colorectal cancer: Tumor registry entry for colorectal cancer.
- Endometrial cancer: Tumor registry entry for endometrial cancer with primary sites C540-C549, C559 AND histology not one of 9590-9989 AND female.

- Lung cancer: Tumor registry entry for lung cancer, any location and any type.
- Melanoma: Three or more mentions of ICD-9 codes 172.* (malignant melanoma of skin) OR tumor registry entry for melanoma.
- Non-Hodgkin's lymphoma: Tumor registry entry for non-Hodgkin's lymphoma with histology in ('9673', '9675', '9684', '9687', '9695', '9705', '9823', '9827'), OR (histology >= '9590' and histology <= '9596'), OR (histology >= '9670' and histology <= '9671'), OR (histology >= '9678' and histology <= '9680'), OR (histology >= '9689' and histology <= '9691'), OR (histology >= '9698' and histology <= '9702'), OR (histology >= '9708' and histology <= '9709'), OR (histology >= '9714' and histology <= '9719'), OR (histology >= '9727' and histology <= '9729').
- Ovarian cancer: Tumor registry entry for ovarian cancer AND female.
- Prostate cancer: Three or more mentions of ICD-9 codes 185.* (malignant neoplasm of prostate) OR tumor registry entry for prostate cancer.

Approximately two control samples were identified per case matched on sex, race/ethnicity, and age (within 5 years). Control samples were required to have at least two clinical narratives (clinical notes, discharge summaries, etc), with preference given to records with at least one fully documented history and physical. Records were excluded as controls if they had one or more codes for neoplasms, ICD-9 codes between 140.* and 239.*, had a tumor registry entry or had the one or more cancer related keywords in the problem list. For breast cancer, endometrial cancer, and ovarian cancer, male controls were also excluded, and for prostate cancer, female controls were excluded.

For specific cancers, controls with additional clinical data were desirable for anticipated analyses. For example, for breast cancer controls among women over 40 years of age, we required that records contain at least one mammography Bi-Rad score as 1 (negative) or 2 (benign). For colorectal cancer controls, we required for patients over 50 years of age the keyword "colonoscopy" in the problem list OR one of the following CPT codes: 45378 (colonoscopy, flexible, proximal to splenic flexure, diagnostic), 45379 (with removal), 45380 (with biopsy, single), 45381 (with directed), 45382 (with control), 45383 (with ablation of), 45384 (with removal of), 45385 (with removal of), 45386 (with dilation by), 45387 (with transendoscopic), 45391 (with endoscopic), and 45392 (with transendoscopic). Finally, for prostate cancer, we required male controls aged 40 years and greater to have at least one prostate specific antigen (PSA) level <4 and that the most recent PSA level is within the normal range.

With these algorithms implemented in the SD in late 2010/early 2011, we identified a total of 7,348 cancer cases for targeted genotyping. Race/ethnicity in the Vanderbilt University EMR and BioVU SD is administratively assigned, which we have shown is highly concordance with genetic ancestry determined by ancestry informative markers (AIMs)(23). As expected based on the overall demographics of BioVU, the majority of case samples were European American (87%). Approximately 4% of the samples were of unknown race/ethnicity and were assigned genetic ancestry via ancestry informative markers for downstream analyses (data not shown). For the first five cancers defined in the SD (breast, colorectal, melanoma, ovarian, and prostate cancers) we identified approximately two controls per case for genotyping as defined in the text above. A total of 8,996 controls were targeted for genotyping. Two controls per case of endometrial cancer, lung cancer, and non-Hodgkin's lymphoma were defined from among the genotyped control samples.

In addition to defining case and control status for genotyping, we have begun to define clinical covariates anticipated for analysis. As described above, screening data has been preferentially represented in controls for select cancers (breast, colorectal, and prostate) and is expected to be defined in cases. Environmental exposures are more difficult to define given that most of these data, if available, exist in the unstructured data (free text or clinical narrative) of the EMR. Work is on-going to define common exposures or other variables that reside in the clinical narrative such as alcohol use, physical activity, and family history using text mining and other approaches. For smoking status, we have applied an implementation of the CTAKES algorithm(24), and have also illustrated that ICD-based smoking definitions are highly specific for identifying smokers(25).

Unlike the epidemiologic collection (NHANES), the clinical collection (BioVU) is relatively free of data access restrictions. Therefore, the clinical collection workflow only utilizes the later stages of the workflow described in Figure 3. Output files from various statistical packages (such as PLINK) are parsed and Results Template files are generated for sharing among PAGE study sites and meta-analysis.

4. Conclusions

We describe here the epidemiologic (NHANES) and clinical (BioVU) collection workflows that enable high-throughput genotype-phenotype association studies and data sharing within EAGLE and the PAGE study. Both workflows were customized based on a variety of factors including data structure and data access. A major strength of this approach is that it provides the infrastructure to conduct systematic genetic analyses resulting in standardized files for data sharing and meta-analysis. A major weakness of this approach is that is requires substantial bioinformatics and computing resources and personnel to create, maintain, and implement the workflow. The preferential accessing of datasets with open access or fewer data use restrictions would assist in easing the effort required for the workflows. However, full access to local or collaborative datasets through dbGaP will still require substantial bioinformatics and computational support to fully mine the genotype-phenotype investments for high returns relevant to human disease and biology.

5. Acknowledgements

This work was supported by NIH U01 HG004798 and its ARRA supplements. The dataset(s) used for the analyses described were obtained from Vanderbilt University Medical Center’s BioVU which is supported by institutional funding and by the Vanderbilt CTSA grant UL1 TR000445 from NCATS/NIH. The Vanderbilt University Center for Human Genetics Research, Computational Genomics Core provided computational and/or analytical support for this work. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

References

1. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S., Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, **106**, 9362-9367.

2. Lettre,G., Palmer,C.D., Young,T., Ejebe,K.G., Allayee,H., Benjamin,E.J., Bennett,F., Bowden,D.W., Chakravarti,A., Dreisbach,A., *et al.* (2011) Genome-Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project. *PLoS Genet*, **7**, e1001300.

3. Dumitrescu,L., Carty,C.L., Taylor,K., Schumacher,F.R., Hindorff,L.A., Ambite,J.-L., Anderson,G., Best,L.G., Brown-Gentry,K., Buzkova,P., *et al.* (2011) Genetic Determinants of Lipid Traits in Diverse Populations from the Population Architecture using Genomics and Epidemiology (PAGE) Study. *PLoS Genet*, **7**, e1002138.

4. Haiman,C.A., Fesinmeyer,M., Spencer,K.L., Buzkova,P., Voruganti,V.S., Wan,P., Haessler,J., Francheschini,N., Monroe,K., Howard,B.V., *et al.* (2012) Consistent direction of effect for established T2D risk variants across populations: The Population Architecture using Genomics and Epidemiology (PAGE) Consortium. *Diabetes*, **61**, 1642-1647.

5. Fesinmeyer,M.D., North,K.E., Ritchie,M.D., Lim,U., Franceschini,N., Wilkens,L.R., Gross,M.D., Buzkova,P., Glenn,K., Quibrera,M., *et al.* Genetic risk factors for body mass index and obesity in an ethnically diverse population: results from the Population Architecture using Genomics and Epidemiology (PAGE) Study *Obesity (Silver Spring)* (in press).

6. Zhang,L., Spencer,K.L., Voruganti,V.S., Jorgensen,N.W., Fornage,M., Best,L.G., Brown-Gentry,K.D., Cole,S.A., Crawford,D.C., Deelman,E., *et al.* Association of functional polymorphism rs2231142 (Q141K) in *ABCG2* gene with serum uric acid and gout in four US populations: the Population Architecture using Genomics and Epidemiology (PAGE) Study *Am J Epidemiol* (in press).

7. Carty,C.L., Buzkova,P., Fornage,M., Franceschini,N., Cole,S., Heiss,G., Hindorff,L.A., Howard,B.V., Mann,S., Martin,L.W., *et al.* (2012) Associations Between Incident Ischemic Stroke Events and Stroke and Cardiovascular Disease-Related Genome-Wide Association Studies Single Nucleotide Polymorphisms in the Population Architecture Using Genomics and Epidemiology Study. *Circulation: Cardiovascular Genetics*, **5**, 210-216.

8. N'Diaye,A., Chen,G.K., Palmer,C.D., Ge,B., Tayo,B., Mathias,R.A., Ding,J., Nalls,M.A., Adeyemo,A., Adoue,V.+., *et al.* (2011) Identification, Replication, and Fine-Mapping of Loci Associated with Adult Height in Individuals of African Ancestry. *PLoS Genet*, **7**, e1002298.

9. Chen,F., Chen,G.K., Millikan,R.C., John,E.M., Ambrosone,C.B., Bernstein,L., Zheng,W., Hu,J.J., Ziegler,R.G., Deming,S.L., *et al.* (2011) Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans. *Human Molecular Genetics*, **20**, 4491-4503.

10. Haiman,C.A., Chen,G.K., Blot,W.J., Strom,S.S., Berndt,S.I., Kittles,R.A., Rybicki,B.A., Isaacs,W.B., Ingles,S.A., Stanford,J.L., *et al.* (2011) Characterizing Genetic Risk at Known Prostate Cancer Susceptibility Loci in African Americans. *PLoS Genet*, **7**, e1001387.

11. Denny,J.C., Ritchie,M.D., Basford,M.A., Pulley,J.M., Bastarache,L., Brown-Gentry,K., Wang,D., Masys,D.R., Roden,D.M., Crawford,D.C. (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, **26**, 1205-1210.

12. Pendergrass,S.A., Brown-Gentry,K., Dudek,S.M., Torstenson,E.S., Ambite,J.L., Avery,C.L., Buyske,S., Cai,C., Fesinmeyer,M.D., Haiman,C., *et al.* (2011) The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. Epidemiol.*, **35**, 410-422.

13. Matise,T.C., Ambite,J.L., Buyske,S., Carlson,C.S., Cole,S.A., Crawford,D.C., Haiman,C.A., Heiss,G., Kooperberg,C., Marchand,L.L., *et al.* (2011) The Next PAGE in Understanding Complex

Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. *American Journal of Epidemiology*, **174**, 849-859.

14. Roden,D.M., Pulley,J.M., Basford,M.A., Bernard,G.R., Clayton,E.W., Balser,J.R., Masys,D.R. (2008) Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin Pharmacol Ther*, **84**, 362-369.

15. Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS) (2012).

16. Pulley,J., Clayton,E., Bernard,G.R., Roden,D.M., Masys,D.R. (2010) Principles of Human Subjects Protections Applied in an Opt-Out, De-identified Biobank. *Clinical and Translational Science*, **3**, 42-48.

17. Mailman,M.D., Feolo,M., Jin,Y., Kimura,M., Tryka,K., Bagoutdinov,R., Hao,L., Kiang,A., Paschall,J., Phan,L., *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*, **39**, 1181-1186.

18. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J., Sham,P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analysis. *Am J Hum Genet*, **81**, 559-575.

19. R Development Core Team (2008) R Foundation for Statistical Computing, Vienna, Austria.

20. Homer,N., Szelinger,S., Redman,M., Duggan,D., Tembe,W., Muehling,J., Pearson,J.V., Stephan,D.A., Nelson,S.F., Craig,D.W. (2008) Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genet*, **4**, e1000167.

21. Pendergrass,S., Dudek,S., Crawford,D., Ritchie,M. (2010) Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis. *BioData Mining*, **3**, 10.

22. Pendergrass,S., Dudek,S., Crawford,D., Ritchie,M. (2012) Visually integrating and exploring high throughput Phenome-Wide Association Study (PheWAS) results using PheWAS-View. *BioData Mining*, **5**, 5.

23. Dumitrescu,L., Ritchie,M.D., Brown-Gentry,K., Pulley,J.M., Basford,M., Denny,J.C., Oksenberg,J.R., Roden,D.M., Haines,J.L., Crawford,D.C. (2010) Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. *Genet Med*, **12**, 648-650.

24. Liu,M., Shah,A., Min,J., Peterson,N.B., Dia,Q., Aldrich,M.C., Chen,Q., Bowton,E.A., Liu,H., Denny,J.C., Xu,H. A study of transportability of an existing smoking status detection module across institutions *AMIA Annu Symp Proc* (in press).

25. Wiley,L.K., Shah,A., Xu,H., and Bush,W.S. (2012) ICD-9 tobacco use codes are effective identifiers of smoking status. Translational Bioinformatics Conference, October 14-17, Jeju Island, Korea.

ATHENA: A TOOL FOR META-DIMENSIONAL ANALYSIS APPLIED TO GENOTYPES AND GENE EXPRESSION DATA TO PREDICT HDL CHOLESTEROL LEVELS

EMILY R. HOLZINGER[†]
Center for Human Genetics Research, Vanderbilt University
Nashville, TN 37232, USA
Email: emily.r.holzinger@vanderbilt.edu

SCOTT M. DUDEK
Center for Systems Genomics, Pennsylvania State University
University Park, PA 16803, USA
Email: sud23@psu.edu

ALEX T. FRASE
Center for Systems Genomics, Pennsylvania State University
University Park, PA 16803, USA
Email: alex.frase@psu.edu

RONALD M. KRAUSS
Children's Hospital Oakland Research Institute
Oakland, CA 94609, USA
Email: rkrauss@chori.org

MARISA W. MEDINA
Children's Hospital Oakland Research Institute
Oakland, CA 94609, USA
Email: mwmedina@chori.org

MARYLYN D. RITCHIE
Center for Systems Genomics, Pennsylvania State University
University Park, PA 16803, USA
Email: marylyn.ritchie@psu.edu

Technology is driving the field of human genetics research with advances in techniques to generate high-throughput data that interrogate various levels of biological regulation. With this massive amount of data comes the important task of using powerful bioinformatics techniques to sift through the noise to find true signals that predict various human traits. A popular analytical method thus far has been the genome-wide association study (GWAS), which assesses the association of single nucleotide polymorphisms (SNPs) with the trait of interest. Unfortunately, GWAS has not been able to explain a substantial proportion of the estimated heritability for most complex traits. Due to the inherently complex nature of biology, this phenomenon could be a factor of the simplistic study design. A more powerful analysis may be a systems biology approach that integrates different types of data, or a *meta-dimensional* analysis. For this study we used the Analysis Tool for Heritable and Environmental Network Associations (ATHENA) to integrate high-throughput SNPs and gene expression variables (EVs) to predict high-density

[†] Work supported by R01 LM010040, T32 GM080178, U01 HL065962 (P-STAR), U19 HL69757-10 (PARC)

lipoprotein cholesterol (HDL-C) levels. We generated multivariable models that consisted of SNPs only, EVs only, and SNPs + EVs with testing r-squared values of 0.16, 0.11, and 0.18, respectively. Additionally, using just the SNPs and EVs from the best models, we generated a model with a testing r-squared of 0.32. A linear regression model with the same variables resulted in an adjusted r-squared of 0.23. With this systems biology approach, we were able to integrate different types of high-throughput data to generate meta-dimensional models that are predictive for the HDL-C in our data set. Additionally, our modeling method was able to capture more of the HDL-C variation than a linear regression model that included the same variables.

1. Introduction

1.1. *A Case for Meta-dimensional Analysis*

Over the past decade, high-throughput technology has become considerably more efficient and less expensive[1]. The human genetics field has reaped the benefits of these advancements via extensive exploratory analyses largely in the form of GWAS. These studies have led to the discovery of thousands of SNPs that are significantly associated with hundreds of common, complex human traits[2]. However, for many of these traits, a large proportion of the estimated heritability remains unexplained by these DNA variants[3].

One of the leading hypotheses regarding this "missing heritability" is that GWAS may not be robust to the inherent complexity of biological processes, and, therefore, may be missing large chunks of the underlying etiology[4]. Two areas where this complexity might lie are in non-additive interactions (gene-gene or gene-environment) and within the different levels of biological regulation. First, because traditional GWAS specifically identify SNPs with large main effects, interactions without large main effects would be missed. Next, complex phenotypes could be under the influence of more than one level of biological regulation. Various types of –omic data (i.e. transcriptomic and methylomic) analyzed simultaneously could take into account trait variation that would be missed by SNP data alone[5]. In order to account for complex etiology, a more powerful *meta-dimensional* analysis would have to be performed. A meta-dimensional analysis is one that integrates different types of high-throughput data while allowing for non-linear interactions in order to identify multi-variable prediction models that include data from from different levels of biological regulation[6]. For example, analyzing microarray gene expression data and SNP genotypes data simultaneously to identify models that predict a complex human disease, such as breast cancer.

In order to successfully perform a meta-dimensional analysis, computational tools need to be able to perform the following tasks successfully: sift through the high level of noise inherent to high-throughput data in order to identify true signals, simultaneously analyze continuous and categorical predictor and outcome variables, and identify main and interaction effects in order to generate a final predictive model. Currently, no single analysis method performs all of these tasks at once. Some candidates that may come together to create a successful analysis pipeline include tree-based methods (i.e. Random Forests[7]), Bayesian networks, computational evolution methods, and various types of clustering and correlation techniques. For this paper, we propose a meta-dimensional analysis tool called ATHENA that combines a tree-based filtering method with a computational evolution modeling method in order to integrate SNP genotypes and gene expression variables to predict HDL-C levels.

1.2. *The Genetics of HDL Cholesterol*

HDL particles are small, dense lipoproteins that circulate throughout the body. Many anti-atherogenic properties have been ascribed to HDL, and low HDL-C levels are strongly and independently associated with increased risk for cardiovascular disease[8]. HDL-C has a relatively large genetic component with heritability estimates between 40-80%[8,9]. Many common variants have been found to be significantly associated with HDL-C in humans, but collectively they only explain a small proportion of the estimated heritability. A recent study used significant GWAS SNPs to perform polygenic scoring and found that the best model only explained ~4.75% of the variation in the HDL-C trait[10]. Some groups have begun to examine a more complex genetic architecture to explain the missing heritability and several gene-gene interactions have been identified[11–13]. In this study, we aim to go a step further by integrating SNPs and gene expression data to find complex models that predict HDL-C levels.

2. Methods

2.1. *The Analysis Tool for Heritable and Environmental Network Associations (ATHENA)*

ATHENA is a multi-functional software package designed by our lab to analyze various types of high-throughput data in order to generate multi-variable models. ATHENA has been tested extensively on simulated data and applied to biological data sets in order to demonstrate its utility on “noisy” data[14–17]. Figure 1 shows the full current and future functionality of ATHENA.

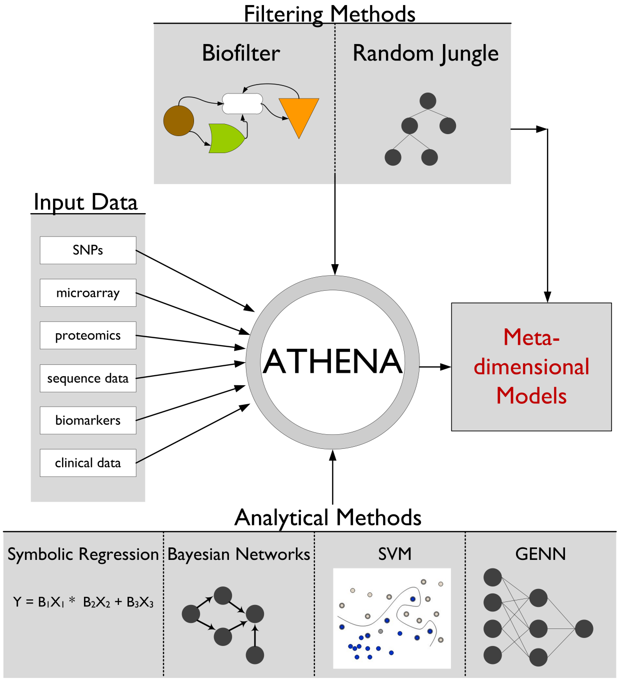


Fig. 1. Components of the ATHENA software package

The main components of ATHENA are a filtering step and a modeling step. The filtering step can be a statistical filter (Random Jungle[18]) or one that prioritizes variables based on their known biological functions (Biofilter[19]). Currently, ATHENA has two different computational evolution modeling techniques--Grammatical Evolution Symbolic Regression (GESR) and Grammatical Evolution Neural Networks (GENN). For this analysis, we used Random Jungle (RJ) as the statistical filter and Grammatical Evolution Neural Networks (GENN) as the modeling technique.

2.1.1. *ATHENA filtering: Random Jungle*

RJ is a faster, parallelized version of the tree-based variable selection method Random Forests (RF). Briefly, RF uses a bootstrap sample of the data to grow a "forest" of decision or regression trees with no pruning. The trees are then tested using the out-of-bag individuals not present in the bootstrap sample to determine which variables are most important for outcome prediction. Importantly, RF can identify main and interaction effects[7]. We chose RJ as the statistical filter because of its capability to analyze millions of quantitative and categorical variables in a relatively computationally efficient manner. Also, the output is a list of variables ranked by an importance score. For this analysis, importance is defined as the percent increase in mean squared error after permuting the variable values while taking into account correlation patterns between the variables[20]. This output lends itself nicely to selecting a subset of variables for input into a modeling technique that is less robust to noise.

2.1.2. *ATHENA modeling: Grammatical Evolution Neural Networks*

GENN uses a variation of genetic programming (GP) called grammatical evolution (GE) to optimize artificial neural networks to identify a model that predicts a given outcome[21–23]. GP is a computational technique that uses concepts of survival of the fittest in order to evolve a fit solution from an original population of random solutions[24]. GE is a more efficient version of GP because the solutions are represented as binary strings, which can be translated into a functional solution, or computer program, via grammar rules[25]. All of the evolutionary operations that are applied to the solutions are done so at the level of the binary string. Below is the algorithm that GENN uses to identify the "fittest" solution:

1. Divide the data into five equal parts for cross-validation (4/5 = training set; 1/5 = testing set).
2. Generate random sub-populations, or demes, of binary strings across multiple processors.
3. Calculate the fitness (i.e. balanced accuracy or mean squared error) of the solutions using the training set.
4. Select the solutions with the highest fitness, which undergo crossover, mutation, migration between demes, and reproduction to create the next generation of solutions.
5. Repeat Steps 3-4 for a user-defined number of generations.
6. Test the final best model using the testing set and save the model.
7. Repeat steps 2-6 for each the other four cross-validation data divisions.

8. Select the overall best model out of the five models using cross-validation consistency first and then testing set fitness to break ties.

The solutions in GENN are artificial neural networks (ANNs). Briefly, ANNs are directed graphs with an input layer (independent variables), hidden layer(s) (processing elements), and an output layer that predicts the outcome of interest[26]. Figure 2 illustrates an example of a two-layer ANN. ANNs are a good candidate for this type of analysis because they are able to model complex, non-linear relationships between variables. Traditionally, ANNs are optimized using a hill-climbing algorithm, such as back-propagation, which iteratively alters the weights (or constants) until prediction no longer improves[23]. This optimization technique is not ideal for a genetic analysis where the correct variables and the network architecture are not known a priori. GENN addresses this issue by evolving the ANNs so that the data drives the optimization of all aspects of the network. GENN has been tested on simulated and biological data and was often found to outperform other prediction techniques[16,22,27].

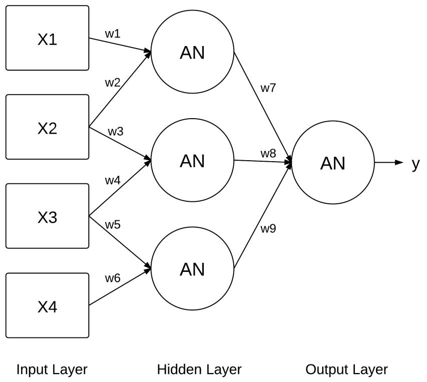


Fig. 2. An example of a two-layer ANN. X=input variable; w=weight; AN=activation node; y=predicted output

2.1.3. *ATHENA filtering-modeling pipeline*

Figure 3 below summarizes the filtering-modeling pipeline that was used for this analysis.

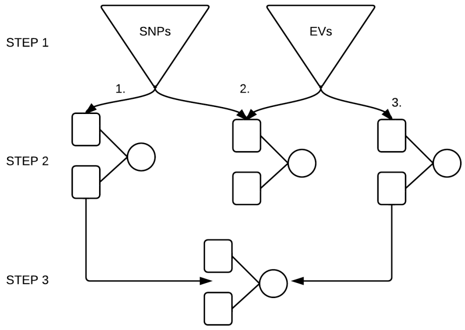


Fig. 3. ATHENA filtering-modeling pipeline for this analysis. Step 1. RJ filtering of SNPs and EVs; Step 2. GENN analysis of filtered SNPs only (2.1), EVs only (2.3) , and SNPs and EVs together (2.2); Step 3. GENN analysis of SNPs and EVs from the best GENN model from Steps 2.1 and 2.3.

In Step 1, we filtered the ~2.7 million SNPs and ~24,000 EVs separately in RJ. This was done because RJ has not been sufficiently tested to determine the effect of the overwhelmingly larger number of SNPs versus EVs that were present in this data set (~112x more SNPs). After filtering, we analyzed the filtered SNPs (Step 2.1), the filtered EVs (Step 2.3), and the filtered SNPs and EVs together (Step 2.2) in GENN. Because GENN has been shown to outperform other methods specifically at prediction modeling when the noise in the data is substantially reduced, we also assessed just the SNPs and EVs that were in the best ANN models from Steps 2.1 and 2.3 in a final GENN analysis (Step 3).

2.2. *Cholesterol and Pharmacogenetics Dataset*

The data for this study comes from the simvastatin clinical trial Cholesterol and Pharmacogenetics (CAP)[28]. The characteristics of the 480 individuals in this analysis are shown in Table 1. The genomic data consists of ~2.7 million SNP genotype dosages and ~24,000 gene expression levels. SNPs were genotyped on Illumina HumanHap 300K BeadChip and Illumina HumanHap 610-Quad BeadChip and imputed to HapMap data using the IMPUTE2 software[29]. Imputation probabilities were used to calculate genotype dosages. Gene expression levels were measured in patient-derived immortalized lymphoblastoid cell lines (LCLs) using the Illumina HumanRef8v3 BeadArray. The gene expression data was corrected for potential confounders by extracting the residuals from a linear regression model that included known covariates (day of assay, cell count, gender, and age) and the top nine principal components for unknown covariates. Our outcome of interest was the mean HDL-C level from the first and follow-up visit before any medication was administered. HDL-C was adjusted for gender, age, body mass index (BMI), and smoking status. All of the individuals in this subset of the cohort were European-American.

Table 1. Data set characteristics

Clinical trait	**Value**
Age in years (mean [sd])	54.4 [12.7]
BMI (mean [sd])	27.6 [5.3]
HDL-C in mg/dl (mean [sd])	53.4 [16.3]
Smoker (% smoker)	13.2
Gender (% male)	54.1

3. Results

3.1. *Random Jungle*

Table 2 below lists the important parameter setting values that were used for RJ for each analysis. Table 2 also displays the computation times and the number of variables that remained after backward elimination. The values for bootstrap sample size and number of trees were previously tuned for each data set as suggested by the method developers[18].

Table 2. RJ filtering parameter settings

Parameter	EV analysis	SNP analysis
Bootstrap Sample Size	11250	684342
Number of Trees	4000	4032
Tree Type	Regression trees	Regression trees
Importance Score	Permutation-based	Permutation-based
Backward Elimination	Discard negative scores	Discard negative scores
Number of Processors	4 (500 trees / processor)	64 (63 trees / processor)
Compute Time (hours)	0.6	52
Remaining Variables	1447	209346

In order to have a comparable threshold for both data sets, we chose an importance score cut-off because it has the same statistical meaning for both the SNPs and EVs. The threshold of 10 was chosen because it generated similar distributions of scores in both data sets. This cut-off resulted in a filtered data set that consisted of 418 SNPs and 241 EVs.

3.2. *GENN*

The filtered EV and SNP variables were analyzed both separately and simultaneously by GENN. In addition, the SNPs and EVs from the best GENN models were analyzed together. Table 3 shows the GENN parameters that were used for these analyses. These parameters were selected based on a tuning analysis where we swept over various settings and selected based on prediction optimization. A detailed description of the parameters can be found in a previous ATHENA publication[14]. The fitness function used by GENN for analysis of quantitative outcomes is shown below:

$$r - squared = 1 - \left[\frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \right] \qquad (1)$$

where y is the observed value, y-hat is the predicted value, and y-bar is the mean value for the quantitative outcome.

Table 3. GENN parameter settings

Parameter	Steps 2.1, 2.3	Steps 2.2, 3
Number of demes (processors)	100	100
Population Size / Deme	3000	1000
Number of generations	1125	250
Number of migrations	45	10
Probability of Crossover	0.9	0.9
Probability of Mutation	0.01	0.01
Fitness	r-squared	r-squared
Analysis time (hours)	8	1

Figure 4 shows the resulting best ANN models from each of the following analyses: a. SNPs only (Step 2.1), b. EVs only (Step 2.3), and c. SNPs and EVs together (Step 2.2). The r-squared values from the testing cross-validation set for each of the models were 0.16, 0.11, and 0.18, respectively.

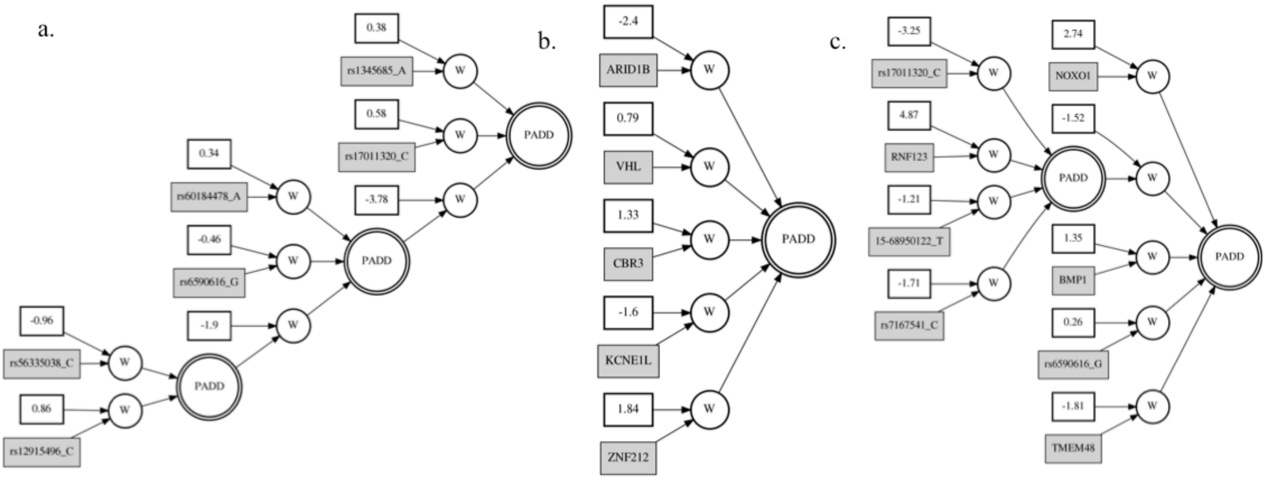


Fig. 4. Best GENN models from the a. SNP, b. EV, and c. SNP and EV integrated analyses. The asterisks in the integrated model denote variables that were present in at least one of the top five cross validation models from the separate SNP and EV analyses. (w = constant and variable are multiplied; PADD = additive activation node)

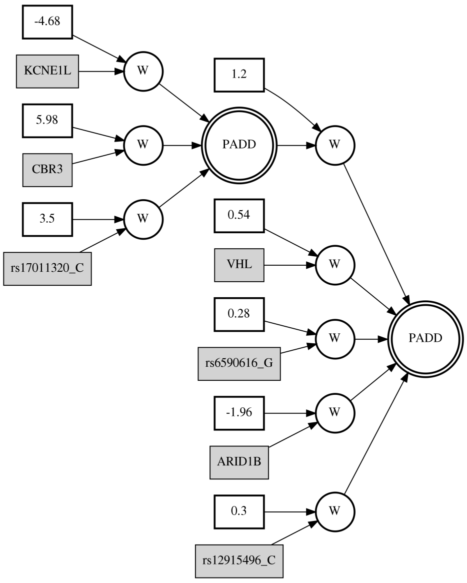


Fig. 5. Best model GENN analysis of variables from best SNP and EV models. Testing r-squared value = 0.32.

Finally, we ran GENN with only the 6 SNPs and 5 EVs that were present in the top models shown Figure 4a. and 4b. Figure 5 shows the resulting network from this analysis (Step 3). The ANN consisted of 3/6 SNPs and 4/5 EVs from the best models and the testing r-squared value was 0.32. This is substantially greater than the three previous networks (Figure 4). Additionally, we tested the same variables using a more traditional statistical prediction method--multivariable linear regression. The adjusted r-squared value from the regression model that included all 6 SNPs and 5 expression variables was 0.23. The full regression model was highly significant, with a p-value of 2.2×10^{-16}.

4. Discussion

In this study, we demonstrate a filtering-modeling pipeline for integrating different types of high-throughput data to generate meta-dimensional prediction models. We were able to build a model that includes variables from different levels of biological regulation and explained more variation than either data-type alone (Figures 4 and 5). Additionally, our best model was more predictive than the commonly used additive modeling technique. Due to its flexibility, this approach is easily extendible to other types of high-throughput data. For example, another quantitative high-throughput measurement such as proteomic data could be added to this analysis by filtering the data using the same RJ method and then adding in these filtered proteomic levels to the GENN analysis.

Notably, although the ANN from the integrated analysis had a higher r-squared value than the analyses that only included SNPs or EVs (Figure 4), it was still less predictive than the analysis that only included just the top SNPs and EVs (Figure 5). This could be a result of the combined increase in pressure on variable selection due to the larger number of predictor variables and on modeling due to the different scales of the EV and SNP values. When we reduced the variable selection pressure by only including the top variables from the EV-only and SNP-only best models, the r-squared value went up substantially. This highlights the ability of GENN to model the variables in an informative way when presented with a limited number of noise variables. Additionally, the GENN model was able to account for more outcome variation than the linear regression model indicating that the more complex modeling method of GENN identifies relationships between the variables that an additive model does not.

One caveat to our approach is that we are not able to explore conditional relationships between the different types of predictor variables. An example would be a model where a SNP in a transcription factor binding site reduces the expression of the targeted gene, which, in turn, affects the phenotype. These types of relationships could be tested by first examining significant correlations between SNPs and EVs and then using this information to guide the modeling analysis. Also, some groups are applying Bayesian networks (BNs) to data integration studies because they are able to capture this type of directionality[30]. Future work will involve incorporating BNs into ATHENA as one of the analysis methods. Other study designs specifically address the hypothesis that SNPs are affecting the phenotype via their association with gene expression levels, such as eQTLs[31–34]. These studies have provided some interesting findings but would not identify SNPs and EVs that have an effect on the phenotype independently of one another.

Interpreting the biological significance of statistical models is not a trivial task for several reasons. First, due the correlation patterns that exist in SNPs and EV data, the variables in the best models could be functional variables or variables that are highly correlated with the functional variables. There is no simple way to determine which is the case. One initial approach could be to map the top ranked SNPs and EVs back to genes to determine if the variables in the best models are representative of any given biological pathway or have similar biological function. We assessed this possibility by analyzing the RJ filtered SNPs and EVs with an online annotation tool called DAVID[35,36]. The most significant biological groups after accounting for redundant pathway information in the databases were those related to immune function. This is interesting because HDL has been shown to play a role in innate and adaptive immune responses[37].

Notably, we did not identify any of the genes known to be highly associated with HDL-C. The gene that is arguably most strongly associated with HDL-C is CETP[38,39]. To determine if our method was not able to find the effects or if the effects were simply not there, we performed a univariate linear regression analysis on each of the SNPs and then ranked the p-values. None of the SNPs in CETP were significantly associated with HDL-C in our data set (data not shown). This suggests that in this subset of individuals, other genes could be more strongly contributing to the variation in HDL-C.

Once a meta-dimensional model has been identified and shown to be predictive, the next step is to replicate the finding in an independent data set. For single SNPs, this process is relatively straightforward. For meta-dimensional models, however, it becomes less trivial due to the increased difficulty of replicating the exact effects of numerous data points simultaneously, especially if the identified variables are not completely correlated with the functional variants. One part of model validation will be to determine if the model is predictive in another data set. Additionally, the functionality of these genes could be tested in vitro or in vivo to determine if perturbation has any phenotypic effect.

The ultimate goal of identifying models that explain the genetic variability of a trait is to use this information to improve therapy or prediction and prevention in a clinical setting. Methods robust to the true nature of complex traits, like the meta-dimensional analysis pipeline presented here, are an initial step towards a more thorough understanding of the genetic architecture of complex human traits like cardiovascular disease.

References

1. Pareek, C. S., Smoczynski, R. & Tretyn, A. Sequencing technologies and genome sequencing. *Journal of Applied Genetics* **52**, 413–435 (2011).
2. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362–9367 (2009).

3. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
4. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21 (2008).
5. Reif, D. M., White, B. C. & Moore, J. H. Integrated analysis of genetic, genomic and proteomic data. *Expert.Rev Proteomics.* **1**, 67–75 (2004).
6. Holzinger, E. R. & Ritchie, M. D. Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics* **13**, 213–222 (2012).
7. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
8. Boes, E., Coassin, S., Kollerits, B., Heid, I. M. & Kronenberg, F. Genetic-epidemiological evidence on genes associated with HDL cholesterol levels: A systematic in-depth review. *Experimental Gerontology* **44**, 136–160 (2009).
9. Weissglas-Volkov, D. & Pajukanta, P. Genetic causes of high and low serum HDL-cholesterol. *The Journal of Lipid Research* **51**, 2032–2057 (2010).
10. Demirkan, A. *et al.* Genetic architecture of circulating lipid levels. *European Journal of Human Genetics* **19**, 813–819 (2011).
11. Turner, S. D. *et al.* Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. *PLoS ONE* **6**, e19586 (2011).
12. Ma, L. *et al.* Knowledge-driven analysis identifies a gene-gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations. *PLoS Genet.* **8**, e1002714 (2012).
13. He, J. *et al.* Gene-based interaction analysis by incorporating external linkage disequilibrium information. *Eur. J. Hum. Genet.* **19**, 164–172 (2011).
14. Holzinger, E. R. *et al.* Initialization Parameter Sweep in ATHENA: Optimizing Neural Networks for Detecting Gene-Gene Interactions in the Presence of Small Main Effects. *Genet Evol Comput Conf.* **12**, 203–210 (2010).
15. Holzinger, E. R., Dudek, S. M., Torstenson, E. C. & Ritchie, M. D. ATHENA Optimization: The Effect of Initial Parameter Settings Across Different Genetic Models. *Lect Notes Comput Sci* **6623**, 48–58 (2011).
16. Holzinger, E. R. *et al.* Comparison of Methods for Meta-dimensional Data Analysis Using in Silico and Biological Data Sets. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics* **7246**, 134–143 (2012).
17. Turner, S. D., Dudek, S. M. & Ritchie, M. D. ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci. *BioData.Min* **3**, 5 (2010).
18. Schwarz, D. F., Konig, I. R. & Ziegler, A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics* **26**, 1752–1758 (2010).
19. Bush, W. S., Dudek, S. M. & Ritchie, M. D. Biofilter: A knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput* **In review**, (2009).
20. Meng, Y. A., Yu, Y., Cupples, L. A., Farrer, L. A. & Lunetta, K. L. Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics* **10**, 78 (2009).
21. Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W. & Ritchie, M. D. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genet Epidemiol* **32**, 325–340 (2008).
22. Motsinger-Reif, A. A., Fanelli, T. J., Davis, A. C. & Ritchie, M. D. Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error. *BMC.Res.Notes* **1**, 65 (2008).
23. Motsinger-Reif, A. A. & Ritchie, M. D. Neural networks for genetic epidemiology: past, present, and future. *BioData.Min* **1**, 3 (2008).
24. Koza, J. *Genetic Programmming*. (MIT Press: Cambridge, Massachusetts, 1993).
25. O'Neill, M. & Ryan, C. Grammatical Evolution. *IEEE Transactions on Evolutionary Computation* **5**, (2001).

26. Anderson, J. A. *An Introduction to Neural Networks*. (MIT Press: Cambridge, Massachusetts, 1995).
27. Spencer, K. L. *et al.* Using genetic variation and environmental risk factor data to identify individuals at high risk for age-related macular degeneration. *PLoS.One.* **6**, e17784 (2011).
28. Simon, J. A. *et al.* Phenotypic predictors of response to simvastatin therapy among African-Americans and Caucasians: the Cholesterol and Pharmacogenetics (CAP) Study. *Am J Cardiol* **97**, 843–850 (2006).
29. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics* **5**, e1000529 (2009).
30. Fridley, B. L., Lund, S., Jenkins, G. D. & Wang, L. A Bayesian integrative genomic model for pathway analysis of complex traits. *Genet. Epidemiol.* **36**, 352–359 (2012).
31. Huang, R. S. *et al.* A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc Natl Acad Sci U S A* **104**, 9758–9763 (2007).
32. Huang, R. S. *et al.* Genetic variants contributing to daunorubicin-induced cytotoxicity. *Cancer Res* **68**, 3161–3168 (2008).
33. Huang, R. S., Duan, S., Kistner, E. O., Hartford, C. M. & Dolan, M. E. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. *Mol Cancer Ther* **7**, 3038–3046 (2008).
34. Huang, R. S. *et al.* Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. *Am J Hum Genet* **81**, 427–437 (2007).
35. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
36. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57 (2009).
37. Norata, G. D., Pirillo, A., Ammirati, E. & Catapano, A. L. Emerging role of high density lipoproteins as a player in the immune system. *Atherosclerosis* **220**, 11–21 (2012).
38. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
39. Dullaart, R. P. F. & Sluiter, W. J. Common variation in the CETP gene and the implications for cardiovascular disease and its treatment: an updated analysis. *Pharmacogenomics* **9**, 747–763 (2008).

STATISTICAL EPISTASIS NETWORKS REDUCE THE COMPUTATIONAL COMPLEXITY OF SEARCHING THREE-LOCUS GENETIC MODELS

Ting Hu
Department of Genetics, Geisel School of Medicine
Dartmouth College, Hanover, NH 03755, USA

Angeline S. Andrew and Margaret R. Karagas
Department of Community and Family Medicine, Geisel School of Medicine
Dartmouth College, Hanover, NH 03755, USA

Jason H. Moore*
Institute for Quantitative Biomedical Sciences
Departments of Genetics and Community and Family Medicine, Geisel School of Medicine
Dartmouth College, Hanover, NH 03755, USA
**E-mail: jason.h.moore@dartmouth.edu*

The rapid development of sequencing technologies makes thousands to millions of genetic attributes available for testing associations with various biological traits. Searching this enormous high-dimensional data space imposes a great computational challenge in genome-wide association studies. We introduce a network-based approach to supervise the search for three-locus models of disease susceptibility. Such statistical epistasis networks (SEN) are built using strong pairwise epistatic interactions and provide a global interaction map to search for higher-order interactions by prioritizing genetic attributes clustered together in the networks. Applying this approach to a population-based bladder cancer dataset, we found a high susceptibility three-way model of genetic variations in DNA repair and immune regulation pathways, which holds great potential for studying the etiology of bladder cancer with further biological validations. We demonstrate that our SEN-supervised search is able to find a small subset of three-locus models with significantly high associations at a substantially reduced computational cost.

Keywords: Epistasis; High-order genetic interactions; GWAS; Statistical epistasis networks; MDR.

1. Introduction

The goal of genome-wide association studies (GWAS) is to identify and characterize susceptibility genes that can help diagnose, treat, and prevent common human diseases.[1–3] However, most existing association analyses employ main-effect-centered strategies that assume a simple genetic architecture and are thus only able to find very limited single-locus effects on disease risks.[4] The non-additive effect of gene-gene interactions, i.e. *epistasis*, has been recognized playing an important role explaining the complex relationship between the genetic and phenotypic variations.[5–7] Thus, identifying and characterizing genetic interactions across multiple loci have become the focus of current association studies.[8–10] However, this imposes a great computational challenge in high-dimensional data analyses. Specifically, for a genetics dataset consisting of n loci, the computational complexity of enumerating all possible two-locus combinations is $O(n^2)$, and it increases exponentially with the order of combinations considered. Given the sizes of current genome-wide data ($n \sim 10^6$) and the next-generation whole-genome sequencing[11] data ($n \sim 10^9$), it requires 3×10^4 to 3×10^{13} years to enumerate and evaluate all three-locus models, using a 1000-node computer cluster where each node is assumed to be able to process 1000 models per second. Therefore, new data-mining technologies with advanced and efficient pre-screening and attribute-selection strategies are needed in large-scale genetic association studies.[12–15]

In this article, we propose a network-based model-prioritization approach that is able to identify high-order association models at a significantly reduced computational cost than exhaustive enumerations. The networks were built by including strong pairwise epistatic interactions as edges and their two end genetic attributes as vertices, as in the framework of statistical epistasis networks (SEN) previously developed by Hu et al.[16] Following the hypothesis that strong pairwise interactions may indicate the existence of higher-order interactions, we propose to i) quantify all pairwise epistatic interactions in a given genetics dataset; ii) construct pairwise statistical epistasis networks; iii) identify attributes that are clustered together by traversing the networks; iv) evaluate the clustered attributes for higher-order interactions. This distinguishes our approach the most from many existing attribute-selection strategies and advances the detection of higher-order interactions since hypothetically it is much less likely for a higher-order interaction to exist without showing any lower-order interactions than without showing any main effects.[17,18]

In the present study, we consider searching for three-locus interaction models and use the multifactor dimensionality reduction (MDR) algorithm and software to evaluate the associations of the models found by our SEN-supervised search. MDR is a data-mining strategy for detecting and characterizing gene-gene interactions associated with a discrete disease status.[19–22] It pools multi-locus genotypes from multiple single-nucleotide polymorphisms (SNPs) into high-risk and low-risk groups. Specifically, a multi-locus genotype combination is considered high-risk if it has subjects with a ratio of cases to controls higher than a given threshold; otherwise it is considered low-risk. The clustering of all multi-locus genotype combinations into high-risk and low-risk is then evaluated for its ability to classify and predict disease status through cross-validations. Population-based data are partitioned into a training set and a testing set. The attribute combination with the highest training accuracy is chosen as the best model and is subsequently evaluated using the testing set. The article by Moore et al[22] provides a good overview of the development of MDR. MDR is model-free, i.e. no particular genetic models are assumed, and non-parametric, i.e. no parameters are estimated, and is thus an ideal independent classifier to evaluate our SEN-supervised model search.

We previously identified a pairwise interaction network by applying SEN to a large population-based bladder cancer dataset.[16] Such a network showed significant topological properties compared to the null networks built from permuted data. We believe that this large connected structure captures the complex genetic architecture of bladder cancer and is a promising guide-map for searching higher-order combinations of attributes that may jointly modify the disease outcome. Here, we use this bladder cancer pairwise interaction network to supervise the search for high-association three-locus models using a fast network traversing algorithm that identifies trios clustered together.

2. Methods

2.1. *Bladder cancer dataset*

The dataset used in this study includes 1422 SNPs from about 500 cancer susceptibility genes for 491 bladder cancer cases and 791 healthy controls.[23,24] The bladder cancer cases were collected among New Hampshire residents of ages 25 to 74 years, diagnosed from July 1, 1994 to June 30, 2001 and identified in the State Cancer Registry. Controls less than 65 years of age were selected using population lists obtained from the New Hampshire Department of Transportation, while controls aged 65 and older were chosen from data files provided by the Centers for Medicare & Medicaid Services (CMS) of New Hampshire. Most ($> 95\%$) of the subjects were of Caucasian origin. Informed consent was obtained from each participant and all procedures and study materials were approved by the Committee for the Protection of Human Subjects at Dartmouth College. DNA was isolated from peripheral circulating blood lymphocyte specimens using Qiagen genomic DNA extraction kits (QIAGEN Inc., Valencia, CA). Genotyping was performed on all DNA samples of sufficient concentration, using the GoldenGate Assay system by Illumina's Custom Genetic Analysis service (Illumina, Inc., San Diego, CA). Out of the submitted samples, 99.5% were successfully genotyped, and samples repeated on multiple plates yielded the same call for 99.9% of the SNPs.

2.2. *Statistical epistasis networks (SEN)*

We have previously developed a network approach to inferring statistical epistasis of bladder cancer.[16] First, entropy-based information-theoretic measures were used to quantify pairwise interactions[22,25–28] for all two-locus models in the bladder cancer dataset. Specifically, for two genetic attributes G_1, G_2, and the phenotypic status C, *mutual information* $I(G_1;C)$ and $I(G_2;C)$ measure the shared information, or dependency, between individual genotypes and the phenotype, i.e. the main effects. In addition, by joining G_1 and G_2 together, $I(G_1,G_2;C)$ measures how much of the phenotypic status that combining G_1 and G_2 can explain. The epistatic interaction strength between G_1 and G_2 can then be defined using *information gain* $IG(G_1;G_2;C) = I(G_1,G_2;C) - I(G_1;C) - I(G_2;C)$. As such, $IG(G_1;G_2;C)$ is the gained mutual information about C from considering genetic attributes G_1 and G_2 together, i.e. the synergy between G_1 and G_2 on the phenotype C. Moreover, normalizing the main effect $I(G_1;C)$ and the interaction effect $IG(G_1;G_2;C)$, by dividing them by the entropy of the phenotype $H(C)$, provides the association of a single attribute or a pairwise interaction with the phenotype C, i.e. the percentage of the phenotypic status that a genotype can explain.

Second, we ranked all possible pairwise interactions between SNPs according to their relative strength and subsequently built a series of statistical epistasis networks by incrementally adding edges if their corresponding pairwise interaction strength was stronger than a given cutoff value. Topological properties were analyzed for the network at each cutoff value including the size of the network (the number of its vertices and the number of its edges), the connectivity of the network (the size of its largest connected component), and its vertex degree distribution. Permutation testing was used to generate a null distribution of those topological properties by building permuted-data networks through the same construction process and

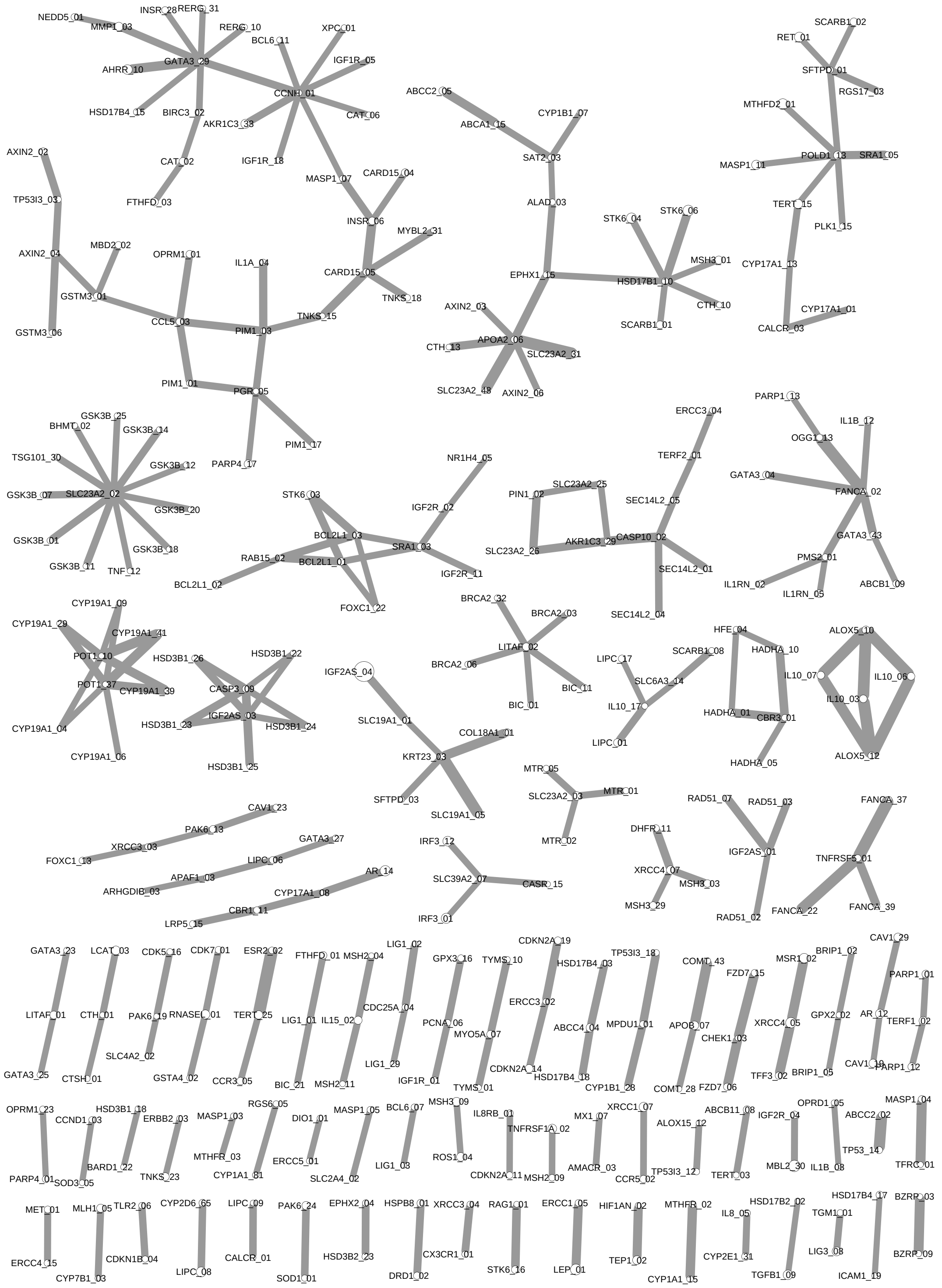


Fig. 1. The derived statistical epistasis network of bladder cancer. The network includes 319 SNPs (vertices) and 255 pairwise interactions (edges). The size of a vertex represents the strength of the main effect of its corresponding SNP, with the disease association ranging from 0.001% to 1.614%. The width of an edge indicates the strength of its corresponding interaction, with the disease association ranging from 1.354% to 2.052%.

using the same cutoffs.

Then, a threshold of the pairwise interaction strength was determined by finding the cutoff when the topological properties of the real-data network differentiated the most from the null distribution.[16] Such a systematically derived and most significant epistasis network of bladder cancer is shown in Fig. 1. This network provided a global map of strong pairwise epistatic interactions associated with bladder cancer. It was able to show not only the neighborhood structure of each attribute, but also the topology of a set of clustered attributes. Thus it serves as a very promising tool to identify higher-order genetic models.

2.3. ***SEN-supervised search for three-locus genetic models***

SEN is essentially an attribute-prioritization approach. However, different from many existing main-effect-centered pruning methods, our network strategy prioritizes attribute pairs that show strong or significant interactions. In addition, organizing these strong interacting pairs in the network format provides a landscape of interaction structures. We hypothesize that the sets of attributes that are clustered together in the bladder cancer network may better explain the case-control outcome than the non-clustered sets. Therefore, we propose to use SEN to supervise the search for multi-locus association models. As the first attempt, in this study, we consider the search for three-locus models and use MDR to assess the associations of three-locus models.

The clustering of vertices, or attributes, in a network is determined based on their pairwise distances. In Graph Theory,[29] the distance $d(v_1, v_2)$ of a pair of vertices v_1 and v_2 is defined as the minimal number of edges for one vertex to reach the other. Two vertices v_1 and v_2 are *neighbors* if $d(v_1, v_2) = 1$. Given three vertices v_1, v_2, and v_3, we define their trio distance $d_{\text{trio}}(v_1, v_2, v_3)$ as the sum of all pairwise distances, i.e. $d_{\text{trio}}(v_1, v_2, v_3) = d(v_1, v_2) + d(v_1, v_3) + d(v_2, v_3)$. Therefore, for trios with $d_{\text{trio}} = 3$, any two of them are directly joined by an edge, and if a trio has $d_{\text{trio}} = 4$, one vertex is directly connected to the other two but the other two are not joined by an edge. We define that a trio of attributes are *clustered* in a network if their $d_{\text{trio}} \leq 4$; otherwise we say that they are not clustered together.

All three-locus models of clustered trios can be identified by traversing the SEN, represented as a graph G with $|V|$ vertices and $|E|$ edges, using the following algorithm. It reads G and outputs a list of trios of vertices that are connected together. The algorithm has a computational complexity $O(|V| \times k^2)$, where k is the maximum number of neighbors of a vertex in G:

```
$vertices = G$.getVertices();
for each $v$ in $vertices$ do
    $neighbors = v$.getNeighbors();
    for each $u_1$ in $neighbors$
        for each $u_2$ in $neighbors$ do
            output $\{u_1, v, u_2\}$;
```

Note that in our bladder cancer epistasis network (Fig. 1) $k = 11 \ll |V|$, so the complexity of the above algorithm $O(|V| \times k^2) \approx O(|V|)$. Thus the SEN-supervised search significantly reduces the computational complexity compared to enumerating all three-locus combinations.

3. Results

We first applied a χ^2 test of independence to identify SNPs with significant main effects. For all 1422 SNPs from the entire dataset, using a Bonferroni-corrected significance level of $\alpha = 0.05$, we found only one significant main-effect attribute *IGF2AS_04* ($p = 1.052 \times 10^{-6}$). This SNP had one interacting neighbor *SLC19A1_01* captured in our SEN (Fig. 1), and this pairwise interaction was previously reported.[30] Thus we removed *IGF2AS_04* from our interaction analysis to avoid its dominance effect when combined with other attributes.

Next, for the other 318 SNPs identified in the bladder cancer network, we ran MDR exhaustively on all 1-way, 2-way ($\binom{318}{2}= 50,403$ pairs), and 3-way ($\binom{318}{3}= 5,309,116$ trios) combinations. We analyzed the correlation between MDR accuracies and SNP neighborhood structures in the network, in order to see whether clustered SNPs in the network have better disease status prediction accuracies than non-clustered ones.

3.1. *MDR accuracy comparison of clustered and non-clustered SNP trios*

We categorized all 5,309,116 trios according to their trio distances and show the MDR accuracies in each distance category (Fig. 2). We observe that, since there are no triangles in the network, the minimal trio distance is 4. In addition, trios of distances greater than 32 are not connected in the network, i.e. at least two out of the three vertices do not have a path connecting them. The clustered trios of distance 4 have significantly higher training and testing accuracies than the trios in all other distance categories, while those other distance categories do not statistically distinguish among themselves. Moreover, the clustered trios have better consistencies between training and testing accuracies (Fig. 2B inset).

We then binned all $d_{\text{trio}} > 4$ three-locus models together as non-clustered trios, and com-

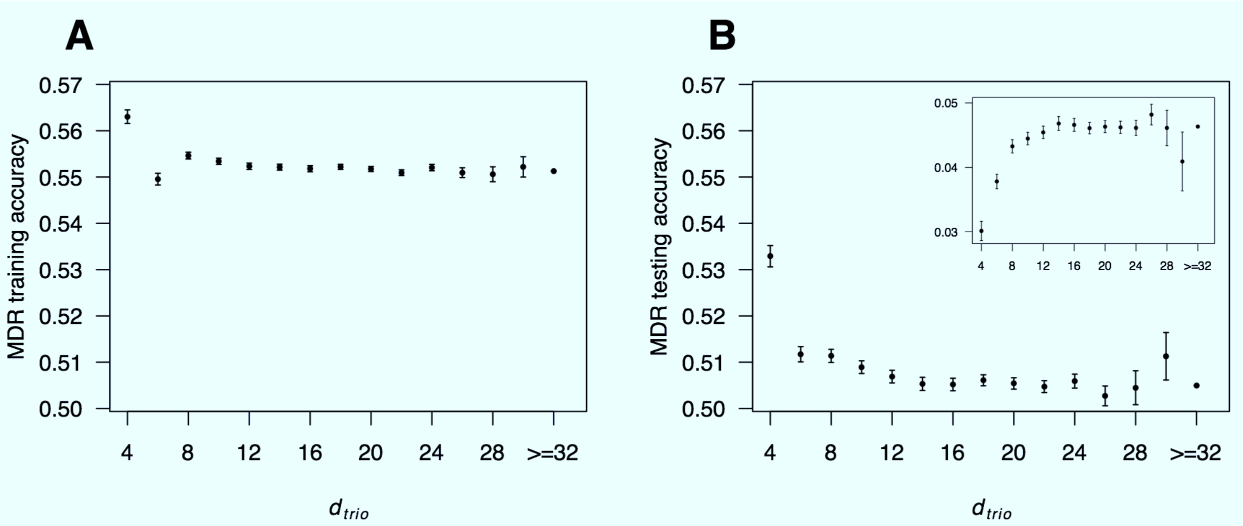


Fig. 2. The 3-way MDR **A**) training accuracy and **B**) testing accuracy relative to the trio distance. Points are mean values and bars show the 95% confidence intervals. The inset depicts Δ = training accuracy − testing accuracy, which indicates the level of over-fitting. A lower value of Δ means a better prediction consistency for training and testing data.

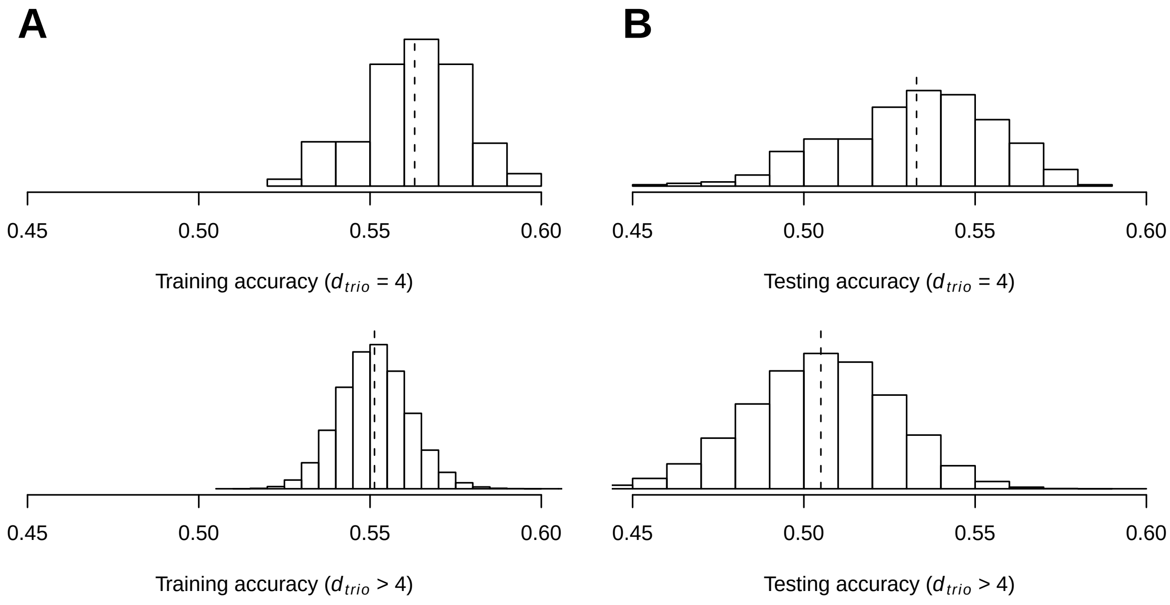


Fig. 3. Distributions of 3-way MDR **A**) training and **B**) testing accuracies for clustered ($d_{\text{trio}} = 4$) and non-clustered ($d_{\text{trio}} > 4$) trios. The mean of each distribution is shown using a vertical dashed line. There are 391 clustered trios and $5,309,116 - 391 = 5,308,725$ non-clustered trios.

pared their distributions of MDR training and testing accuracies to those of the clustered trios (Fig. 3). As seen from the figure, clustered trios have both better training and testing accuracies compared to non-clustered trios. Therefore, using the pairwise SEN structure was able to identify a good subset of three-locus combinations that improved the phenotypic status prediction accuracy.

We also performed a correlation analysis on the MDR accuracies at different combination orders. Table 1 shows that, in general, three-way accuracies had stronger correlations with two-way accuracies than those with one-way accuracies. Compared to non-clustered trios, the three-way accuracies of clustered trios were less correlated with one-way accuracies. That is, three-locus models of clustered trios were less biased towards high main-effects attributes. When correlating two-way with three-way accuracies, compared to non-clustered trios, clustered trios had a lower dependency on training data but a higher dependency on testing data.

3.2. *SEN-supervised MDR three-locus models*

As shown previously, SEN-supervised search yielded a small subset of three-locus combinations (391 out of 5,309,116) based on their clustering structure in the network, and this small subset had significantly better three-way MDR accuracies compared to the others. In this section, we examined the results of these SEN-supervised MDR models, and tested whether the observations from such a model-selection process were statistically significant.

For these 391 SEN-filtered trios, their best and average MDR accuracies are reported in Table 2. We performed two sets of significance tests to assess the p-values for each observation. First, we randomly resampled 391 trios out of the total 5,309,116 and repeated it 1000 times. Second, on the 318 vertices identified in the network, we permuted their neighborhood

Table 1. Spearman's rank correlation of MDR accuracies at different model orders

	1-way vs. 3-way	**2-way vs. 3-way**
	Training balanced accuracy	
Clustered trios	$\rho = 0.1863$ ($p = 1.27 \times 10^{-10}$)	$\rho = 0.4319$ ($p < 2.2 \times 10^{-16}$)
Non-clustered trios	$\rho = 0.2934$ ($p < 2.2 \times 10^{-16}$)	$\rho = 0.5897$ ($p < 2.2 \times 10^{-16}$)
	Testing balanced accuracy	
Clustered trios	$\rho = 0.1060$ ($p = 2.77 \times 10^{-4}$)	$\rho = 0.4027$ ($p < 2.2 \times 10^{-16}$)
Non-clustered trios	$\rho = 0.1946$ ($p < 2.2 \times 10^{-16}$)	$\rho = 0.3795$ ($p < 2.2 \times 10^{-16}$)

Table 2. MDR results of the clustered trios and their levels of statistical significance

	Observed-value	**Significance**	
		random-resample	edge-swap
Best training accuracy	0.5992	$p = 0.005$	$p = 0.002$
Best testing accuracy	0.5873	$p = 0.002$	$p < 0.001$
Average training accuracy	0.5630	$p < 0.001$	$p < 0.001$
Average testing accuracy	0.5329	$p < 0.001$	$p < 0.001$

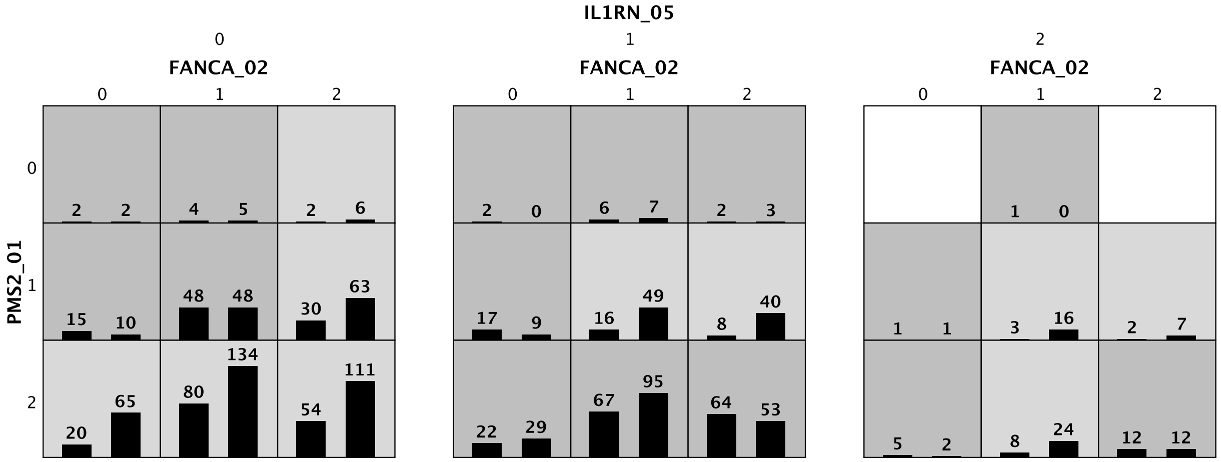


Fig. 4. Summary of the best MDR model using SEN-supervised search. A three-locus model has 27 multifactorial cells, each of which is filled with the distribution of cases (left bars) and controls (right bars) for the corresponding genotypes. A cell is left blank if there are no samples falling into its genotype. Each non-empty cell is labeled either "high-risk" (dark grey) or "low-risk" (light grey) based on its case-control ratio.

structures by swapping edges. For each edge swapping, two edges, e.g. $e_1 = \{v_{11}, v_{12}\}$ and $e_2 = \{v_{21}, v_{22}\}$, were picked randomly, and then their end vertices were swapped to form two new edges $e'_1 = \{v_{11}, v_{22}\}$ and $e'_2 = \{v_{21}, v_{12}\}$. This was a standard network randomization procedure where the total number of neighbors for each vertex was preserved but its interact-

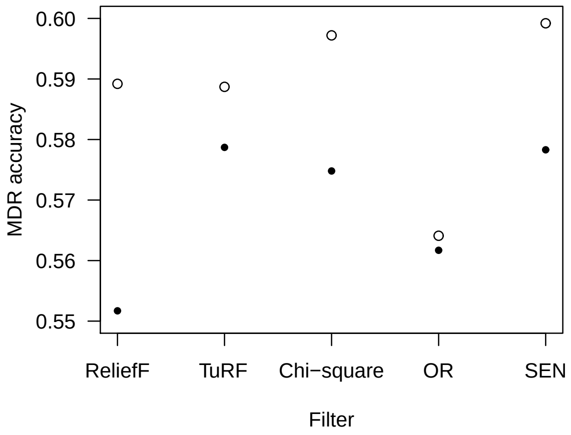


Fig. 5. Results of the best three-locus MDR models using five different attribute-selection or model-prioritization techniques. Circles represent training balanced accuracies and solid points are testing balanced accuracies.

ing partners were randomized. For each permutation, we performed edge swapping $10 \times |E|$ times, where $|E|$ is the total number of edges in the network (Fig. 1). Such a permutation process provided null networks with randomized pairwise interactions. Again, we generated 1000 permuted networks and used them to identify the clustered-trio subsets. Then MDR analyses were applied to both sets of permuted data and the assessed significances of the real observations are shown in Table 2. As we can see, all observations from the subset found by SEN-supervised search were statistically significant.

The best three-locus MDR model using SEN-supervised search was *FANCA_02, PMS2_01*, and *IL1RN_05*, with a training balanced accuracy 0.5992 and a testing balanced accuracy 0.5783 ($p = 1 \times 10^{-5}$ using a standard permutation test). This model included two DNA repair genes and one immune regulation gene. Fig. 4 summarizes the MDR analysis for the best model. Out of all 27 possible genotype combinations, 25 had observed samples, 15 genotypes were predicted as high-risks (dark-grey cells), and 10 genotypes were predicted as low-risks (light-grey cells).

3.3. *Comparing SEN-supervised search to other common MDR filters*

Due to the exhaustive enumeration nature of MDR, attribute-selection is usually used for large genome-wide data. We implemented four most commonly used filters, ReliefF,[31] TuRF,[32] Chi-square, and Odds Ratio (OR), on the bladder cancer data (1422 SNPs), and compared the best models they found to our best model using SEN-supervised search (Fig. 5). For each of the four other filters, we chose its top 15 most important attributes and ran MDR on all three-locus combinations ($\binom{15}{3}$= 455) of them. This also provided a comparable number of models for MDR to evaluate since SEN-supervised search yielded 391 three-locus combinations. As seen in the figure, our SEN-supervised search found the best three-locus model compared to all the other common attribute-selection strategies.

4. Discussion

Epistasis has been recognized playing an important role in understanding the mapping between genetic and phenotypic variations.[8–10] Detecting and characterizing epistasis is a very challenging data-mining task due to the fact that the epistatic interactions could involve multiple genetic attributes from a pair to a large set, and this undetermined order of interactions imposes enormous computational complexities for enumerating all possible combinations of genetic attributes for varying orders in genome-wide data.[15] Various pre-screening techniques have been proposed to filter potentially important attributes for further higher-order combination analyses. However, most of them adopt main-effect-centered strategies and may overlook attributes that are important in interactions but only show weak main effects.[17]

In this article, we proposed a network-guided approach to searching three-locus genetic models for association studies. The network was built by including strong pairwise epistatic interactions, and we were able to show that trios clustered together in this network have higher associations than those non-clustered ones. Traversing the pairwise statistical epistasis networks (SEN) to search clustered three-locus models significantly reduces the computational complexity of enumerating all possible three-locus combinations. Thus our SEN-supervised model search can serve a very promising prioritization method and can be combined with many existing association-mining techniques, such as MDR used in this study.

We had previously developed a network approach to characterizing statistical epistasis interactions in genetic association studies.[16] In this framework, all pairwise interactions in a genetic dataset were quantified using information gain, an information-theoretic measure based on Shannon entropy.[33] Then networks were built by including pairs of attributes, as edges and two end vertices, if their pairwise interaction strengths were greater than a theoretically-derived threshold. This threshold was determined systematically by analyzing network topological properties and comparing them to null networks built using permuted data through the same construction process. This SEN approach advanced many existing genetic association methods by focusing on interactions rather than individual genetic factors. Moreover, by organizing interactions in the form of networks, SEN provided a global connection map and suggested clustering of multiple attributes that might have joint effects on the phenotype.

The present study explored the clustering structure captured in our previous SEN application to a bladder cancer dataset (Fig. 1). Using a fast network-traversing algorithm, the three-locus models of clustered trios were identified and further evaluated using MDR. These models were shown having both significantly higher training and testing MDR accuracies than the three-locus models of non-clustered trios (Fig. 2 and Fig. 3). Moreover, the clustered models had less over-fitting (Fig. 2B inset). These results show that the SEN-supervised search was able to identify a small subset of three-locus models with significantly high associations at a very moderate computational cost. Note that even if the computational complexity of building a pairwise interaction network ($O(|V|^2)$) is considered together with the SEN-supervised search ($O(|V| \times k^2) \approx O(|V|)$), where $|V|$ is the total number of attributes and k is the maximum number of neighbors of an attribute in the network, the computational cost is still far less than enumerating all possible three-locus combinations ($O(|V|^3)$). This reduction of computational complexity is even more encouraging in the era of genome-wide and whole-genome studies where thousands to millions of genetic attributes are considered.

The best three-locus MDR model identified using the SEN-supervised search includes *FANCA_02* (rs2239359), *PMS2_01* (rs3735295), and *IL1RN_05* (rs419598). All three SNPs had very limited main effects with one-way MDR testing accuracies 0.4929, 0.5110, and 0.5276, respectively. The falcon anemia complementation group A (FANCA) gene produces DNA repair protein that may operate in a post replication repair or a cell cycle checkpoint function. Postmeiotic segregation increased 2 (PMS2) is a component of the post-replicative DNA mismatch repair system. Interleukin 1 receptor antagonist (IL1RN) encodes the protein that inhibits the activities of interleukin 1 alpha (IL1A) and interleukin 1 beta (IL1B), and modulates a variety of interleukin 1 related immune and inflammatory responses. The three genes have moderate biological relationships,[34] all have been found associated with various cancers, and both DNA repair and immune regulation are considered major biological processes involved in bladder carcinogenesis.[35–37] However, the interaction effect among the three genes associated with bladder cancer has never been reported previously. One could speculate, nevertheless, that defects in the protective cell cycle checkpoint and DNA repair functions could lead to attempts to replicate damaged DNA. Immune surveillance would be the remaining protective mechanism to eliminate potential tumor cells. Thus, this trio of genetic variations could increase the probability of tumor cell expansion. We expect that with further biological validations, our findings could help explain the etiology and the complex genetic architecture of bladder cancer.

With the fast development of sequencing technologies, more and more large-scale biomedical data are becoming available. Although this presents exciting opportunities for genetic association studies to explain many common human diseases, mining these high-dimensional data to identify important genetic factors with non-linear interaction effects is a daunting endeavor. In this article, we proposed a network-guided search approach that is able to efficiently identify high-association three-locus genetic models. Our approach prioritizes genetic attributes that have strong pairwise interaction effects. This differentiates our method from most existing pre-screening strategies that focus on individual attributes with significant main effects. The effectiveness of our approach was validated using MDR. In future research, we expect to extend our SEN-supervised approach to the search for higher-order models and to expand its applications to more data-mining and classification techniques.

Acknowledgements

This work was supported by the National Institute of Health (NIH) grants R01-LM009012, R01-LM010098, and R01-AI59694 to JHM, and P42-ES007373 and R01-CA5749367 to MRK.

References

1. J. N. Hirschhorn and M. J. Daly, *Nature Review Genetics* **6**, 95 (2005).
2. W. Y. S. Wang, B. J. Barratt, D. G. Clayton and J. A. Todd, *Nature Review Genetics* **6**, 109 (2005).
3. J. Hardy and A. Singleton, *New England Journal of Medicine* **360**, 1759 (2009).
4. T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff and et al., *Nature* **461**, 747 (2009).

5. J. H. Moore, *Human Heredity* **56**, 73 (2003).
6. O. Carlborg and C. S. Haley, *Nature Review Genetics* **5**, 618 (2004).
7. J. H. Moore and S. M. Williams, *BioEssays* **27**, 637 (2005).
8. H. J. Cordell, *Human Molecular Genetics* **11**, 2463 (2002).
9. H. J. Cordell, *Nature Review Genetics* **10**, 392 (2009).
10. J. H. Moore and S. M. Williams, *The American Journal of Human Genetics* **85**, 309 (2009).
11. J. Shendure and H. Ji, *Nature Biotechnology* **26**, 1135 (2008).
12. J. H. Moore and M. D. Ritchie, *Journal of the Amarican Medical Association* **291**, 1642 (2004).
13. J. H. Moore and B. C. White, *Genetic Programming Theory and Practice IV* , 969 (2005).
14. R. Nunkesser, T. Bernholt, H. Schwender, K. Ickstadt and I. Wegener, *Bioinformatics* **23**, 3280 (2007).
15. J. H. Moore, F. W. Asselbergs and S. M. Williams, *Bioinformatics* **26**, 445 (2010).
16. T. Hu, N. A. Sinnott-Armstrong, J. W. Kiralis, A. S. Andrew, M. R. Karagas and J. H. Moore, *BMC Bioinformatics* **12**, p. 364 (2011).
17. R. Culverhouse, B. K. Suarez, J. Lin and T. Reich, *American Journal of Human Genetics* **70**, 461 (2002).
18. W. Wongseree, A. Assawamakin, T. Piroonratana, S. Sinsomros, C. Limwongse and N. Chaiyaratana, *BMC Bioinformatics* **10**, p. 294 (2009).
19. M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl and J. H. Moore, *The American Journal of Human Genetics* **69**, 138 (2001).
20. L. W. Hahn, M. D. Ritchie and J. H. Moore, *Bioinformatics* **19**, 376 (2003).
21. M. D. Ritchie, L. W. Hahn and J. H. Moore, *Genetic Epidemiology* **24**, 150 (2003).
22. J. H. Moore, J. C. Gilbert, C.-T. Tsai, F.-T. Chiang, T. Holden, N. Barney and B. C. White, *Journal of Theoretical Biology* **241**, 252 (2006).
23. M. R. Karagas, T. D. Tosteson, J. Blum, J. S. Morris, J. A. Baron and B. Klaue, *Environmental Health Perspectives* **106**, 1047 (1998).
24. A. S. Andrew, H. H. Nelson, K. T. Kelsey, J. H. Moore, A. C. Meng, D. P. Casella, T. D. Tosteson, A. R. Schned and M. R. Karagas, *Carcinogenesis* **27**, 1030 (2006).
25. A. Jakulin and I. Bratko, *Lecture Notes in Artificial Intelligence* **2838**, 229 (2003).
26. D. Anastassiou, *Molecular Systems Biology* **3**, p. 83 (2007).
27. J. H. Moore, N. Barney, C.-T. Tsai, F.-T. Chiang, J. Gui and B. C. White, *Human Heredity* **63**, 120 (2007).
28. B. A. McKinney, J. E. Crowe, J. Guo and D. Tian, *PLoS Genetics* **5**, p. e1000432 (2009).
29. D. B. West, *Introduction to Graph Theory: Second edition* (Prentice Hall, 2001).
30. A. S. Andrew, J. Gui, A. C. Sanderson, R. A. Mason, E. V. Morlock, A. R. Schned, K. T. Kelsey, C. J. Marsit, J. H. Moore and M. R. Karagas, *Human Genetics* **125**, 527 (2009).
31. I. Kononenko, *Lecture Notes in Computer Science* **784**, 171 (1994).
32. J. H. Moore and B. C. White, *Lecture Notes in Computer Sceicne* **4447**, 166 (2007).
33. T. M. Cover and J. A. Thomas, *Elements of Information Theory: Second Edition* (Wiley, 2006).
34. A. K. Wong, C. Y. Park, C. S. Greene, L. A. Bongo, Y. Guan and O. G. Troyanskaya, *Nucleic Acids Research* **40**, W484 (2012).
35. E. M. El-Omar, M. Carrington, W.-H. Chow, K. E. L. McCol, J. H. Bream, H. A. Young, J. Herrera, J. Lissowska, C.-C. Yuan, N. Rothman, G. Lanyon, M. Martin, J. F. Fraumeni Jr and C. S. Rabkin, *Nature* **404**, 398 (2000).
36. M. C. Southey, M. A. Jenkins, L. Mead, J. Whitty, M. Trivett and et al., *Journal Of Clinical Oncology* **23**, 6524 (2005).
37. S. Michiels, A. Laplanche, T. Boulet, P. Dessen, B. Guillonneau, A. Mejean, F. Desgrandchamps, M. Lathrop, A. Sarasin and S. Benhamou, *Carcinogenesis* **30**, 763 (2009).

EVALUATION OF LINEAR CLASSIFIERS ON ARTICLES CONTAINING PHARMACOKINETIC EVIDENCE OF DRUG-DRUG INTERACTIONS

A. KOLCHINSKY

School of Informatics and Computing, Indiana University
Bloomington, IN, USA
E-mail: akolchin@indiana.edu

A. LOURENÇO

Institute for Biotechnology & Bioengineering, Centre of Biological Engineering, University of Minho
Braga, Portugal
E-mail:analia@deb.uminho.pt

L. LI

Department of Medical and Molecular Genetics, Indiana Univeristy School of Medicine
Indianapolis, IN, USA
E-mail: lali@iupui.edu

L. M. ROCHA*

School of Informatics and Computing, Indiana University
Bloomington, IN, USA
**E-mail: rocha@indiana.edu*

Background. Drug-drug interaction (DDI) is a major cause of morbidity and mortality. DDI research includes the study of different aspects of drug interactions, from *in vitro* pharmacology, which deals with drug interaction mechanisms, to pharmaco-epidemiology, which investigates the effects of DDI on drug efficacy and adverse drug reactions. Biomedical literature mining can aid both kinds of approaches by extracting relevant DDI signals from either the published literature or large clinical databases. However, though drug interaction is an ideal area for translational research, the inclusion of literature mining methodologies in DDI workflows is still very preliminary. One area that can benefit from literature mining is the automatic identification of a large number of potential DDIs, whose pharmacological mechanisms and clinical significance can then be studied via *in vitro* pharmacology and *in populo* pharmaco-epidemiology.

Experiments. We implemented a set of classifiers for identifying published articles relevant to experimental pharmacokinetic DDI evidence. These documents are important for identifying causal mechanisms behind putative drug-drug interactions, an important step in the extraction of large numbers of potential DDIs. We evaluate performance of several linear classifiers on PubMed abstracts, under different feature transformation and dimensionality reduction methods. In addition, we investigate the performance benefits of including various publicly-available named entity recognition features, as well as a set of internally-developed pharmacokinetic dictionaries.

Results. We found that several classifiers performed well in distinguishing relevant and irrelevant abstracts. We found that the combination of unigram and bigram textual features gave better performance than unigram features alone, and also that normalization transforms that adjusted for feature frequency and document length improved classification. For some classifiers, such as linear discriminant analysis (LDA), proper dimensionality reduction had a large impact on performance. Finally, the inclusion of NER features and dictionaries was found not to help classification.

1. Introduction

Drug-drug interaction (DDI) has been implicated in nearly 3% of all hospital admissions[1] and 4.8% of admissions among the elderly;[2] it is also a common form of medical error, representing 3% to 5% of all inpatient medication errors.[3] With increasing rates of polypharmacy, which refers to the use of multiple medications or more medications than are clinically indicated,[4] the incidence of DDI will likely increase in the coming years.

DDI research includes the study of different aspects of drug interactions. *In vitro* pharmacology experiments use intact cells (e.g. hepatocytes), microsomal protein fractions, or recombinant systems to investigate drug interaction mechanisms. Pharmaco-epidemiology (*in populo*) uses a population based approach and large electronic medical record databases to investigate the contribution of a DDI to drug efficacy and adverse drug reactions.

Biomedical literature mining (BLM) can be used to detect novel DDI signals from either the published literature or large clinical databases.[5] BLM is becoming an important biomedical informatics methodology for large scale information extraction from repositories of textual documents, as well as for integrating information available in various domain-specific databases and ontologies, ultimately leading to knowledge discovery.[6–8] It has seen applications in research areas that range from protein-protein interaction,[9,10] protein structure,[11] genomic locations associated with cancer,[12] drug targets,[13] and many others. BLM holds the promise of tapping into the biomedical collective knowledge and uncovering relationships buried in the literature and databases, especially those relationships present in global information but unreported in individual experiments.[14]

Although pharmaco-epidemiology and BLM approaches are complementary, they are usually conducted independently. DDI is thus an exemplary case of translational research that can benefit from interdisciplinary collaboration. In particular, automated literature mining methods allow for the extraction of a large number of potential DDIs whose pharmacological mechanisms and clinical significance can be studied in conjunction with *in vitro* pharmacology and *in populo* pharmaco-epidemiology.

Though BLM has previously been used for DDI information extraction,[15,16] much remains to be done before it can integrated into translational workflows. One gap is in the extraction of DDI information from a pharmacokinetics perspective, since existing methods do not explicitly capture pharmacokinetics parameters and do not consider knowledge from *in vitro* and *in vivo* DDI experimental designs, especially the selection of enzyme-specific probe substrates and inhibitors. For instance, important pharmacokinetic parameters such as Ki, IC50, and AUCR have not been included in existing text mining approaches to DDI. Yet this kind of pharmacokinetic information may be particularly relevant when seeking evidence of causal mechanisms behind DDIs, and as a complement to DDI text mining of patient records, where reporting biases and confounds often give rise to non-causal correlations.[17]

We have previously showed that BLM can be used for automatic extraction of numerical pharmacokinetics (PK) parameters from the literature.[18] However, that work was not oriented specifically toward extraction of DDI information. In order to perform DDI information extraction from a pharmacokinetics perspective, we first need to be able to identify the relevant documents that contain such information. Here, we evaluate the performance of text classification methods on documents that may contain pharmacology experiments in which evidence for DDIs is reported. Our goal is to develop and evaluate automated methods of identifying DDIs backed by reported pharmacokinetic evidence, which we believe is an essential first step towards the integration of literature mining methods into translational DDI workflows. A collaboration between Rocha's lab, working on BLM, and Li's lab, working on *in vitro* pharmacokinetics, was developed in order to pursue this goal.

In this paper, we report on the performance of a set of classifiers on a manually-annotated corpus produced by Li's lab. We consider a wide range of linear classifiers, among them logistic regression, support vector machines (SVM), binomial Naive Bayes, linear discriminant analysis, and a modification of our 'Variable Trigonometric Threshold' (VTT) classifier, which was previously found to perform well on protein-protein interaction text mining tasks.[14,19,20] In addition, we compare different feature transformation methods, including normalization techniques such as TFIDF and PCA-based dimensionality reduction. We also compare performance when using features generated by several Named Entity Recognition (NER) tools.

In the next section, we describe the corpus used in this study. Section 3 discusses the evaluated classifiers, while section 4 deals with dimensionality reduction and feature transforms. Section 5 covers our methods of cross-validation and performance evaluation. Section 6 provides classification performance results for textual features, while section 7 does so for the combination of textual and NER features. We conclude with a discussion in section 8.

2. Corpus

Li's lab selected 1213 PubMed pharmacokinetics-related abstracts for the training corpus. Documents were obtained by first searching PubMed using terms from an ontology previously developed for automatic extraction of numerical PK pharmacokinetics parameters.[18] The retrieved articles were manually classified into two groups: abstracts that explicitly mentioned evidence for the presence or absence of drug-drug interactions were labeled as DDI-relevant (602 abstracts), while the rest were labeled as DDI-irrelevant (611 abstracts). DDI-relevance was established if articles contained one of the four primary classes of pharmacokinetics studies: clinical PK studies, clinical pharmacogenetic studies, *in vivo* DDI studies, and *in vitro* drug interaction studies. The classification was initially done by three graduate students with M.S. degrees and one postdoctoral annotator. Any inter-annotator conflicts were further checked by a Pharm D. and an M.D. scientist with extensive pharmacological training. The corpus, as well as further details,[21] is available upon request.

We extracted textual features from the abstract title and abstract text, as well as several other PubMed fields. These included the author names, the journal title, the Medical Subject Heading (MeSH) terms, the 'registry number/EC number' (RN) field, and the 'secondary source' field (SI) (the latter two contain identification codes for relevant chemical and biological entities). For each PubMed entry, the content of the above fields was tokenized, processed by Porter stemming, and converted into textual features (unigrams and, in certain runs, bigrams). Strings of numbers were converted into '#', while short textual features (those with a length of less than 2 characters) and infrequent features (those that occurred in less than 2 documents) were omitted. Each MeSH term was treated as a single textual token. Finally, the occurrence of different features in different documents was recorded in binary occurrence matrices. We evaluated performance using unigram features only (the unigram runs), as well as using a combination of unigram and bigram features (the bigram runs).

3. Classifiers

Six different linear classifiers were implemented:

(1) VTT: a simplified, angle-domain version of our 'Variable Trigonometric Threshold' Classifier (VTT).[14,19,20] Given a binary document vector $\mathbf{x} = \langle x_1, \ldots, x_K \rangle$, with its features (i.e. dimensions) indexed by i, the VTT separating hyperplane is:

$$\sum_i \theta_i x_i - \lambda = 0$$

Here, λ is a threshold (bias) and θ_i is the 'angle' of feature i in class space:

$$\theta_i = \arctan \frac{p_i}{n_i} - \frac{\pi}{4}$$

where p_i is the proportion of positive-class documents in which feature i occurs, and n_i is the proportion of negative-class documents in which features i occurs. θ_i is positive when $p_i \geq n_i$ and negative otherwise. The threshold parameter λ is chosen via cross-validation. The full version of VTT, previously used in protein-protein interaction tasks, includes additional parameters to account for named entity occurrences and is used in section 7 below. VTT performs best on sparse data sets, in which most feature values x_i are set to 0; for this reason, we do not evaluate it on dense dimensionality-reduced datasets (see below).

(2) SVM: a linear Support Vector Machine (SVM) classifier (provided by the sklearn[22] library's interface to the LIBLINEAR package[23]) with a cross-validated regularization parameter.

(3) Logistic regression: a logistic regression classifier (also provided by sklearn's interface to LIBLINEAR) with a cross-validated regularization parameter.

(4) Naive Bayes: a binomial Naive Bayes classifier with a Beta-distributed prior for smoothing. The prior's concentration parameter was determined by cross-validation.

(5) LDA: a Linear Discriminant Analysis (LDA) classifier, where the data covariance matrix was shrunk toward a diagonal, equal-variance structured estimate. The shrinkage parameter was determined by cross-validation.

(6) dLDA: a 'diagonal' version of LDA, where only the diagonal entries of the covariance matrix are estimated and the off-diagonal entries are taken to be 0. A cross-validated parameter determines shrinkage toward a diagonal, equal-variance estimate. This classifier provides a more robust estimate of feature variances; it is equivalent to a Naive Bayes classifier for multivariate Gaussian features.[24]

4. Feature Transforms

For both unigram and bigram runs, the classifiers were applied to the following data matrices:

(1) No transform: the raw binary occurrence matrices, as described in section 2. For LDA, when the number of documents (N) was less than the number of dimensions (giving rise

to singular covariance matrices), the occurrence matrices were projected onto their first N principal components.

(2) IDF: occurrences of feature i were transformed into that feature's Inverse Document Frequency (IDF) value:

$$\mathrm{idf}\,(i) = \log \frac{N}{c_i + 1}$$

where c_i is the total number of occurrences of features i among all documents. This reduced the influence of common words on classification.

(3) TFIDF: the Term Frequency, Inverse Document Frequency (TFIDF) transform applies the above IDF transform, and then divides each document's feature values by the total number of that document's features. This attempts to minimize differences between documents of different sizes (i.e. with different numbers of features).

(4) Normalization: here the non-transformed, IDF, and TFIDF document matrices underwent a length-normalization transform, where each document vector was inversely scaled by its L2 norm. This normalization has been argued to be especially important for good SVM performance.[25]

(5) PCA-based dimensionality reduction: The above matrices were run through a Principal Component Analysis (PCA) dimensionality reduction step. Projections onto the first 100, 200, 400, 600, 800, and 1000 components were applied.

5. Performance evaluation

We evaluated the performance of the classifiers using three different measures: the commonly-used F1 score, the area under the interpolated precision/recall curve[26] (here called iAUC), and Matthews Correlation Coefficient[27] (MCC).

In this task, only one corpus was provided. Thus, we had to use it both for training classifiers and for measuring generalization performance on out-of-sample documents. We performed the following cross-validation procedure to estimate generalization performance:

(1) The documents of the entire corpus were partitioned into 4 folds (75%-25% splits). This was repeated 4 times, giving a total of 16 folds (we call these the *outer folds*).

(2) For each fold, classifiers were trained on 75% block of the corpus and tested on the 25% block of the corpus.

(3) The 16 sets of testing results were averaged to produce an estimate of generalization performance.

In addition, all of the classifiers mentioned in section 3 contain cross-validated parameters: for VTT, this is the bias parameter, while the other classifiers have regularization or smoothing parameters. In order to fully separate training from testing data and accurately estimate generalization performance, nested cross-validation was done within each of the 75% blocks of the above outer folds:

(1) The 75% block is itself partitioned into 4 folds (75%-25% splits of the 75% block). This is repeated 4 times, producing a total of 16 folds (we call these the *inner folds*)

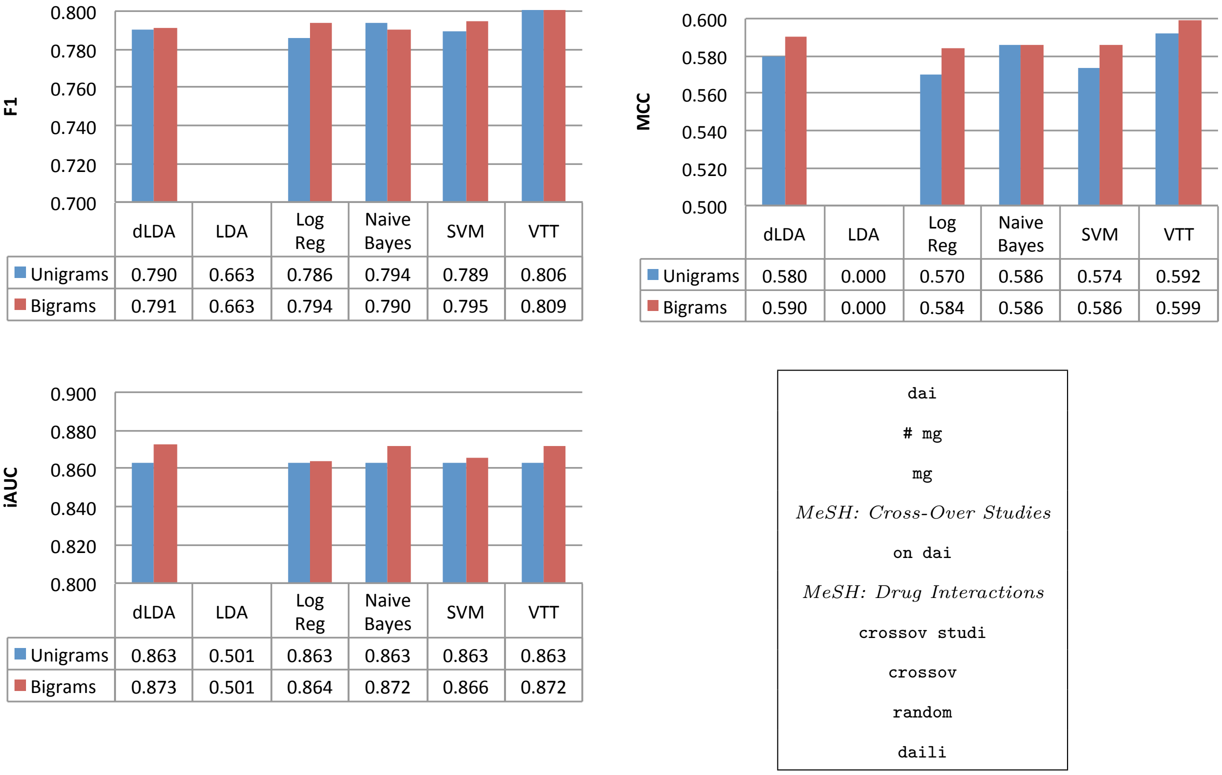
(2) For each searched value of the cross-validated parameter, a classifier is trained on each of the 16 inner folds' 75% block and tested on its 25% block.

(3) The value giving the best average performance (here, according to the MCC metric) is chosen as the cross-validated parameter value for this outer fold.

An outer fold's cross-validated parameter value is then used to train on the fold's 75% block and test on its 25% block.

6. Classification performance

6.1. *Overall performance*



F1	dLDA	LDA	Log Reg	Naive Bayes	SVM	VTT
🟦 Unigrams	0.790	0.663	0.786	0.794	0.789	0.806
🟥 Bigrams	0.791	0.663	0.794	0.790	0.795	0.809

MCC	dLDA	LDA	Log Reg	Naive Bayes	SVM	VTT
🟦 Unigrams	0.580	0.000	0.570	0.586	0.574	0.592
🟥 Bigrams	0.590	0.000	0.584	0.586	0.586	0.599

iAUC	dLDA	LDA	Log Reg	Naive Bayes	SVM	VTT
🟦 Unigrams	0.863	0.501	0.863	0.863	0.863	0.863
🟥 Bigrams	0.873	0.501	0.864	0.872	0.866	0.872

dai

# mg

mg

MeSH: Cross-Over Studies

on dai

MeSH: Drug Interactions

crossov studi

crossov

random

daili

Fig. 1. Classification performance using non-transformed features, for both unigram and bigram runs. Top left is the F1 measure, top right is the MCC measure, and lower left is the iAUC measure. LDA performed poorly and is below the charts' lower cutoff value. Lower right shows the top 10 features identified in a typical bigram fold, ranked according to the information gain criteria.

Figure 1 shows the performance of the classifiers in unigram runs (which included only unigram features) and bigram runs (which included both unigram and bigram features), without any feature transforms applied. In addition, it also shows the top 10 features identified in a typical bigram fold, ranked according to the information gain criteria.[28]

With the exception of LDA, all of the classifiers performed similarly on the task. VTT performed slightly better than the other classifiers according to the F1 and MCC measures. LDA's performance was dismal, suggesting that in such a high-dimensional setting there is not enough data to estimate the feature covariance matrix, even under covariance matrix shrinkage. This is supported by the fact that the dLDA (diagonal LDA) classifier, which estimates only the diagonal entries of the covariance matrix, performed well on the task.

The difference between unigram and bigram runs was not major, but bigram performance showed a consistent small improvement, indicating that the advantage in predictability provided by bigrams outweighs their cost in additional parameters. For the rest of this work, we will only report on the bigram run performance. The pattern of performance for the unigram runs was similar to that of bigram runs.

6.2. *Feature transforms*

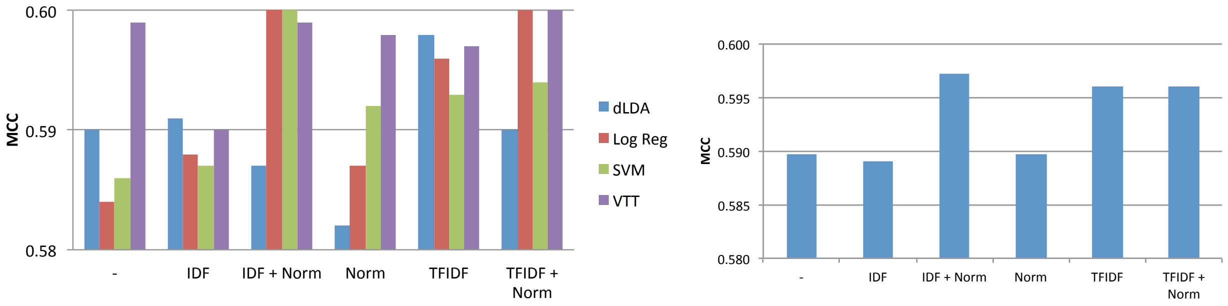


Fig. 2. MCC performance using bigram features under various transforms. '-' refers to no transform, IDF and TFIDF refer to transforms described in section 4, while IDF+Norm and TFIDF+Norm refer to those same transforms followed by unit-length normalizations. Results are shown for 4 well-performing classifiers (left); average MCC values across those 4 classifiers (right).

For simplicity, in the following sections we present performance results in terms of MCC values only. It is important to note that in most of the conditions, the 16-fold estimate of MCC performance gave a standard error on the order of 0.01; differences in performance of this scale can be ascribed to statistical fluctuations.

In figure 2, we plot the performance of the classifiers under different feature transform methods on the bigram runs. We tested these transforms under 4 classifiers: diagonal LDA (dLDA), SVM, Logistic Regression (Log Reg), and VTT. LDA performance is not reported, since as previously seen it performs badly on high-dimensional data. The binomial Naive Bayes classifier was omitted because it is not applicable to non-binary data.

The different transforms did not change performance dramatically, but some did offer advantages. VTT performed consistently well across different kinds of transforms, except for the IDF transform, where its performance decreased. As expected, SVM benefited from length normalization (whether L2-type unit-length normalization, or L1-type normalization offered by the term-frequency part of TFIDF). As seen in the bottom section of figure 2, the transforms offering good performance across a range of classifiers seemed to be those combining an IDF correction with some kind of length normalization: either IDF+Norm or TFIDF (with or without unit-length normalization).

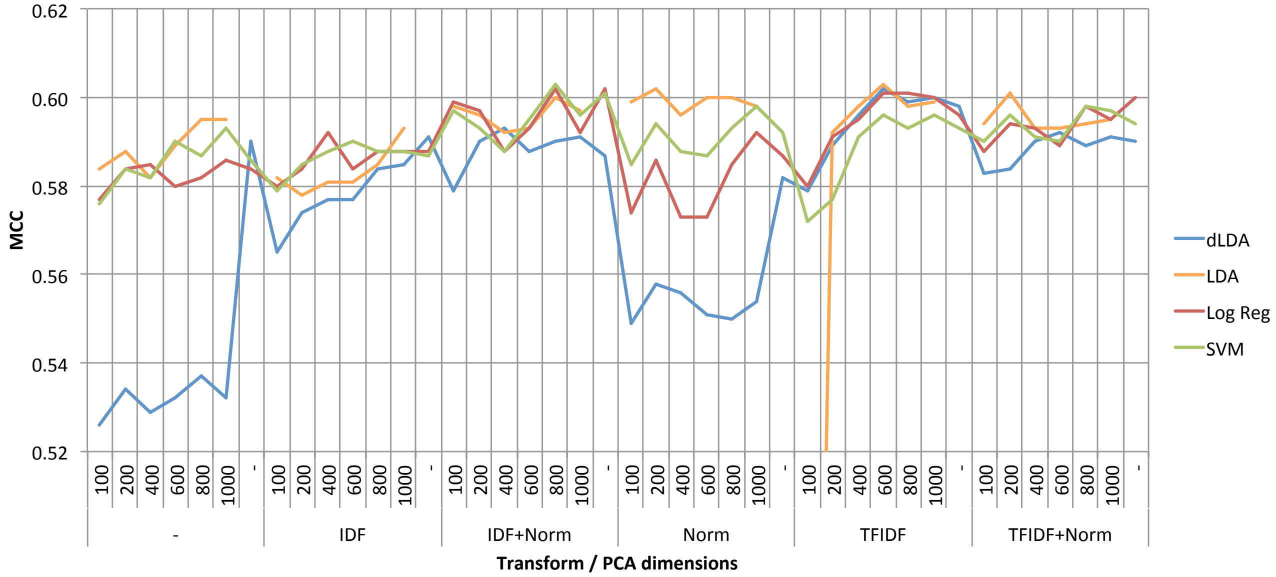


Fig. 3. MCC performance on abstracts under different feature transforms and PCA-based dimensionality reductions, bigram runs. The very bottom lists different transforms, while the numbers refer to the number of principal components kept. '-' refers to both no transform (original data matrix) and to no dimensionality reduction, as appropriate.

6.3. *Dimensionality reduction*

Figure 3 shows the performance of 4 classifiers under PCA-based dimensionality reduction on the bigram runs. Here, after applying the previously described transforms, the data matrices are projected onto their principal components. This generates smaller-dimensional, non-sparse data matrices. In this case, we have omitted the VTT classifier, since it does not generalize to non-sparse datasets. We have also omitted the binomial Naive Bayes classifier, since it is not applicable to non-binary data.

Dimensionality reduction only has a significant effects on performance for LDA, where this is expected. Because LDA requires an estimate of the full feature covariance matrix, it does not perform well when the data is very high-dimensional (and hence, the covariance matrix is difficult to estimate). However, under dimensionality reduction LDA performs extremely well, often outperforming other classifiers. Figure 4 shows the performance of different classifiers under different dimensionality reductions, now averaged across the 6 feature transforms described previously. Interestingly performance tends to increase as more principal components are kept. With 1000 principal components, LDA has the best on-average performance, though SVM also does well here. On the other hand, Diagonal LDA – which does not take into account feature covariances – does not perform well under dimensionality reduction.

7. Classification performance on abstracts with NER

The above runs used the occurrences of unigrams and bigrams as features. We have previously used features extracted using Named Entity Recognition (NER) tools in order to improve classification performance on a protein-protein interaction text mining task.[14,19,20] NER identifies

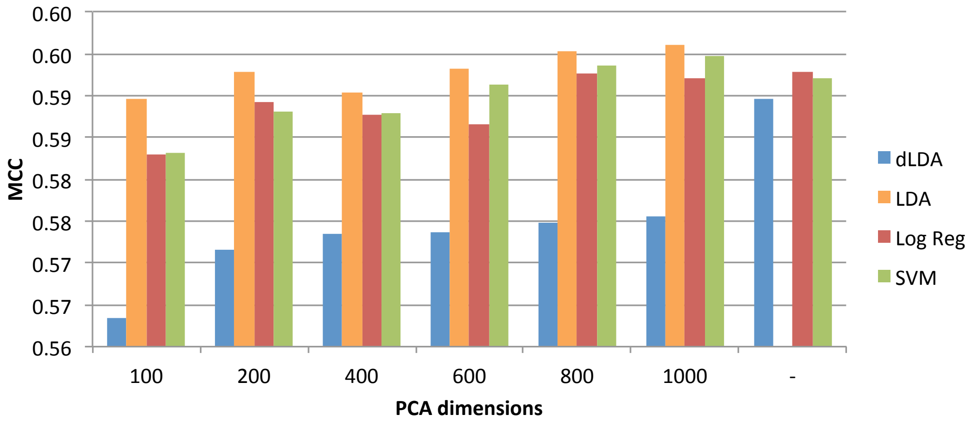


Fig. 4. MCC performance of different classifiers under feature transforms and dimensionality reduction condition, but now averaged across different feature transforms, bigram runs. The bottom axis refers to number of principal components kept, and '-' refers to no dimensionality reduction.

occurrences of named entities (for example, drugs, proteins, or chemical names) in documents. We applied a set of NER extraction tools and used the count of named entities identified in each document as an additional document feature, on top of the textual occurrence features previously discussed.

The following publicly-available tools were used to identify named entities:

- OSCAR4:[29] a recognizer of chemical names
- ABNER:[30] biomedical named entity recognizer for proteins
- DrugBank:[31] a database of drug names
- BICEPP:[32] a recognizer of clinical characteristics associated with drugs

We also identified named entities using the following dictionaries, provided by Li's lab:[21]

- i-CYPS: a dictionary of cytochrome P450 [CYP] protein names, a group of enzymes centrally involved in drug metabolism
- i-PkParams: a dictionary of pharmacokinetic parameters
- i-Transporters: a dictionary of proteins involved in transport
- i-Drugs: a dictionary of Food and Drug Administration's drug names

For SVM, Logistic Regression, and LDA, the NER counts were treated as any other feature. Diagonal LDA was omitted since it was outperformed by dimensionality-reduced LDA, and binomial Naive Bayes was omitted since NER-count features are non-binary. VTT incorporates NER-count features via a modified separating hyperplane equation:

$$\sum_i \theta_i x_i - \sum_j \frac{\beta_j - c_j}{\beta_j} - \lambda = 0$$

where x_i represent non-NER feature occurrences, θ_i and λ are textual feature weighting and bias parameters as described in section 3, c_j is the count of NER features produced for the current document by NER tool j, and β_j is a cross-validated weighting term for NER tool j.

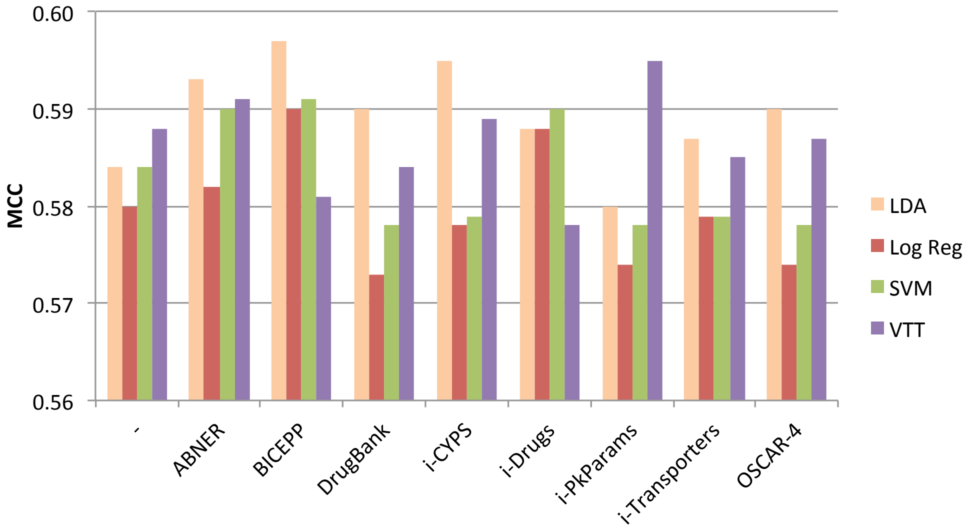


Fig. 5. MCC performance of the classifiers in combination with different NER features on the bigram runs. Classifiers used non-transformed data matrices, apart from LDA which was applied to an occurrence matrix projected onto its first 800 principal components.

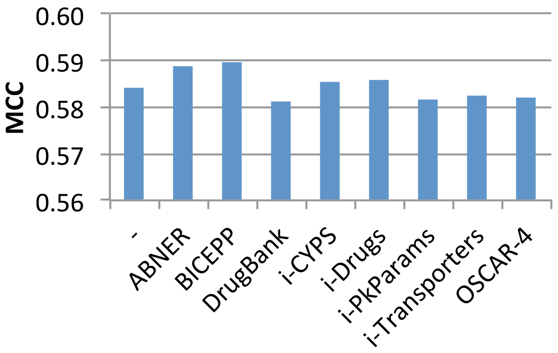


Fig. 6. MCC performance when using NER features on the bigram runs, averaged across the 4 classifiers shown in figure 5.

The classifiers were run on occurrence matrices with no transform applied, except for LDA, which was run on occurrence matrices projected onto their first 800 principal components. Each run utilizes NER features from a single tool, to test their individual merit on this task. It is important to note that in the presence of NER count features, whose values are of a different magnitude from those of binary occurrence features, length normalization can significantly hurt classifier performance (data not shown).

Figure 5 shows the performance of the different classifiers on a combination of bigram and NER features, while figure 6 shows the same performance averaged across classifiers. Given the scale of standard errors of MCC performance estimates (~0.01), it does not appear that NER features offer a significant improvement in classification rates. We also attempted to use combinations (pairs) of NER features in classification, but this also failed to improve performance (data not shown). We discuss possible reasons for this in the final section.

8. Discussion

We studied the performance of BLM on the problem of automatically identifying DDI-relevant PubMed abstracts, that is those containing pharmacokinetic evidence for the presence or absence of drug-drug interactions (DDI). We compared the performance of several linear classifiers using different combinations of unigrams, bigrams, and NER features. We also tested the effect several feature transformation and normalization methods, as well as dimensionality-reductions to different numbers of principal components.

Several of the classifiers achieved high levels of performance, reaching MCC scores of ~0.6, F1 scores of ~0.8, and iAUC scores of ~0.86. Bigrams in combination with unigrams tended to perform better than unigrams alone, and the combination of document-frequency and length normalization also tended to have a slight positive effect on performance. This effect may have been more pronounced if we had used count (instead of occurrence) matrices, in which document vector magnitudes are more variable. In addition, we also implemented PCA-based dimensionality reduction. Its effect on performance was mild for most classifiers, except for linear discriminant analysis (LDA). We observed dismal LDA performance with no dimensionality reduction, and high performance when data matrices were projected onto their first 800-1000 principal components. This is consistent with the well-known weakness of LDA in high-dimensional classification contexts.

Both relevant and irrelevant training sets came from the field of pharmacokinetics and, for this reason, shared very similar feature statistics. This makes distinguishing between them quite a difficult text classification problem – though also a more practically relevant one (such as in a situation where a researcher needs to automatically label a pre-filtered a list of potentially relevant documents). It may also explain why the NER features did not make a positive impact on classification performance: the documents in both classes would be expected to have similar counts of drug names, proteins, and other named entities, and so these counts would not help class separation. It is possible, of course, that the use of NER more finely tuned to DDI, relation extraction, or some other more sophisticated feature-generation technique could improve performance.

To conclude, the best performing classifiers and feature-transforms led to similar upper limits of performance, suggesting a fundamental limit on the amount of statistical signal present in the labels and feature distributions of the corpus. However, to achieve near-optimal generalization performance, selecting the proper combination of classifier, feature transforms, and dimensionality-reduction is necessary. When working with classifiers that contain cross-validated parameters, this can be done through the use of nested cross-validation. We provide a thorough report of the performance of supervised classifiers on this text classification scenario. Linear classifiers with common feature transforms provide a justifiable, well-understood "lower-bound" for classification performance.

Using such procedures, given the reasonable performance achieved here, we show that under realistic classification scenarios, automatic BLM techniques can identify reports of DDIs backed by pharmacokinetic evidence in PubMed abstracts. These reports can be essential in identifying causal mechanics of putative DDIs, and can serve as input for further *in vitro* pharmacological and *in populo* pharmaco-epidemiological investigation. Thus, our work shows that this text classification task is tractable, providing an essential step in enabling further development of interdisciplinary translational research in DDI.

Acknowledgments

This work was supported by the *Indiana University Collaborative Research Grant* "Drug-Drug Interaction Prediction from Large-scale Mining of Literature and Patient Records."

References

1. C. Jankel and L. Fitterman, *Drug safety* **9**, p. 51 (1993).
2. M. Becker *et al.*, *Pharmacoepidemiol Drug Saf.* **16**, 641 (2007).
3. L. Leape *et al.*, *JAMA* **274**, 35 (1995).
4. E. Hajjar, A. Cafiero and J. Hanlon, *Am. J. Geriatr. Pharmacother.* **5**, 345 (2007).
5. N. Tatonetti *et al.*, *Clinical Pharmacology & Therapeutics* **90**, 133 (2011).
6. H. Shatkay and R. Feldman, *Journal of Computational Biology* **10**, 821 (2003).
7. L. Jensen, J. Saric and P. Bork, *Nature Reviews Genetics* **7**, 119 (2006).
8. K. Cohen and L. Hunter, *PLoS Comput. Biol.* **4**, p. e20 (2008).
9. F. Leitner *et al.*, *Nature Biotechnology* **28**, 897 (2010).
10. M. Krallinger *et al.*, *BMC bioinformatics* **12**, p. S3 (2011).
11. A. Rechtsteiner *et al.*, Use of text mining for protein structure prediction and functional annotation in lack of sequence homology, in *Joint BioLINK and Bio-Ontologies Meeting (ISMB Special Interest Group)*, 2006.
12. R. McDonald *et al.*, *Bioinformatics* **20**, 3249 (2004).
13. H. El-Shishiny, T. Soliman and M. El-Asmar, Mining drug targets based on microarray experiments, in *Computers and Communications, IEEE Symposium on*, 2008.
14. A. Abi-Haidar *et al.*, *Genome Biology* **9**, p. S11 (2008).
15. I. Segura-Bedmar *et al.*, *BMC Bioinformatics* **11**, p. S1 (2010).
16. B. Percha, Y. Garten and R. Altman, Discovery and explanation of drug-drug interactions via text mining., in *Pacific Symposium on Biocomputing*, 2012.
17. N. Tatonetti, P. Patrick, R. Daneshjou and R. Altman, *Sci. Transl. Med.* **4**, 125ra31 (2012).
18. Z. Wang *et al.*, *J. Biomed. Inform.* **42**, 726 (2009).
19. A. Lourenço *et al.*, *BMC Bioinformatics* **12**, p. S12 (2011).
20. A. Kolchinsky *et al.*, *IEEE/ACM Trans. Comput. Biol. Bioinf.* **7**, 400 (2010).
21. H. Wu *et al.*, *BMC Bioinformatics (under revision)* (2012).
22. F. Pedregosa *et al.*, *JMLR* **12**, 2825 (2011).
23. R. Fan, K. Chang, C. Hsieh, X. Wang and C. Lin, *JMLR* **9**, 1871 (2008).
24. P. Bickel and E. Levina, *Bernoulli* **10**, 989 (2004).
25. E. Leopold and J. Kindermann, *Machine Learning* **46**, 423 (2002).
26. J. Davis and M. Goadrich, The relationship between precision-recall and roc curves, in *Proc of the 23rd International Conference on Machine Learning*, 2006.
27. B. Matthews *et al.*, *Biochimica et biophysica acta* **405**, p. 442 (1975).
28. Y. Yang and J. Pedersen, A comparative study on feature selection in text categorization, in *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997.
29. D. Jessop, S. Adams, E. Willighagen, L. Hawizy and P. Murray-Rust, *J. Cheminf.* **3**, 1 (2011).
30. B. Settles, *Bioinformatics* **21**, 3191 (2005).
31. D. Wishart *et al.*, *Nucleic Acids Research* **34**, D668 (2006).
32. F. Lin, S. Anthony, T. Polasek, G. Tsafnat and M. Doogue, *BMC Bioinformatics* **12**, p. 112 (2011).

INCORPORATING EXPERT TERMINOLOGY AND DISEASE RISK FACTORS INTO CONSUMER HEALTH VOCABULARIES

Michael Seedorff*
University of Iowa
240 Schaeffer Hall, Iowa City, IA, USA
Email: michael-seedorff@uiowa.edu

Kevin J. Peterson, BS
Department of Information Technology, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: peterson.kevin@mayo.edu

Laurie A. Nelsen, MS
Mayo Clinic Global Business Solutions, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: nelsen.laurie@mayo.edu

Cristian Cocos, PhD
Mayo Clinic Global Business Solutions, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: cocos.cristian@mayo.edu

Jennifer B. McCormick, PhD, MPP
Department of General Internal Medicine, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: mccormick.jb@mayo.edu

Christopher G. Chute, MD, DrPH
Department of Health Sciences Research, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: chute@mayo.edu

Jyotishman Pathak, PhD
Department of Health Sciences Research, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: pathak.jyotishman@mayo.edu

It is well-known that the general health information seeking lay-person, regardless of his/her education, cultural background, and economic status, is not as familiar with—or comfortable using—the technical terms commonly used by healthcare professionals. One of the primary reasons for this is due to the differences in perspectives and understanding of the vocabulary used by patients and providers even when referring to the same health concept. To bridge this "knowledge gap," consumer health vocabularies are presented as a solution. In this study, we introduce the Mayo Consumer Health

*This work was done while the author was an undergraduate summer intern at Mayo Clinic.

Vocabulary (MCV)—a taxonomy of approximately 5,000 consumer health terms and concepts—and develop text-mining techniques to expand its coverage by integrating disease concepts (from UMLS) as well as non-genetic (from deCODEme) and genetic (from GeneWiki+ and PharmGKB) risk factors to diseases. These steps led to adding at least one synonym for 97% of MCV concepts with an average of 43 consumer friendly terms per concept. We were also able to associate risk factors to 38 common diseases, as well as establish 5,361 Disease:Gene pairings. The expanded MCV provides a robust resource for facilitating online health information searching and retrieval as well as building consumer-oriented healthcare applications.

Keywords: Information Extraction; Consumer Health Vocabularies; Disease Risk Factors

1. Introduction

In the age of individualized medicine, it is becoming increasingly evident that more and more consumers are using the Internet and the World Wide Web to seek medical and health related information.[1,2] According to surveys by the Jupiter Organization and Harris Interactive, in 2007, 71% of people who used the Internet, also used it to seek health information (an increase by 37% since 2005).[3] Furthermore, it has been reported that 70% of people who obtain health information online say that it has influenced a decision about their treatment.[4] However, often due to various educational, economical, cultural, and language differences between patients and healthcare professionals, there exists a barrier in the process of gathering and interpreting health related information. One of the primary reasons for this is due to differences in perspectives and understanding of healthcare between patients and providers, as well as a significant disconnect in the vocabulary used even when they are referring to the same health concept.

Since various aspects of healthcare outcomes, including empowering consumers to make better-informed decisions and increasing patient compliance, can be affected due to this information disconnect, addressing the consumer health vocabulary problem has emerged as an important research activity in the recent past[5–7] as evidenced by services such as MedLine Plus[8] provided by the NIH. Cole et al.[9] proposed using a standardized biomedical terminology, SNOMED-CT,[10] and a commercially developed consumer health vocabulary, Intelligent Medical Object's Personal Health Terminology (PHT^{TM}), to assist patients and physicians who use common language terms to find specialist physicians with a particular clinical expertise. In particular, based on a user's input string, PHT was searched for term matching to acquire the SNOMED-CT codes (via PHT^{TM}–SNOMED map) that were in turn used to find physicians with the appropriate clinical specialty. In more recent work, the Open Access Collaboratory Consumer Health Vocabulary (OAC-CHV[11]) developed at the University of Utah contains more than 150,000 consumer health terms that are mapped to clinically oriented terms from the UMLS.[12] OAC-CHV has also been demonstrated in successfully translating clinical text from electronic medical records to consumers.[13]

While the above research has shown promising outcomes, there are several limitations hindering widespread adoption and application of these results. First, excluding OAC-CHV, the existing vocabularies for consumer health, such as PHT^{TM}, are either closed sourced, or have a commercial license. This not only prevents them from being leveraged in consumer health applications and tools, but also limits further development and community input. Second, with the recent advances in genomic medicine, the science and the role of non-genetic and genetic risk factors in disease etiology is becoming clearer. Consequently, there is an increasing need to incorporate such information within consumer health vocabularies–a requirement not adequately met by existing vocabularies. Finally, best practices in modeling vocabularies require explicit specification of relationships between the terms and concepts, as well as providing appropriate metadata (e.g., synonyms, definitions, provenance). This impacts the vocabulary management and development to semantics-based querying and navigation leveraging the vocabulary. Our preliminary findings indicate that none of the existing consumer health vocabularies, including freely available OAC-CHV, adopt such methodologies, and are developed using ad-hoc vocabulary modeling formalisms. For example, OAC-CHV is modeled and maintained using Microsoft Excel files, instead of a more formal knowledge representation language, such as OWL (Web Ontology Language).[14]

In this study, we attempt to address the first two limitations. Specifically, we introduce the Mayo Consumer Health Vocabulary (MCV) developed and maintained by the ontology team at Mayo Clinic Global Products and Services to support annotation on MayoClinic.com (`http://www.mayoclinic.com`) health portal initially launched in 1995. Currently, MCV comprises approximately 5,000 consumer health terms arranged in a taxonomy, and includes mappings to SNOMED-CT and ICD-9[15] for some of the core concepts. The terminology extends beyond the typical medical terminologies to include lifestyle terms representing consumer health concepts related to nutrition, exercise and other lifestyle behaviors that influence a persons health. While successfully used to annotate health related information (articles, documents, blog entries, multimedia etc.) within the MayoClinic.com portal[b], MCV currently lacks the coverage for several disease concepts as well as relevant disease risk factors. The current study addresses these requirements by developing text mining approaches for integrating disease concepts (from OAC-CHV[16]) as well as non-genetic (from deCODEme[17]) and genetic (from GeneWiki+[18] and PharmGKB[19]) risk factors to diseases. The integration led to adding at least one synonym for 97% of MCV concepts with an average of 43 consumer friendly terms per concept, an important step in increasing search result coverage for future versions of MayoClinic.com. We were also able to associate non-genetic risk factors to 38 common diseases, as well as establish 5,361 Disease:Gene pairings. We discuss the details of our methods and findings in the remainder of this manuscript.

2. Resources and Tools

The following resources and tools were leveraged to conduct this study.

2.1. ***Open Access Collaboratory Consumer Health Vocabulary***

The Open Access Collaboratory Consumer Health Vocabulary (OAC-CHV[16]) is created and maintained by the Consumer Health Vocabulary Initiative. It is a relationship file that links commonly used real-world vocabulary to associated medical terminology. Additionally, it provides the associated UMLS CUIs as well as understandability scores for each term and whether

[b] Our recent Web analytics statistics indicate that the MayoClinic.com portal is, on average, visited by more than 22 million unique visitors every month.

a term is disparaged (has an abnormality, such as a misspelling). In total there are 158,519 terms and 57,819 unique UMLS CUIs (2.7 terms per UMLS CUI). We used this file for finding near-matching terms to those in MCV and retrieving the connected terms based on common UMLS CUIs. We also used the UMLS CUIs connected to retrieved terms for comparison of similarity between MCV and OAC-CHV terms.

2.2. *SNOMED-CT*

The Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT[10]) was created by the College of American Pathologists and is maintained by the International Health Terminology Standards Development Organisation. It is a hierarchical ontology of medical terms. Similarity information can be gathered to compare two items in SNOMED-CT using several ontology-based algorithms such as Wu-Palmer.[20] We used SNOMED-CT as the UMLS source for comparison of UMLS CUIs and retrieval of UMLS CUI synonyms within the UMLS::Similarity and UMLS::Interface modules, respectively (see below).

2.3. *PharmGKB*

The Pharmacogenomics Knowledge Base (PharmGKB) is managed at Stanford University and focuses on maintaining information about gene:drug relationships and the corresponding gene variations, but also includes limited information on gene:disease relationships.[19] The data is collected from literature and other databases that report study results having to do with gene:drug interactions. It uses its own ID system for genes and diseases but provides data sets that allow for translation of genes into Entrez Gene IDs and diseases into SNOMED or UMLS IDs. We retrieved all Disease:Entrez Gene ID relationships and used this as a basis for our list of genetic risk factors by disease.

2.4. *deCODE genetics*

deCODE genetics[17] is a pharmaceutical company with an interest in genetic effects on disease and medicine. They sell a Direct-to-Consumer genetic testing service, called deCODEme, for sequencing a portion of an individual's genome to estimate genetic risk of various diseases. deCODEme has a website that contains information on the 47 diseases that are being tested, including information on both non-genetic and genetic factors that increase an individual's risk. We use the non-genetic factors portion of these disease pages to mine risk factors.

2.5. *UMLS::Interface*

UMLS::Interface is a Perl module that retrieves the position of UMLS CUIs from a UMLS ontology source (i.e. SNOMED-CT).[21] It provides tools for translating medical terms given as strings into the corresponding UMLS CUIs, getting positions in the ontology based on the UMLS CUI, and returning related UMLS CUIs and associated medical terms. Position in UMLS can be retrieved using a UMLS CUI or, if no UMLS CUI is available, one can be estimated based on an input string. It requires UMLS be loaded into a MySQL Database for access. We used this module to retrieve sister nodes (synonyms) for each MCV term.

2.6. *UMLS::Similarity*

UMLS::Similarity is a Perl module that retrieves a similarity score between two concepts based on their positioning in the hierarchical UMLS source (i.e. SNOMED-CT).[21] It has several options for evaluating either similarity or relatedness for two UMLS CUIs. Eight similarity measures, based on location in the ontology, were incorporated into the module (including Wu-Palmer, the similarity measure used in this study) as well as various relatedness measures that were not used in this study. This module was used for computing the similarity of MCV and OAC-CHV terms to indicate whether the relationship was valid (should be maintained) or invalid (should be deleted).

2.7. *MetaMap*

MetaMap is a program designed to extract biomedical terminology from text and map it to appropriate UMLS concepts.[22] It splits input text into minimal phrases and provides potential UMLS matches for the terms, indicating a score from 0-1000 with a higher score meaning a better match, as well as the semantic type (i.e. disease, substance, ...), UMLS source, and UMLS CUI. We used this program to extract non-genetic risk factors from plain text with the ability to divide sentences into phrases and indicate the semantic type being crucial.

3. Materials and Methods

3.1. *Materials*

The primary materials used in this study are the following:

- The February 4, 2011 OAC-CHV data set, available for download via `http://consumerhealthvocab.org`. The data set contains 158,519 mappings between medical concepts and terms along with several measures of understandability for each term. There is a one–to–many relationship between UMLS CUIs and OAC-CHV terms.
- The July 3, 2012 GeneWiki+ relationships data set, available for download via `http://genewikiplus.org/wiki/GeneWiki:Data`. The data set contains 18,230 relationships between genes and diseases, referencing the diseases using a Disease Ontology ID (DOID).[23]
- The June 13, 2012 Human Disease Ontology data set, available for download via `http://obofoundry.org`. The data set contains 8,631 entries, each with at least one DOID, and a total of 14,311 SNOMED IDs mapped to the entries.
- The July 5, 2012 PharmGKB relationships data set, available by request via `http://www.pharmgkb.org/downloads.jsp`. The data set contains 11,706 unique relationships between drugs, diseases, genes, haplotypes, and gene variant locations (see Table 1). It includes information on whether pharmacokinetic and pharmacodynamic effects play a part in the relationship as well as PubMed IDs for articles that provide evidence supporting the relationship. Also available are gene and disease data sets, providing mappings between genes and Entrez Gene IDs, and diseases and SNOMED-CT IDs, respectively.
- The MCV data set and MCV-SNOMED relationship data set, not publicly available for this study but, in the future, will be made available for public use. MCV includes a list of around 5,000 medical terms, 2,126 of which are considered core terms (directly associated with

Table 1. PharmGKB Relationships (Highlighted fields indicate relationships studied in this work)

	Haplotype	Gene	Variant Location	Drug	Disease	Entrez Gene ID	SNOMED-CT
Haplotype	0	0	0	762	169	0	0
Gene		684	0	2,578	1,541	27,421	0
Variant Location			0	3,147	2,053	0	0
Drug				0	772	0	0
Disease					0	0	4,348
Entrez Gene ID						0	0
SNOMED-CT							0

clinical concepts) and were the basis of this effort. These core terms are identified by MCV IDs and divided into 4 groups: diseases (1,443), first aid (63), symptoms (102), and test procedures (518). The MCV-SNOMED relationship data set contains 1,476 relationships between MCV IDs and SNOMED IDs.

3.2. *Methods for integrating disease concepts*

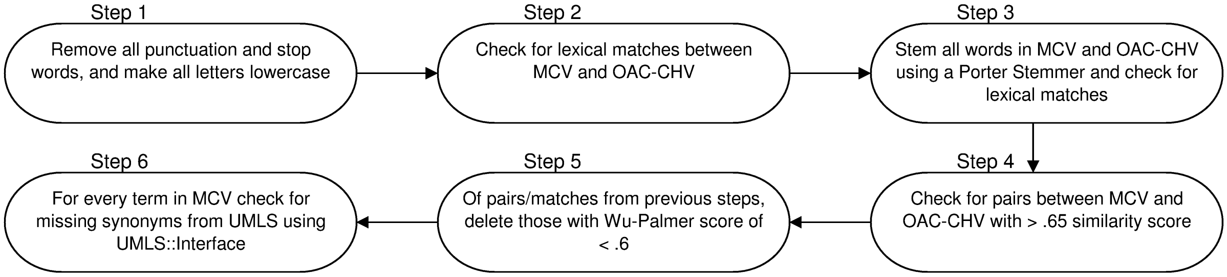

Step 1
Remove all punctuation and stop words, and make all letters lowercase
Step 2
Check for lexical matches between MCV and OAC-CHV
Step 3
Stem all words in MCV and OAC-CHV using a Porter Stemmer and check for lexical matches
Step 4
Check for pairs between MCV and OAC-CHV with > .65 similarity score
Step 5
Of pairs/matches from previous steps, delete those with Wu-Palmer score of < .6
Step 6
For every term in MCV check for missing synonyms from UMLS using UMLS::Interface

Fig. 1. Outline for linking MCV and OAC-CHV terms

For this study, we compared biomedical terms in MCV and OAC-CHV to expand the list of word alternatives for MCV. Note that traditional methods for ontology matching and alignment are not applicable here because they rely primarily on relationships between concepts as well as the hierarchical structure in the source and target ontologies (which are "metadata–based"), whereas both OAC-CHV and MCV are at present a nearly flat list of terms with minimal relationships and hierarchies. A general outline for integrating MCV and OAC-CHV is given in Fig. 1. For the strings in MCV and OAC-CHV, we removed all punctuation and stop words and made all letters lowercase. We used a specific subset of stop words that showed up often in the data to avoid deleting good words (i.e. 'a' in "vitamin a deficiency"). Because every term in OAC-CHV was paired with a UMLS CUI and a medically preferred term, we were able to create sets of potential phrases for each UMLS CUI which allowed us to retrieve a list of synonyms quickly for any entry in OAC-CHV. We began by simply seeing if any terms in MCV were exact matches to terms in OAC-CHV. This was followed by stemming all words in every term using a Porter Stemmer[24] and checking for exact matches between the two sets.

All matches were added to a matched list.

We then created a similarity score for each pair of terms between OAC-CHV and MCV. This score was calculated by giving one point to each word that was in the other term and .75 points to each stemmed word that was in the other stemmed term, summing these points, and dividing by the total number of words between the two terms. For example, the terms 'knee knees injury' and 'knee injuries' would receive a score of $(.75 + 1 + .75 + 1 + .75)/5 = .85$ (Fig. 2). Based on outcome observations, an empirical threshold of .65 was set where any pair that achieved a score equal to or over this threshold was considered to be matching and was added to the matched list.

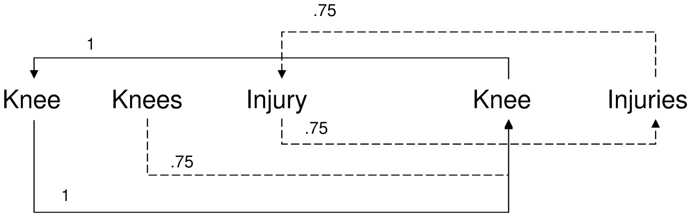


Fig. 2. Comparison scoring of two example terms

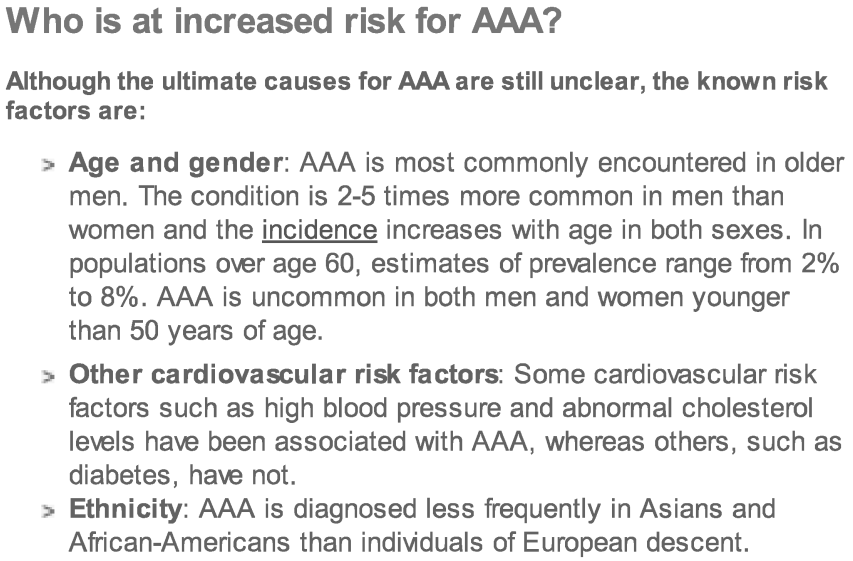
The next step was to get UMLS CUI codes for every term that had been paired in the matched list. For OAC-CHV terms, that information was already included in the file. For MCV terms, we used a relationships file developed by Mayo to get the connected SNOMED IDs. With those SNOMED IDs, we queried the BioPortal REST service which returned the appropriate UMLS CUIs.[25]

Next we evaluated the strength of the SNOMED relationship between each pair, using their UMLS CUI codes and the UMLS::Similarity module. MCV terms that were connected to multiple UMLS CUIs had the highest similarity score counted for each pairing. MCV terms which were connected to no UMLS CUIs did not go through this step. The similarity measure used was the Wu-Palmer Similarity score,[20] a measure that ranged from 0 (exclusive) to 1 (inclusive) with a larger number indicating two UMLS CUIs being more similar. Based on output observations, we set a threshold of .6 where any pair scoring below that would be deleted from the list of pairings.

Once all pairings had been computed, we began gathering synonyms for MCV terms. For every pairing between MCV and OAC-CHV, the OAC-CHV term was connected to a group of terms with the same UMLS CUI. For every pairing, this group of OAC-CHV terms was added to the correct MCV. UMLS::Interface was then queried for equivalent terms to every MCV term. These two groups of synonyms were combined for each MCV term and duplicate synonyms were deleted.

3.3. *Methods for integrating non-genetic and genetic disease risk factors*

For the second part of this study, we integrated non-genetic and genetic risk factors to diseases in MCV. Non-genetic factors were obtained by mining information from deCODEme's website for most of the 47 medical conditions that they do genetic testing on. The text mining algorithm was implemented using the XML and RCurl packages in R.[26–28] First, a list of diseases was queried from the "about deCODEme" page of their website. The page for each individual disease was then accessed and the associated factors were retrieved. Because the information on non-genetic risk factors was stored in consistent locations within deCODEMe's website templates (usually in bold text; as seen in Fig. 3), our retrieval algorithm processed just the relevant text area. For factors that included ambiguous terms such as 'age,' 'ethnicity,' and 'gender,' we developed the following heuristics based on typical structures of the paragraphs that followed the highlighted function:



Who is at increased risk for AAA?

Although the ultimate causes for AAA are still unclear, the known risk factors are:

- **Age and gender**: AAA is most commonly encountered in older men. The condition is 2-5 times more common in men than women and the incidence increases with age in both sexes. In populations over age 60, estimates of prevalence range from 2% to 8%. AAA is uncommon in both men and women younger than 50 years of age.
- **Other cardiovascular risk factors**: Some cardiovascular risk factors such as high blood pressure and abnormal cholesterol levels have been associated with AAA, whereas others, such as diabetes, have not.
- **Ethnicity**: AAA is diagnosed less frequently in Asians and African-Americans than individuals of European descent.

Fig. 3. Sample of risk factor portion of deCODEme site

- Gender – Typically the first gender to show up in the paragraph was at higher risk. When no gender was at higher risk, then either no gender was named or the first instance of a gender in the paragraph was accompanied by a conjunction and the opposite gender. For instance, in Fig. 3, "AAA is most commonly encountered in older men" would give us 'men' as the higher risk group, because it is spotted first in the paragraph. However, if the sentence were to instead say "AAA is most commonly encountered in older men and women," we would not assume a higher risk group.
- Age – There were many different structures for ages being described. We made a list of the typical ones for querying the text such as "over age ##," "between the ages of ## and ##," and "in their ##s." For instance, in Fig. 3, the phrase "over age 60" indicates that 60+ is a high risk group.
- Ethnicity – Typically there were many ethnicities mentioned and there was a rough ordering indicated by the comparison words used. Words such as 'more,' 'highest,' and 'fourfold'

indicated that the earliest ethnicities in the paragraph were at higher risk while the word 'less' indicated that the earliest ethnicities following the word 'than' represented high risk groups. Our method deleted all words that did not have to do with this ordering and were not ethnicities, allowing us to extract ethnicities based on locations of comparison words. For example, in Fig. 3 the ethnicity sentence is reduced to "less Asians African-Americans than European" and European would be chosen as the high risk group.

- 'Other' Categories – Categories that included the word 'other' in their title often listed many risk factors but did not have a uniform structure, making it much more difficult to extract the factors. To solve this problem we ran the paragraphs through MetaMap, a biomedical terminology extraction tool which split the paragraph up into concepts and provided expected semantic categories as well as goodness-of-fit scores. We took the terms which were substance, disease, or injury related, based on their semantic categories, and, if they had a perfect fit score of 1000, added them to the non-genetic factors list. In addition, if the words 'smoking,' 'alcohol,' or 'cocaine' were found, they were added to the factors list, even without a perfect goodness-of-fit score.

The second type of factor that we looked at was genetic. Initially we extracted all SNOMED IDs that were linked to each MCV ID by processing MCV's relationships file. The Human Disease Ontology[23] holds relationships between SNOMED IDs and DOIDs, allowing us to extend our connections between MCV IDs and DOIDs (Disease Ontology IDs). Using these relationships, we queried the GeneWiki+ data set to retrieve genes that were correlated to each DOID, and by extending that relationship to MCV and accumulating the genes, we created a relationship file between MCV IDs and Entrez Gene IDs.

In addition to using the GeneWiki+ data set, we also had access to PharmGKB relationships files which, among other things, linked diseases and genes through their PharmGKB Accession IDs. Subsequently, by using the PharmGKB genes relationships file, we replaced the listed genes with their Entrez Gene IDs. Similarly, by using the PharmGKB diseases relationships file, we replaced the diseases in the relationships with the connected SNOMED-CT IDs. We then replaced these SNOMED-CT IDs with the connected MCV IDs from MCV's relationships file and added any MCV:Entrez Gene ID pairs that were missing from GeneWiki+ to our list of MCV ID:Entrez Gene ID relationships.

4. Results

The MCV file we began with included 2,126 terms. After just looking for exact matches or stemmed perfect matches, 1,677 terms had found matches in OAC-CHV and 449 had not. When we did not use UMLS::Similarity to evaluate matches, we had 2,092 terms that found matches and 34 that did not. After using UMLS:Simliarity to eliminate weak or incorrect matches we had 2,069 terms that had matches and 57 that did not. Table 2 shows a summary of these findings.

On average, each term in MCV had 50.2 synonyms when not checking against UMLS::Similarity, but just 38.5 synonyms after incorporating this extra measure. UMLS::Interface averaged adding 4.5 synonyms to each term in MCV with a final average output of 43 synonyms per MCV term.

Table 2. Summary of MCV terms mapping results

	MCV Terms mapped to OAC-CHV	MCV Terms *not* mapped to OAC-CHV
Perfect Matches	1,646	480
Perfect Matches after stemming	1,677	449
Close matches using algorithm	2,092	34
Matches after UMLS::Similarity	2,069	57

deCODEme contained information on 47 diseases or conditions. Of these, five either did not have non-genetic factor information in the usual area (in lists within the main text area) or did not have any non-genetic factor information at all. Of the 42 that did contain non-genetic factor information, 38 matched either an MCV name or one of the synonyms previously created. On average each of these 38 diseases had 6.7 non-genetic factors gathered from deCODEme.

GeneWiki+ contained information on 18,230 Gene:Disease relationships and a total of 10,084 unique Entrez Gene ID:DOID relationships. There were a total of 361 diseases and seven symptoms from MCV that mapped to at least one gene and a total of 4,884 mappings between MCV entries and Entrez Gene IDs (once the MCV IDs had been processed into SNOMED IDs and then DOIDs).

The PharmGKB relationships file contained a total of 11,706 unique relationships, but only 1,541 of those were between diseases and genes. There were 570 MCV ID:Entrez Gene ID relationships recorded after tracking the DOIDs to the corresponding SNOMED-CT IDs and then MCV IDs. Of these, 93 already existed in the GeneWiki+ information and 477 were new. See Table 3 for a summary of these results. After including the PharmGKB information, coverage of MCV terms was the same (361 diseases and seven symptoms).

Table 3. Matching between Diseases and Genes

	MCV:Entrez Gene Pairs
Only in GeneWiki+	4,791
In both GeneWiki+ and PharmGKB	93
Only in PharmGKB	477
Total	5,361

5. Discussion

The principle goal of this study was to map terms and concepts from MCV to synonyms or near-synonyms from publicly available sources. Connecting similar terms from OAC-CHV, checking the quality of these matches using UMLS::Similarity, and extracting close relations from UMLS::Interface expanded the base list of terms by more than 43 times and over 97% of terms in MCV added at least one synonym. Having such a list will allow for improved search results that minimize the difficulty of finding an exact phrase to retrieve information on an expected medical concept.

Our extraction of genetic factors was also very helpful in adding to MCV. GeneWiki+ and PharmGKB each added a valuable amount of gene:disease matchings with GeneWiki+ contributing somewhat more, reasonable considering PharmGKB specializes in gene:drug relationships. A large number of relationships presented in these files were unable to be mapped to any diseases in MCV due to either MCV lacking the disease or one of the ID relationship files being incomplete. With only 42 diseases from deCODEme having non-genetic risk information, it may have been more valuable to just manually edit those relationships. Extraction of ethnicity, gender, and age information was valuable but many factors were included in the 'other' categories and were not always correctly retrieved by MetaMap. It may be worthwhile to map these to a database of risk factors at some point, but that was not considered in this study.

6. Conclusion

In this study we integrated synonyms for medical terminologies as well as both non-genetic and genetic risk factors for diseases into MCV. Bringing this information into medical query services oriented towards consumers is an important step to providing better results and risk information that is growing in importance, especially as genetic risks become better known. The expanded version of MCV created in this exercise provides a solid basis for creation of consumer-oriented healthcare applications and online health information searching. With MCV becoming publicly available in the future, current limitations due to many consumer health vocabulary sources being closed source should be reduced.

7. Acknowledgments

This study was funded in part by the Mayo-NIH Relief Fund Award (FP00068486) and Mayo Clinic Early Career Development Award (FP00058504).

References

1. D. Borzekowski and V. Rickert. Adolescent Cybersurfing for Health Information: A New Resource That Crosses Barriers. *Archives of Pediatrics and Adolescent Medicine*, 155(7):813–817, 2001.
2. R. J. W. Cline and K. M. Haynes. Consumer Health Information Seeking on the Internet: The State of the Art. *Health Education Research*, 16(6):671–692, 2001.
3. Harris Interactive: Consumer Health Care Survey Reveals Mixed Bag of Results. Last accessed: 6th October, 2009.
4. Robert J. Bensley and Jodi Brookins Fisher. *Community Health Education Methods: A Practical Guide*. Jones and Bartlett Publishers, 2008.
5. Rita D. Zielstorff. Controlled Vocabularies for Consumer Health. *Journal of Biomedical Informatics*, 36(4-5):326–333, 2003.
6. Catherine Smith and P. Stavri. Consumer health vocabulary. *Consumer Health Informatics: Informing Consumers and Improving Health Care*, pages 122–128, 2005.
7. Q.T. Zeng and T. Tse. Exploring and Developing Consumer Health Vocabularies. *Journal of American Medical Informatics Association*, 13(1):24–29, 2006.
8. N. Miller, E. M. Lacroix, and J. E. Backus. MEDLINEplus: Building and Maintaining the Na-

tional Library of Medicine's Consumer Health Web Service. *Bulletin of the Medical Library Association*, 88(1):11–17, 2000.

9. Curtis L. Cole, Andrew S. Kanter, Michael Cummens, Sean Vostinara, and Frank Naeymi-Rad. Using a Terminology Server and Consumer Search Phrases to Help Patients Find Physicians with Particular Expertise. *Proceedings of the 11th World Congress on Medical Informatics: MED-INFO*, 107:492–496, 2004.
10. SNOMED-CT: Systematized Nomenclature of Medicine-Clinical Terms. Last accessed: 10th July, 2012.
11. Q.T. Zeng, T. Tse, G. Divita, A. Keselman, and et al. Term Identification Methods for Consumer Health Vocabulary Development. *Journal of Medical Internet Research*, 9(1):e4, 2007.
12. Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32(Database issue):267–270, 2004.
13. Qing T. Zeng, S. Goryachev, H. Kim, A. Keselman, and S. Rosendale. Making Texts in Electronic Health Records Comprehensible to Consumers: A Prototype Translator. In *AMIA Annual Syposium*, pages 846–850, 2007.
14. Deborah L. McGuinness, Frank van Harmelen, and et al. OWL: Web Ontology Language. In *http://www.w3.org/2004/OWL/*, 2004.
15. World Health Organization International Classification of Diseases (ICD-9) Clinical Modification. Last accessed: 7th October, 2009.
16. Q. T. Zeng, T. Tse, G. Divita, A. Keselman, J. Crowell, A. C. Browne, S. Goryachev, and L. Ngo. Term identification methods for consumer health vocabulary development. *J. Med. Internet Res.*, 9(1):e4, 2007.
17. Decodeme. `http://www.decodeme.com/`.
18. Genewiki+. `http://genewikiplus.org/`.
19. R. B. Altman. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat. Genet.*, 39(4):426, Apr 2007.
20. Z. Wu and M. Palmer. Verb semantics and lexical selection. In *In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pages 133–138, Las Cruces, NM, 1994.
21. McInnes, Pedersen, and Pakhomov. Umls-interface and umls-similarity : Open source software for measuring paths and semantic similarity. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 431–435, San Francisco, CA, November 2009.
22. A. R. Aronson and F. M. Lang. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3):229–236, 2010.
23. J. D. Osborne, J. Flatow, M. Holko, S. M. Lin, W. A. Kibbe, L. J. Zhu, M. I. Danila, G. Feng, and R. L. Chisholm. Annotating the human genome with Disease Ontology. *BMC Genomics*, 10 Suppl 1:S6, 2009.
24. Kurt Hornik. *Snowball: Snowball Stemmers*, 2012. R package version 0.0-8.
25. P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.*, 39(Web Server issue):W541–545, Jul 2011.
26. Duncan Temple Lang. *XML: Tools for parsing and generating XML within R and S-Plus.*, 2012. R package version 3.9-4.1.
27. Duncan Temple Lang. *RCurl: General network (HTTP/FTP/...) client interface for R*, 2012. R package version 1.91-1.1.
28. R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

DETECTION OF PROTEIN CATALYTIC SITES IN THE BIOMEDICAL LITERATURE

KARIN VERSPOOR* and ANDREW MACKINLAY

National ICT Australia, Victoria Research Lab
Parkville, VIC 3010 Australia
*E-mail: *karin.verspoor@nicta.com.au, andrew.mackinlay@nicta.com.au*

JUDITH D. COHN and MICHAEL E. WALL

Computer and Computational Sciences Division, Los Alamos National Laboratory,
Los Alamos, NM 87545 USA
E-mail: jcohn@lanl.gov, mewall@lanl.gov

This paper explores the application of text mining to the problem of detecting protein functional sites in the biomedical literature, and specifically considers the task of identifying catalytic sites in that literature. We provide strong evidence for the need for text mining techniques that address residue-level protein function annotation through an analysis of two corpora in terms of their coverage of curated data sources. We also explore the viability of building a text-based classifier for identifying protein functional sites, identifying the low coverage of curated data sources and the potential ambiguity of information about protein functional sites as challenges that must be addressed. Nevertheless we produce a simple classifier that achieves a reasonable ∼69% F-score on our full text silver corpus on the first attempt to address this classification task. The work has application in computational prediction of the functional significance of protein sites as well as in curation workflows for databases that capture this information.

Keywords: text mining, information extraction, machine learning, catalytic site, biomedical literature, biomedical natural language processing, protein functional sites

1. Introduction

To facilitate progress in understanding and prediction of protein function, it is critical to populate databases with information about the physical aspects of protein function,[1,2] including the location of functionally important residues on the protein and the biochemical properties of ligand-protein interactions. Drug discovery for treatment of diseases proceeds systematically from this information; drugs can be designed to target a specific functionally important site on the protein and can become the basis for large-scale drug screening experiments. However, such physical information is currently scarce compared to more qualitative information about protein function, such as pathway assignments or Gene Ontology annotations, despite its critical importance for characterization and eventual manipulation of protein behavior.

In previous work, we have shown that text mining can be integrated with protein structure-based methods for prediction of protein functional sites to identify high-quality predictions that are supported by evidence in the biomedical literature.[3] The method we developed in that work is called Literature-Enhanced Automated Prediction of Functional Sites, or LEAP-FS. While that work showed that we were able to recover a good proportion of curated functional site annotations in existing databases, it did not attempt to classify the functional importance of each site more specifically, e.g., identifying catalytic sites. Automated identification of catalytic sites in the biomedical literature has application, for example, in genome annotation pipelines and in drug design. Such predictions provide fine-grained information regarding the biological significance of a specific functional site, influencing both the overall understanding of the role of a protein in a biological process, and how that protein might be modulated through drug intervention. The classification can also be employed within the curation pipeline for the development of resources such as the Catalytic Site Atlas[4] to assist in identifying meaningful literature to be curated, and to highlight specific residue mentions within that literature that should be considered for inclusion in the resource.

Development of a functional site classifier would lay the foundation for generalizing the methods we previously developed in LEAP-FS. First, it supports the generalization of the methods to a broader set of the biomedical literature; we would like to be able to identify functionally important protein residues through analysis of literature that is not directly connected to proteins via curated links. The classifier would play a role in that generalization by assisting in the recognition of literature where residues are specifically discussed as being catalytically active. Second, classification of a residue mention as within a catalytic site gives us increased confidence that the residue mention is relevant for prediction of functional sites – it provides evidence that the residue is mentioned due to its functional importance.

In this work, we take a step towards finer-grained analysis of amino acid residues mentioned in text. We provide an analysis of the residues identified in publications linked to proteins in the Protein Data Bank in order to understand what relevant information is readily available in curated data sources. We further explore the development of a classifier which can classify amino acid residues as catalytic residues based on the textual context of the residue mention, e.g. for the positive cases below.

(1) We propose a mechanism involving general base catalysis by the carboxy-terminal **Trp270** carboxyl group [PMID 12356304]
(2) it is possible that **Arg 381** is one of the catalytic bases previously observed [PMID 9174368]

We find that while our classifier does a reasonable job in classifying positive instances, there is significant ambiguity around negative instances, both for the purpose of developing training material and during classification of held-out test data.

2. Related Work

BindingMOAD is a database of protein ligand-binding sites, which is updated through curation of approximately 2000 full text publications each year.[5] To assist with this curation, a natural language processing system called BUDA, using the rule-based GATE system[6] (gate.ac.uk) was developed to identify articles relevant to binding, and to extract protein-ligand interactions along with quantitative binding affinities. However, it has been noted that the curation of BindingMOAD cannot rely completely on this automated information extraction, due to ambiguities that persist in the extracted information.[7] Furthermore, specific evaluation results of BindingMOAD have not been made available so we are unable to make detailed comparisons.

The Open Mutation Miner[8] system extracts mutation information from full text publications and identifies mutation impact information including protein properties such as kinetic and stability data. The system also aims to capture protein function impacts, through detec-

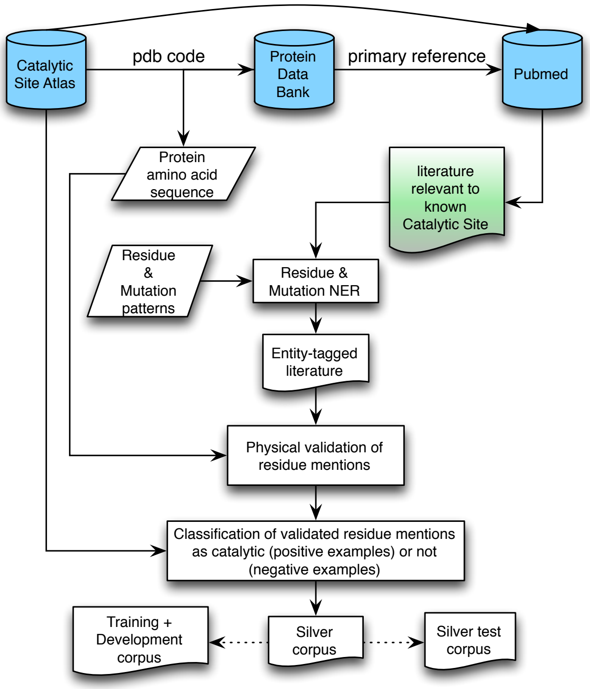

Catalytic Site Atlas
pdb code
Protein Data Bank
primary reference
Pubmed
Protein amino acid sequence
literature relevant to known Catalytic Site
Residue & Mutation patterns
Residue & Mutation NER
Entity-tagged literature
Physical validation of residue mentions
Classification of validated residue mentions as catalytic (positive examples) or not (negative examples)
Training + Development corpus
Silver corpus
Silver test corpus

Fig. 1. Architecture for distant learning silver corpus creation

tion of Gene Ontology[9] molecular function terms via dictionary look-up with some morphological processing. A rule-based strategy is employed for association (grounding) of an impact to a mutation. We refer the reader to their work[8] for a thorough review of other text mining systems that focus on extracting mutation information. The scope of these systems is different than our work, as we are interested in detecting all specific protein residue mentions, not only mutation sites, and we focus on catalytic and ligand binding sites.

Nagel et al[10] address annotation of individual protein residues, focusing on binding and enzymatic activity. The goals of that work are closest to our own goals. They develop a mutual information-based classifier for detection and categorization of functional terms within a sentence, achieving F-scores of 0.57 for binding and 0.27 for enzymatic activity on a very small corpus of 100 manually annotated abstracts; association of residues to these terms is achieved using syntactic relationships. However, only 16 extracted functional annotations are produced by the system. The authors compared those annotations to information in UniProt and did find that that the text mining identified correct annotations, of which 11 were not already present in the resource. The evaluation of this work was very limited in scope; we will present a much larger-scale evaluation.

3. Methods

3.1. *Amino Acid residue detection in text*

To detect residues in biomedical publications, we employ a set of patterns that take advantage of regularities in how protein residues and mutations are expressed in text. These patterns are described in detail in previous work;[3,11] that work showed that we were able to achieve F-scores of over 94% on a third-party gold standard.[10] For the analysis reported here, we re-used the previous results we generated, which included detection of single-letter abbreviations for mutations (e.g. *D199S*) but ignored single-letter abbreviations for individual residues. All residue and mutation mentions were normalised to a three-letter residue abbreviation with the specific position of the residue.

3.2. *Corpus creation*

We created two corpora for processing: a corpus of abstracts and a corpus of full text publications. The starting point for the creation of each of these corpora was the set of 17,595 PubMed identifiers referenced as the primary citation of records in the Protein Data Bank[12] (PDB; www.pdb.org), using PDB data downloaded in May 2010 for the LEAP-FS system.

We were able to retrieve the full set of abstracts from a local Medline repository. We were also able to successfully retrieve 11,560 full text publications from this set. After running the residue and mutation detection step over these corpora, we identified 6,109 abstracts and 8,491 full text publications with residue mentions. We next applied a physical verification step for those residue mentions, in which each amino acid mention identified in the text must be matched to a physical residue in the corresponding PDB record, with both the position in the protein sequence and the specific amino acid matching (for mutations, either the wild type or the mutated amino acid was allowed to match). This step ensures that residues identified in the text are grounded to the appropriate protein sequence, and resulted in identifying 5,236 abstracts and 7,309 full text publications with physically verified text residues (**PVTR**) (in each case representing 86% of the original corpus).

3.3. *Distant learning for training data creation*

To avoid costly manual annotation of training data, we take advantage of high quality external knowledge to automatically generate appropriate training data. We have previously explored this strategy for creation of training data for extraction of protein-residue associations from text.[11] We extend that approach here to create a "silver standard" data set – i.e. training data that we believe to be highly reliable, but which has not been manually verified. The architecture of the approach is outlined in Figure 1.

The silver corpus creation starts with the abstract and full text corpora we collected, with each physically-verified text residue in the corpus serving as a potential training example. The annotation of the text residues in the corpus as well as the sub-selection of publications for the silver standard data set relies on external curated data. The external knowledge we rely on to build our training and test data is the curated links to the literature from the Protein Data Bank which form the basis of our corpus creation, coupled with literature-curated annotations of catalytic sites available in the Catalytic Site Atlas[4] (CSA; www.ebi.ac.uk/thornton-srv/databases/CSA/). We refer to the subset of CSA annotations that are marked as coming from the literature as **CSA-Lit**. The annotated catalytic sites in CSA-Lit represent highly reliable positive information about the residues in PDB records that are catalytically active. Our training data focuses on the publications that have at least one physically-verified text residue that is annotated in CSA-Lit.

Because catalytic sites are also binding sites, and generally are a subset of functional sites, there is significant potential ambiguity about whether a given functional site is catalytic. To ensure that our training data cleanly captures specifically catalytic sites as positive instances but does not inadvertently include a catalytic site as a negative instance, we refer to other functional site resources to discard potentially ambiguous cases. Catalytic sites that are not in the CSA-Lit subset but are annotated in CSA (usually on the basis of sequence alignment with a known catalytic site) are discarded as they are not definitively catalytic, but very likely to be. We also consider any site identified in BindingMOAD as a binding site and any residue that is near a small molecule (NSM) in the corresponding PDB structure (see[3] for details on how this is formally determined) as ambiguous. The logic we employ is formalized as follows:

For each PubMed article with at least one physically-verified text residue in CSA-Lit, for each physically-verified text residue in the article,

(1) is it in CSA-Lit? (if yes, annotate as positive instance)
(2) is it in CSA? (if yes, discard)
(3) is it annotated in BindingMOAD? (if yes, discard)
(4) is it a residue near a small molecule in the PDB structure? (if yes, discard)
(5) otherwise, annotate the text residue as a negative instance.

Application of this strategy results in an imbalanced silver corpus for the abstracts, with 749 positive instances and 179 negative instances (in 259 abstracts), and a significantly larger and more balanced silver corpus for the full texts, with 5846 positive instances and 6095 negative instances (over 312 articles).

3.4. *Applying basic machine-learning classification*

The silver standard we created is designed to resemble the judgments which would be produced by a human without requiring an explicit annotation stage. The curators of CSA determined on the basis of a particular article whether a particular site was catalytic or not, which suggests that this information is available explicitly or implicitly in the text of the article. This in turn suggests that a machine-learning algorithm may be able to successfully classify some of the residue mentions on the basis of this textually-encoded information as catalytic.

In this section, we describe a fairly simplistic machine learning approach to this problem. This approach was designed to determine how readily the annotations could be determined using simple features based on the textual context surrounding the residue mention. In addition to this, it was desirable to have the features selected so the classifier could be trained on the relatively small abstracts-only portion of the corpus, since these are far more easily accessible than full text data. Because of the small size of this portion of the dataset, feature types which tend to suffer from data sparseness were not explored extensively. In particular, for word n-grams, from the few hundred instances in abstracts we have available there is unlikely to be enough information to meaningfully populate feature vectors for $n > 1$, so we only experimented with features based on word unigrams.

Table 1. **Statistics of PDB-PMID-Residue relationships in CSA**. PDB = Protein Data Bank. CSA = Catalytic Site Atlas. CSA-Lit = the subset of CSA annotations marked as based on literature. PMID = PubMed ID. A verified text residue is a residue that has been identified through text mining, and mapped to a physical residue in the corresponding PDB protein sequence. "Site" refers to a particular numbered location in a protein sequence.

Source	Set	Residues	PDB	PMIDs	(PMID,Site)
1. PDB	PDB residues, with abstract	17904740	30816	17595	4797110
2. PDB	PDB residues with verified text residues (abstracts)	44701	9923	5236	14127
3. PDB	PDB residues with verified text residues (full text)			7309	107153
4. CSA	PDB residues in CSA	112031	17524		
5. CSA	PDB residues in CSA, with abstract	94327	14673	7587	29447
6. CSA	Verified text residues; match to CSA (abstracts)	9059	3163	1630	2708
7. CSA-Lit	PDB residues in CSA-Lit	6372	942		
8. CSA-Lit	PDB residues in CSA-Lit, with abstract	5586	831	823	2799
9. CSA-Lit	PDB residues in CSA-Lit, with abstract with at least one verified text residue	2116	343	341	1139
10. CSA-Lit	Verified text residues; match to CSA-Lit (abstracts)	878	259	259	476
11. CSA-Lit	Verified text residues; match to CSA-Lit (full text)			312	805
12. CSA-Lit	Verified text residues; match to CSA-Lit (full text + abstract)			444	1052

We report classification results with a model built using Zhang Le's Maxent Toolkit[a]. In preliminary experiments, we achieved superior performance using this toolkit than with other tools. The corpora were preprocessed; sentences were identified using the Jena Sentence Boundary Detector tool[13] and the text was tokenized and tagged with part-of-speech (POS) tags using the GENIA tagger.[14] We experimented with the following feature sets:

- Tokens(b,e): the set of tokens in the range (b, e) relative to a physically-verified text residues (PVTR) token, not crossing sentence boundaries. For example, Tokens$(-2, -1)$ denotes the two tokens immediately preceding the PVTR. The value \$ denotes the sentence boundary, so Tokens$(\$, -1)$ means all preceding tokens.
- Lemma(b,e) and BioLem(b,e): the same as Tokens(b,e), but using *lemmas* of the tokens derived from the GENIA tagger and the BioLemmatizer,[15] respectively.
- MT: Match Type. Whether the PVTR was identified via mutation pattern such as *Cys42Ala* or a bare amino acid residue pattern such as *Cys42*.

Use of the full preceding sentential context, e.g. Lemma$(\$, -1)$, was found to be the most effective; smaller ranges tended to be detrimental. Features based on POS-tags were unhelpful. Unlemmatized tokens performed worse than lemmas (results not reported).

4. Results

4.1. *Literature-based recovery of CSA annotations*

To establish the context for the task of classifying functional sites as catalytic, we undertook an analysis of the data we had generated for our initial experiments with the LEAP-FS method. The aim of this analysis was to understand the proportion of the residues extracted from

[a] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

Table 2. Analysis of the overlap of the physically verified text residues (PVTR) in our full text corpus with functional site annotations. Percentiles in parentheses are relative to the category.

Category	Num PVTR	% PVTR	Num PMID	PMID %
All PVTR	107153	100.0%	7309	100.0%
in abstract	6085	(5.7%)	2477	(33.9%)
not in abstract	101068	(94.3%)	4832	(66.1%)
PVTR CSA-Lit	805	0.8%	312	4.3%
in abstract	237	(29.4%)	127	(40.7%)
not in abstract	568	(70.6%)	185	(59.3%)
PVTR any CSA	5821	5.4%	2413	33.0%
in abstract	1252	(21.5%)	759	(31.5%)
not in abstract	4569	(78.5%)	1654	(68.5%)
PVTR BindingMOAD	5652	5.3%	698	9.5%
in abstract	537	(9.5%)	239	(34.2%)
not in abstract	5115	(90.5%)	459	(65.8%)
PVTR NSM	42603	39.8%	5254	71.9%
in abstract	3540	(8.3%)	1653	(31.5%)
not in abstract	39063	(91.7%)	3601	(68.5%)
PVTR any annotation	44428	41.5%	5566	76.2%
in abstract	3900	(8.8%)	1804	(32.4%)
not in abstract	40528	(91.2%)	3762	(67.6%)

carefully selected biomedical publications that correspond to known catalytic sites. In our previous work,[3] we established that a significant proportion of the residues identified in the publications we analysed corresponded to functionally active sites as recorded in both CSA and BindingMOAD and used this as evidence supporting the hypothesis that residue mentions in the literature have functional significance. Here, we specifically examine how well we are able to recover the functional site annotations in the CSA, essentially measuring the method's recall of curated annotations.

Table 1 summarises the results. Line 1 indicates the total amount of information in the Protein Data Bank that is linked to a PubMed ID. The last column (PMID,Site) indicates the number of unique combinations of PMIDs and residue locations in the Protein Data Bank. This represents an upper bound on the number of residues mentioned in text that we would expect to find. Line 2 represents the results of the analysis in our previous work; we identified 14,127 residue mentions in 5,236 PubMed abstracts; those residue mentions correspond to 44,701 physical residues in the PDB. Line 3 extends that to processing of full text publications. Lines 4-6 focus on the subset of PDB residues that are included in the Catalytic Site Atlas; we take these residues to be the set of known (or presumed) catalytic sites. We can see here that text mining of PubMed abstracts (Line 6) is only able to identify a small proportion ($\sim$9%) of the catalytic residues that could be mentioned in some publication (Line 5). Restricting our analysis to those residues in the CSA that have been explicitly marked as having supporting evidence in the literature (the CSA-Lit subset; Lines 7-12) we find that we are able to recover a somewhat higher proportion, $\sim$17% for abstracts (476/2799). When we process full text as well, our coverage of CSA-Lit improves dramatically, to $\sim$38% (1052/2799) for all literature evidence we were able to access and detect.

Table 3. Results using 8-fold cross-validation over the development and *test* sets. BL = Baseline; ME refers to the MaxEnt classification engine. A = abstracts; F = full text.

			Sections		Catalytic	Non-catalytic	F-score
Eng.	σ^2	Features	Train	Test	P / R / F	P / R / F	Mic / Mac
ME	0.0	MT, Lemma	A	A	**86.8**/ 93.0 / 89.8	51.7 / **34.6**/**41.4**	81.5 / **66.4**
ME	1.0	MT, Lemma	A	A	82.6 / **97.1** / 89.3	30.8 / 6.0 / 10.1	76.9 / 54.0
ME	4.0	MT, Lemma	A	A	85.6 / 95.8 /**90.4**	**56.7**/ 25.6 / 35.2	**81.8**/ 65.5
ME	0.0	MT, BioLem	A	A-*test*	*79.2 / 86.9 / 82.9*	*17.4 / 10.8 / 13.3*	*69.0 / 48.6*
BL			A, F	A	82.2 /**100.0**/**90.2**	0.0 / 0.0 / 0.0	74.1 / 45.1
ME	1.0	BioLem	A, F	A	**90.4**/ **81.2** / 85.6	**41.0**/ 60.2 /**48.8**	**79.5**/**68.1**
ME	4.0	MT, Lemma	A, F	A	90.0 / 74.6 / 81.5	34.5 /**61.7**/ 44.2	76.0 / 65.0
ME	0.0	MT, BioLem	A, F	A-*test*	*89.2 / 80.0 / 84.4*	*44.2 / 62.2 / 51.7*	*78.2 / 68.8*
BL			A	F	48.9 /**100.0**/ 65.7	0.0 / 0.0 / 0.0	32.1 / 32.8
ME	0.0	Lemma	A	F	51.9 / 87.2 / 65.1	64.8 /**22.5**/ 33.4	56.2 / 56.5
ME	0.0	MT, BioLem	A	F	**52.4**/ 89.4 / 66.1	68.8 / 22.2 /**33.6**	**57.8**/**58.1**
ME	1.0	MT, BioLem	A	F	50.6 / **96.4** /**66.3**	**73.9**/ 9.7 / 17.2	56.8 / 57.3
ME	0.0	MT, BioLem	A	F-*test*	*51.7 / 88.8 / 65.3*	*64.8 / 19.9 / 30.5*	*56.0 / 56.2*
BL			A, F	F	48.9 /**100.0**/ 65.7	0.0 / 0.0 / 0.0	32.1 / 32.8
ME	1.0	Lemma	A, F	F	**63.5**/ 72.6 /**67.7**	69.6 / 60.0 / 64.5	**66.4**/**66.4**
ME	0.0	BioLem	A, F	F	62.4 / **72.7** / 67.2	69.0 / 58.1 / 63.0	65.5 / 65.5
ME	1.0	BioLem	A, F	F	62.7 / 73.6 /**67.7**	**69.7**/ 58.1 / 63.3	66.0 / 66.0
ME	4.0	MT, Lemma	A, F	F	**63.5**/ 60.9 / 62.2	64.0 /**66.5**/**65.2**	63.8 / 63.7
ME	0.0	MT, BioLem	A, F	F-*test*	*69.2 / 68.4 / 68.8*	*69.9 / 70.7 / 70.3*	*69.6 / 69.5*

We further wish to understand the extent of the ambiguity we face in attempting the classification task. To assess this, we examined the annotation status of the physically verified text residues (PVTRs) in the full text data set. While the annotations of the CSA-Lit subset are clearly the most relevant to our classification task, representing literature-curated catalytic site annotations, all of the annotations in CSA are very likely to be valid catalytic sites, as they were derived through alignment with known catalytic sites in closely related structures. As we suggested above, many binding sites are also catalytic sites. Hence sites in PDB protein structures which are in close proximity with a small molecule (NSM = near small molecule), a characteristic strongly suggestive of a ligand binding site, as well as the curated subset of those sites represented in the BindingMOAD database,[5] are also potentially catalytic. The overlap of the PVTRs with each of these sources is summarized in Table 2. The results show that a large proportion of the PVTRs overlap with some existing annotation for those sites (41.5%), despite only a small fraction having been formally curated as catalytic sites (0.8%). While this is a strong result for LEAP-FS – supporting the hypothesis that text residues are likely to be functionally important – it means that we have a large ambiguity set for our catalytic site classification task.

4.2. *Classification results over silver corpus*

The abstracts and full text corpora were split into a training subset and a test subset, with 80% of the articles in each corpus randomly selected for training and 20% reserved as a held-out test set. Table 3 provides a selection of the classification results over these subsets (test results in italics). Our experiments used 8-fold cross-validation. We include results for

Table 4. Results using 8-fold cross-validation over the development and *test* sets, aggregated by unique physically-verified text residue. See caption Table 3 for abbreviation definitions.

			Sections		Catalytic	Non-catalytic	F-score
Eng.	σ^2	Features	Train	Test	P / R / F	P / R / F	Mic / Mac
BL			A	A	78.9 / **100.0** / 88.2	0.0 / 0.0 / 0.0	69.6 / 44.1
ME	0.0	MT, Lemma	A	A	**82.5** / 96.7 / **89.1**	65.8 / **23.6** / **34.7**	**80.1** / **66.4**
ME	4.0	MT, Lemma	A	A	80.9 / 98.5 / 88.8	**70.0** / 13.2 / 22.2	79.5 / 64.2
ME	1.0	MT, BioLem	A	A	79.7 / **99.0** / 88.3	60.0 / 5.7 / 10.3	77.4 / 59.8
ME	0.0	MT, BioLem	A	A-*test*	*73.3 / 92.8 / 81.9*	*33.3 / 9.7 / 15.0*	*66.1 / 52.3*
BL			A, F	A	78.9 / **100.0** / **88.2**	0.0 / 0.0 / 0.0	69.6 / 44.1
ME	0.0	Lemma	A, F	A	87.9 / **85.9** / 86.8	51.3 / 55.7 / 53.4	79.8 / 70.2
ME	4.0	BioLem	A, F	A	**88.7** / 85.1 / **86.9**	**51.6** / 59.4 / **55.3**	**80.3** / **71.2**
ME	0.0	MT, Lemma	A, F	A	88.5 / 77.8 / 82.8	42.9 / **62.3** / 50.8	76.6 / 67.8
ME	0.0	MT, BioLem	A, F	A-*test*	*85.9 / 80.7 / 83.2*	*55.6 / 64.5 / 59.7*	*77.0 / 71.7*
BL			A	F	19.3 / **100.0** / 32.3	0.0 / 0.0 / 0.0	6.2 / 16.2
ME	0.0	Lemma	A	F	22.6 / 94.5 / 36.5	94.6 / **22.9** / **36.9**	50.5 / 58.7
ME	0.0	MT, BioLem	A	F	**22.8** / 95.8 / **36.9**	95.8 / 22.7 / 36.8	**50.8** / **59.3**
ME	1.0	MT, BioLem	A	F	20.8 / **99.7** / 34.5	**99.2** / 9.6 / 17.5	40.8 / 57.2
ME	0.0	MT, BioLem	A	F-*test*	*24.3 / 94.1 / 38.6*	*92.1 / 19.1 / 31.6*	*48.5 / 57.4*
BL			A, F	F	19.3 / **100.0** / 32.3	0.0 / 0.0 / 0.0	6.2 / 16.2
ME	1.0	Lemma	A, F	F	**36.7** / 78.6 / **50.1**	93.0 / 67.6 / 78.3	**75.5** / **68.7**
ME	1.0	BioLem	A, F	F	35.2 / **80.1** / 48.9	**93.2** / 64.7 / 76.4	74.2 / 68.0
ME	4.0	MT, Lemma	A, F	F	**36.7** / 65.2 / 47.0	89.8 / **73.2** / **80.6**	75.4 / 66.1
ME	0.0	MT, BioLem	A, F	F-*test*	*50.4 / 75.7 / 60.5*	*92.2 / 79.4 / 85.3*	*80.8 / 74.3*

a baseline system, labelled BL, which is a majority-class classifier. Other lower-performing scenarios are not included. All results reported for Lemma and BioLem are Lemma($\$, -1$) and BioLem($\$, -1$), respectively. The column σ^2 indicates the value for the σ^2 Gaussian smoothing parameter to the MaxEnt learner (0.0, 1.0, 4.0 were tested).

The abstract development set contains 613 catalytic (positive) and 133 non-catalytic (negative) text instances, while the full text development set contains 4641 catalytic and 4846 non-catalytic text instances. The standard measurements of precision, recall, and f-score are calculated over these text instances. F-score is calculated through micro-averaging (Mic), i.e. across all text residues in the test set, and macro-averaging (Mac), i.e. averaging performance over the two categories catalytic vs. non-catalytic.

We note that the baseline system BL has non-zero values for the non-majority class (therefore, less than 100% majority class recall) in the cases of training on abstracts and full text together. This is because the majority class is calculated from the input data in the training fold. The folds are randomly determined at the document level rather than the text instance level. Coupled with more balanced instances in the full text data, this can result in the majority class in a given fold not being the same as the majority class in the entire data set.

5. Discussion

5.1. *Full text versus abstracts*

Examination of Table 1 shows a clear advantage when processing full publications as compared to abstracts, despite having access to a smaller proportion of the relevant literature (66% of relevant publications). This demonstrates that the increased level of detail that is available in full text publications[16] is important for understanding of specific physical residues. Our results also indicate an advantage in processing both the abstracts and full text together.

Our results show some discrepancies between what is identified in abstracts as compared to the corresponding full text publications we were able to access. While it would be typical for a full text publication to contain the abstract, we found that our system was not able to identify (minimally) the same text residues as in the corresponding abstracts for 497 full text files. Further investigation revealed that more than half (250) of those full text publications were spurious – while there was text downloaded for a given PMID, it was not the actual publication context. This typically resulted from an error in the logic of our full text retrieval script or a subscription firewall. We found that in 157 publications we missed residue mentions due to conventions in the HTML to plain text conversion script that our residue detection patterns were not sensitive to. An additional 88 publications had no results in the full text data set because single letter mutations were not included in the full text processing.

5.2. *Classifier performance*

Examination of Table 3 shows several consistent patterns. First, the classifiers based on machine learning all easily outperform the baseline classifier; this effect is most pronounced on the more balanced full text test set. Second, the classifiers trained on more data (combining both abstracts and full text) outperform the classifiers trained on abstracts alone. The lack of complete subsumption of the abstract data in the full text data, as discussed above, likely contributes to this effect, but it also demonstrates the advantage of more training contexts to learn from. Third, the MT (match type) feature improves performance in most cases. Fourth, the results on the held-out test set are slightly lower than the corresponding results for the development set, except in the case of the final full text test run, trained on all the development data. This again shows the benefit of more data. However, the differences between the various feature sets we experimented with were small and not fully consistent across the system combinations we considered – sometimes BioLem gives an advantage over Lemma and sometimes not, and various settings for the σ^2 parameter affected P/R/F across the two categories inconsistently. We have experimented with a limited set of features in this work to test the viability of the approach; application of other features and other approaches to named entity recognition is warranted to achieve improved performance.

One complicating factor for the classifier arises from the distinction between a catalytic *text mention* of a given site in a protein, and a catalytic site. A catalytic site may be discussed in text for some reason that has nothing to do with its function and therefore a given text mention may not be appropriately categorized as "catalytic", even if the corresponding protein site is a catalytic site. However, given our distance-based methodology for producing the training and test data for the classifier, we cannot discriminate between these cases. We annotate individual text mentions of PVTRs based on site-level information rather than considering whether the specific local textual context provides evidence of function.

An analysis of the classifier's performance at the level of a unique PVTR, rather than at the level of text mentions, is shown in Table 4. Here, we have aggregated over the classification of all text mention instances of a given (PMID, Site) pair. We have employed a simple majority vote of classifications over the instances – that is, if the majority of the individual text mentions are classified as catalytic, then the PVTR is classified as catalytic as well. In the case of a tie, we examine the scores of the classification. When the data is viewed this way, we see improved performance on the recall of catalytic sites, at a significant cost to precision. In contrast, the classification of non-catalytic sites has improved overall. These results therefore confirm the catalytic text mention/catalytic site difference, suggesting that many text mentions of catalytic sites are not clear references to its catalytic status, while it is possible to reliably rule a PVTR out as non-catalytic due to a lack of catalytic text mentions.

Addressing this problem could lead to overall improvement of the classifier, which we plan to explore in future work. We could build a classifier which aggregates information across text mentions to support classification of a unique PVTR rather than classification of each individual text mention. We could also explore a two-part solution, where the first part is to identify sentences that contain functional information about a site in a protein, and the second part is to classify that functional information more specifically. This would require development of a training corpus which provided more specific functional information. This could be done manually or by filtering text mentions in our existing corpus according to whether there is a detected Gene Ontology molecular function term within the same sentence, similar to the mutation grounding strategy of Naderi and Witte.[8]

5.3. *Use of the classifier for improving curation of catalytic sites*

Our data highlights (a) the low coverage of curated information about both catalytic sites and binding sites more generally, and (b) the significant ambiguity of functional sites.

The gap between curated information and the amount of inferred information in genomic databases is a well known problem,[17] and we see clear evidence of that gap here. In comparing Lines 4 and 7 of Table 1, we see that the CSA-Lit curated subset represents less than 6% of the full CSA database. BindingMOAD is the curated subset of the PDB NSM data, focused on protein ligand binding sites. For the PVTRs in our corpus, only 13% of NSM sites we recover are also captured in BindingMOAD (PVTR NSM vs. PVTR BindingMOAD data in 2). While this difference in coverage could be because many of those NSM sites are not high-quality binding sites, it is more likely a reflection of the time and resources that manual curation requires. The fact that we are able to recover over forty thousand NSM sites in our corpus of 7,309 full text publications suggest that text mining can play a powerful role in closing this gap, by highlighting sites that have literature evidence of functional importance.

However, to close the annotation gap we must go one step further than identification, and perform finer-grained categorization of those PVTRs. The catalytic site classifier we have developed on the basis of our training data could be applied more broadly to the full set of 7,309 articles for which we have identified PVTRs. This would identify those PVTRs not in CSA-Lit that are most likely to be catalytic sites; those in turn can be prioritized for curation, and the specific document with the catalytic mention of the site can be provided to a database curator. These developments would enable higher-throughput in the curation process.

6. Conclusion

This paper has explored the applicability of text mining from the biomedical literature to the problem of detecting catalytic sites. We have presented two corpora in which protein residue mentions were annotated using reliable external knowledge about catalytic residues. Our analysis of these corpora according to their coverage of existing annotated resources showed that the literature is a good source of information about functionally significant protein sites, and furthermore that processing of full text publications is particularly important for achieving good recall of these sites from the literature. With respect to classification of these functional sites as catalytic, we observed that there is considerable ambiguity in assigning the functional role of a given site.

Nevertheless, we explored development of a classifier learned from our annotated silver corpora to enable automatic annotation of catalytic sites in the biomedical literature. Despite the ambiguity of catalytic sites, and the evaluation of the annotation at the level of individual text mentions of protein sites rather than aggregated over unique physical sites, the classifiers were able to achieve reasonably good performance with a simple set of features. Having established the viability of the approach, and having identified some of the challenges that arise in this task, we are confident that we in future work we will be able to develop new methods that improve upon the initial results presented here. This work represents an important step in the development of effective strategies for understanding functional characteristics of proteins at the level of specific residues, and for supporting curation of that information in databases, by exploiting the information available in the published literature.

References

1. S.-A. Marashi, *EXCLI Journal* **4**, 87 (2005).
2. Y. Ofran, M. Punta, R. Schneider and B. Rost, *Drug Discov Today* **10**, 1475 (2005).
3. K. M. Verspoor, J. D. Cohn, K. Ravikumar and M. E. Wall, *PLoS One* **7**, e32171 (2012).
4. C. T. Porter, G. J. Bartlett and J. M. Thornton, *Nucleic Acids Research* **32**, D129 (2004).
5. M. L. Benson and et al, *New Mathematics and Natural Computation* **06**, 49 (2010).
6. H. Cunningham, D. Maynard, K. Bontcheva and et al, *Text Processing with GATE (V. 6)* 2011.
7. M. L. Benson, R. D. Smith, N. A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, J. Nerothin and H. A. Carlson, *Nucleic Acids Research* **36**, D674 (2008).
8. N. Naderi and R. Witte, *BMC Genomics* **13**, S20 (2012).
9. The Gene Ontology Consortium, *Nat Genet* **25**, 25 (2000).
10. K. Nagel, A. Jimeno-Yepes and D. Rebholz-Schuhmann, *BMC Bioinf* **10 Suppl 8**, S4 (2009).
11. K. Ravikumar, H. Liu, J. D. Cohn, M. E. Wall and K. M. Verspoor, *J Biomed Sem* **3**, S2 (2012).
12. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov and P. Bourne, *Nucleic Acids Research* **28**, 235 (2000).
13. K. Tomanek, J. Wermter and U. Hahn, A reappraisal of sentence and token splitting for life science documents, in *POT 12th MEDINFO*, (IOS Press, 2007).
14. Y. Tsuruoka, Y. Tateishi, J. D. Kim, T. Ohta, J. Mcnaught, S. Ananiadou and J. Tsujii, Developing a robust part-of-speech tagger for biomedical text, in *POT 10th Panhellenic Conf on Informatics*, 2005.
15. H. Liu, T. Christiansen, W. Baumgartner and K. Verspoor, *J Biomed Sem* **3**, 3 (2012).
16. K. B. Cohen, H. L. Johnson, K. Verspoor, C. Roeder and L. Hunter, *BMC Bioinf* **11**, 1 (2010).
17. W. Baumgartner, K. Cohen, L. Fox, G. Acquaah-Mensah and L. Hunter, *Bioinf* **23**, i41 (2007).

MODELING CELL HETEROGENEITY: FROM SINGLE-CELL VARIATIONS TO MIXED CELLS

ERIC BATCHELOR
Center for Cancer Research, National Cancer Institute, National Institutes of Health
Bethesda, MD 20894, USA
Email: eric.batchelor@nih.gov

MARICEL G. KANN
Department of Biological Sciences, University of Maryland, Baltimore County
Baltimore, MD, 21250, USA
Email: mkann@umbc.edu

TERESA M. PRZYTYCKA
National Center for Biotechnology Information, National Institutes of Health
Bethesda, MD 20894, USA
Email: przytyck@ncbi.nlm.nih.gov

BENJAMIN J. RAPHAEL
Department of Computer Science, Brown University
Providence, RI 02912, USA
Email: braphael@cs.brown.edu

DAMIAN WOJTOWICZ
National Center for Biotechnology Information, National Institutes of Health
Bethesda, MD 20894, USA
Email: wojtowda@ncbi.nlm.nih.gov

Emerging technologies such as single cell gene expression analysis and single cell genome sequencing provide an unprecedented opportunity to quantitatively probe biological interactions at the single cell level. This new level of insight has begun to reveal a more accurate picture of cellular behavior, and to highlight the importance of understanding cellular variation in a wide range of biological contexts. The aim of this workshop is to bring together researchers working on identifying and modeling cell heterogeneity that arises by a variety of mechanisms, including but not limited to cell-to-cell noise, cell-state switches and cell differentiation, heterogeneity in immune responses, cancer evolution, and heterogeneity in disease progression.

1. Background

Quantifying the molecular mechanisms underlying cell behaviors and functions is one of the ultimate goals of biology and medicine. Until recently, most current measures to classify and characterize cellular behavior have been performed on the average of all cells in a sample instead of a single cell. However, measurements derived from pooled populations of cells lack the specificity to capture outlier cell behavior that might explain cell differentiation and transitions from normal to disease cellular states. The noise, or variance, between the genomic state of different cells -- even among cells assumed to be homogenous -- has been shown to be strongly correlated with protein expression and function [1]. Furthermore, it has been argued that neglecting cell heterogeneity is one of the major causes of error in disease classification [2]. Emergence of cell heterogeneity might be sporadic (e.g., cell-to-cell variation in an isogenic cell population [3]), programmed (e.g., cell differentiation [4]), or, to some extent, driven by selection pressure (e.g., in cancer [5]).

Emerging technologies like single cell gene expression and single cell sequencing provide unprecedented opportunities to quantify single cell level differences. These technologies will be able to provide a wealth of various forms of new information including protein abundance, methylation patterns, promoter structure, gene expression, copy number variations, gene function and essentiality, DNA structure, evolutionary plasticity, and selective advantage. These data can all be leveraged in the quest to understand the emergence and consequences of heterogeneity. However, simple accrual of data from these various single cell experimental techniques is not enough to obtain a clear understanding of the diverse range of biological processes affected by cellular heterogeneity. Synthesis and interpretation of the wealth of single cell-level data depends on novel computational approaches to uncover and model the biological principles that underlie the emergence of cell heterogeneity. Most importantly, computational methods are needed to provide a system-level view of the interplay of diverse, fluctuating biological components.

The increasing interest in the mechanisms underlying diseases is complemented by the emergence of single cell technologies. Due to these factors, we expect increasing efforts to develop computational models and tools for the analysis of cell heterogeneity. We believe that this workshop gives the community of computational biologists the chance to discuss this theme, which may soon become dominant in biomedical science.

2. Main directions and challenges

The focus of this workshop is on uncovering and modeling cell heterogeneity that arises by any of the above-mentioned mechanisms – sporadically, programmed, and through evolution. The main topics covered by this workshop are questions related to cell-to-cell noise, cell-state switches and cell differentiation, heterogeneity in immune responses, cancer evolution, and disease progression and heterogeneity.

2.1. *Cell-to-cell noise*

The stochastic nature of gene expression leads to cell-to-cell differences in protein level, commonly referred to as noise. Expression noise can be disadvantageous, by affecting the precision of performing biological functions, but it may also be advantageous by enabling heterogeneous stress-response programs to environmental changes [6]. Therefore, various genes and gene groups might display various levels of expression noise. Importantly, gene expression is a multi-step process and the stochasticity of its individual steps, including transcription and translation, contributes to the resulting variability. Untangling different components of expression noise is highly nontrivial and requires a concerted effort of experiment and computational modeling.

2.2. *Cell state switches and cell differentiation*

Heterogeneity has profound implications for cellular differentiation, in which cells must commit to one of a finite number of possible cell states. It has long been appreciated that cells must have molecular mechanisms for counteracting fluctuations in both environmental conditions and cellular components to reliably affect developmental programs [7]. However, recent work suggests that utilization of heterogeneity in the activation of cell differentiation programs in a population of cells can be evolutionarily advantageous. Such advantages are being found in a broad range of biological systems. For example, combined experimental and computational work in *B. subtilis* has begun to characterize the benefits of cell-to-cell variability for the survival of prokaryote populations [4,8]. The importance of heterogeneity in differentiation programs in higher organisms is also being highlighted by recent advances in single cell technologies. New data from analysis of both embryonic stem cells (e.g., [9]) and adult stem cells (e.g., [10]) highlight the necessity for novel computational approaches to provide a deeper understanding of the role of heterogeneity in these important areas of biology and medicine.

2.3. *Heterogeneity in immune responses*

Innate and adaptive immune responses depend on the proper utilization and regulation of cellular heterogeneity. Recognition of a wide variety of antigens necessitates a heterogeneous population of immune cells. However, if the heterogeneity is too great, discrimination between “self” and “other” antigens may be compromised leading to autoimmune diseases. Advances in flow cytometry and single cell proteomics are beginning to elucidate the molecular mechanisms governing the proper regulation on this heterogeneity [11]. Complementary approaches in computational methods to analyze this important aspect of cellular immunity will be key to further our understanding of immune responses in healthy and disease states [12].

2.4. *Cancer evolution*

A tumor is formed from a heterogeneous mass of cells with different complements of somatic mutations and possibly different differentiated states. The implications of this heterogeneity for cancer progression and treatment are not well understood. While mutational heterogeneity could merely be a consequence of the somatic mutation process that drives cancer development, heterogeneity may itself be an important, or even essential, contributor to tumor evolution [5,13,14]. Single-cell analyses are necessary to understand the role heterogeneity plays in cancer evolution. For instance, single-cell sequencing is likely to profoundly impact cancer diagnostics and prognosis through the detection of rare tumor cells or through the monitoring of circulating tumor cells [15]. Single-cell next generation sequencing can also be used to investigate tumor subpopulations and to delineate the differences between primary and metastasic tumors [5,14].

3. Workshop Contributions

The workshop includes two invited speakers and five accepted submissions.

Dana Pe'er (invited speaker) is an Associate Professor in the Department of Computer Science and the Department of Biological Sciences at Columbia University. Dr. Pe'er received her doctoral degree in computational biology from the Hebrew University. Her research focuses on understanding the organization, function and evolution of molecular networks. Dr. Pe'er and her team develop computational methods to integrate diverse high-throughput genomic data and to discover the general principles governing cellular signal processing, propagation of small changes in regulatory networks and how they alter cellular functioning that can lead to diseases such as autoimmune disease and cancer.

Sylvia Plevritis (invited speaker) is an Associate Professor in the Department of Radiology in the Stanford School of Medicine. Dr. Plevritis received her doctoral degree in Electrical Engineering and master's degree in health services from Stanford University. Her research focuses on computational and mathematical modeling of cancer biology and cancer outcomes. Using diverse sources of information from genomic and proteomic data to clinical data, her laboratory was able to infer natural histories of cancer, to estimate cell subpopulations after cancer treatment, and to identify perturbations of molecular networks during different stages of cancer. Dr. Plevritis is also the Program Director of the Stanford Center for Cancer Systems Biology (CCSB), and the co-Section Chief of Information Sciences in Imaging at Stanford (ISIS).

Julian Candia, Jayanth Banavar and Wolfgang Losert. "From molecules to cells to organisms: understanding health and disease with multidimensional single-cell methods." This works describes a newly developed framework to investigate multicolor data from fluorescence-activated cell sorting (FACS). The method integrates several approaches to gain different perspectives on the data: singular value decomposition to reduce data representation, machine learning to separate patients into classes and improve diagnosis, and network analysis to infer cell subpopulations.

Michael Januszyk, Jason P. Glotzbach, Michael Sorkin, Atul J. Butte and Geoffrey C. Gurtner. "Automated Functional Profiling of Progenitor Cell Populations using High-Resolution Single Cell Gene Expression Data." The authors show how information from thousands of publicly available microarray datasets of gene expression can be used to increase the power of single cell gene expression data analysis, which has high-resolution but is limited to only several dozen target genes that can be measured at the same time. The method is based on higher-order covariance of gene expression retrieved from the Gene Expression Omnibus (GEO) database, and functional classification based on Gene Ontology. Applied to murine bone marrow-derived mesenchymal stem cells, the method finds significant associations between cell subpopulations and known functional categories.

Layla Oesper, Ahmad Mahmoody and Ben Raphael. "Estimating Tumor Clonal Populations from Copy Number Data." The authors introduce an algorithm to infer tumor subpopulations directly from high-throughput DNA sequencing data. Their method decomposes a mixture of normal cells, clonal cells, and sub-clonal populations to maximize the probability of observed data using techniques from convex optimization. The algorithm was applied to nine breast cancer samples and successfully recovered the heterogeneity of these tumor cell populations.

Kyoungmin Roh and Stephen Proulx. "The role of positive and negative feedback loops of p53 pathway." This work describes simulations of different scenarios built from deterministic p53 feedback loops, both negative and positive, based on Puszynski's model. Using simulated annealing to find the optimal response of p53 to DNA damage, the authors demonstrate the ability of p53 feedback loops to reduce the chance of cell apoptosis, making the cell less sensitive to DNA damage. The analysis provides new insight on p53 feedback loops and DNA damage pathways.

Damian Wojtowicz, Daniela Ganelin, Raheleh Salari, Jie Zheng, David Lavens, Yitzhak Pilpel and Teresa Przytycka. "Teasing apart sources of stochastic variations in eukaryotic gene expression." In this project, the authors develop a novel computational approach to delineate the relative impact of transcription and translation processes on cell-to-cell variations, noise, in protein abundance, and apply it to large-scale gene expression data from yeast (Newman et al., 2006). Interestingly, they show that translation-related genomic features, such as codon usage and 5'UTR secondary structure, have higher impact on noise than previously appreciated.

4. Acknowledgments

We would like to thank Dana Pe'er and Sylvia Plevritis for kindly accepting our invitation and generously supporting our workshop. M.G.K was supported by the National Institutes of Health (NIH), 1K22CA143148. E.B. was supported in part by Intramural Research Program of the NIH, NCI, Center for Cancer Research. T.M.P. and D.W. research was supported in part by the Intramural Research Program of the NIH, National Library of Medicine. B.J.R. is supported by the National Institutes of Health grant R01HG5690, a CAREER Award from the National Science Foundation (CCF-1053753), and Burroughs Wellcome Career Award at the Scientific Interface.

References

1. Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, et al. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature 441: 840-846.
2. Marko NF, Quackenbush J, Weil RJ (2011) Why is there a lack of consensus on molecular subgroups of glioblastoma? Understanding the nature of biological and statistical variability in glioblastoma expression data. PLoS One 6: e20826.
3. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK (2009) Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature 459: 428-432.
4. Suel GM, Kulkarni RP, Dworkin J, Garcia-Ojalvo J, Elowitz MB (2007) Tunability and noise dependence in differentiation dynamics. Science 315: 1716-1719.
5. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature 472: 90-94.
6. Eldar A, Elowitz MB (2010) Functional roles for noise in genetic circuits. Nature 467: 167-173.
7. Waddington CH (1942) Canalization of development and the inheritance of acquired characters. Nature 150: 563-565.

8. Kuchina A, Espinar L, Garcia-Ojalvo J, Suel GM (2011) Reversible and noisy progression towards a commitment point enables adaptable and reliable cellular decision-making. PLoS Comput Biol 7: e1002273.
9. Phanstiel DH, Brumbaugh J, Wenger CD, Tian S, Probasco MD, et al. (2011) Proteomic and phosphoproteomic comparison of human ES and iPS cells. Nat Methods 8: 821-827.
10. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, et al. (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell 144: 296-309.
11. Feinerman O, Jentsch G, Tkach KE, Coward JW, Hathorn MM, et al. (2010) Single-cell quantification of IL-2 response by effector and regulatory T cells reveals critical plasticity in immune response. Molecular systems biology 6: 437.
12. Coward J, Germain RN, Altan-Bonnet G (2010) Perspectives for computer modeling in the study of T cell activation. Cold Spring Harb Perspect Biol 2: a005538.
13. Michor F, Polyak K (2010) The origins and implications of intratumor heterogeneity. Cancer Prev Res (Phila) 3: 1361-1364.
14. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, et al. (2010) Inferring tumor progression from genomic heterogeneity. Genome Res 20: 68-80.
15. Navin N, Hicks J (2011) Future medical applications of single-cell sequencing in cancer. Genome Med 3: 31.

COMPUTATIONAL BIOLOGY IN THE CLOUD: METHODS AND NEW INSIGHTS FROM COMPUTING AT SCALE

PETER M. KASSON

Departments of Molecular Physiology and Biomedical Engineerng,
University of Virginia, Box 800886
Charlottesville, VA 22908, USA

The past few years have seen both explosions in the size of biological data sets and the proliferation of new, highly flexible on-demand computing capabilities. The sheer amount of information available from genomic and metagenomic sequencing, high-throughput proteomics, experimental and simulation datasets on molecular structure and dynamics affords an opportunity for greatly expanded insight, but it creates new challenges of scale for computation, storage, and interpretation of petascale data. Cloud computing resources have the potential to help solve these problems by offering a utility model of computing and storage: near-unlimited capacity, the ability to burst usage, and cheap and flexible payment models. Effective use of cloud computing on large biological datasets requires dealing with non-trivial problems of scale and robustness, since performance-limiting factors can change substantially when a dataset grows by a factor of 10,000 or more. New computing paradigms are thus often needed. The use of cloud platforms also creates new opportunities to share data, reduce duplication, and to provide easy reproducibility by making the datasets and computational methods easily available.

1. Challenges and opportunities of massive data

In recent years, large-scale datasets have become increasingly common in many biological fields. There have been tremendous strides in the throughput capacity and affordability of genomic sequencing. RNAi and similar techniques have allowed broad surveys of genetic regulation and host-pathogen interaction [1-3]. Multiple techniques for protein-protein association have gone large-scale, leading to "interactome" scale analyses of HIV infection [4] and other processes. "Brain atlas" projects are making available datasets that combine gene expression profiles with detailed anatomic and localization data [5]. And structural genomics projects have steadily increased the number of high-resolution macromolecular structures available for the proteins involved in all these processes. In addition to statistical analysis of all these datasets, more compute-intensive approaches such as large-scale simulation have made it possible to simulate many mutants of a drug target or combine data sources to examine the structure and dynamics of large subcellular structures. All these advances offer the possibility for tremendous insight into biology but pose challenges for effective analysis to maximize this insight.

The particulars involved in analyzing each of these domain-specific datasets have been treated elsewhere; we will discuss some of the common themes, particularly as they relate to cloud computing. Dataset size has increased greatly. Simply transmitting and storing many sequencing datasets is non-trivial. Sharing access to these data is yet more complicated when they are stored on individual researchers' or centers' computer systems. The challenges of sharing are compounded when patient data are concerned and access should be restricted and monitored. Analyzing these large datasets can also be very computationally intensive. Most analyses are at best case O(N); anything that leverages the comprehensive nature of large-scale datasets to examine pairwise or higher-order association often scales substantially worse. So the "classical" paradigms of storing datasets on local clusters and running analysis there are challenged three times: by transfer and archival capacity, by storage capacity, and by compute capacity.

2. Cloud solutions for problems of scale

We will briefly discuss how cloud-computing paradigms offer new solutions for these challenges; the workshop will illustrate a number of these as well as give an opportunity to discuss challenges and future directions. This subject has also been reviewed in [6-8] and elsewhere.

2.1. *Flexible capacity: the utility model of computing*

One substantial advantage of cloud-computing solutions is the ability to adjust computational resources according to requirements. Doubling the size of a traditional cluster is both expensive and time-consuming; in a cloud model, it simply involves requesting (and paying for) twice the capacity. The capacity limit of services such as Amazon's EC2 and Google Compute Engine has not been made public, but a recent demonstration showed the Institute for Systems Biology Genome Explorer running on 600,000 cores (Google IO 2012). In terms of raw compute capacity, this alone would likely rank among the top 5 supercomputers in the world. An additional advantage for cloud computing is the ability to dynamically adjust utilization. Most analyses do not run at a constant rate over time—one would like to request a large amount of resources to run a calculation and then release those resources while the results are evaluated and the next analysis is designed. It is thus rare to see a cluster at 100% utilization all the time, and like spare airline seat capacity or spare hotel rooms, the spare cycles are wasted money. In current cloud paradigms, a user pays only for the jobs (or virtual machines) that are running. This provides a much better fit to a fluctuating usage pattern.

2.2. *Storage*

Cloud storage solutions such as Amazon's S3 or Google Cloud Storage offer similar scalability and flexibility to the matching compute solutions. More importantly, they allow large and potentially shared datasets to be stored on the same infrastructure where large-scale analyses are run. This can obviously be achieved if one has a copy of a dataset on one's local cluster, but such an approach quickly becomes redundant when the dataset is held in common to many disparate users—the NCBI Short Read Archive is a good example. Caching a copy of this dataset on each cluster where analysis is run quickly becomes an expensive and redundant exercise. Companies such as DNAnexus have utilized cloud resources to offer storage, access to shared datasets, and transparent sharing of data. Cloud storage also provides enhanced reliability, as the data are backed up in several geographical locations.

2.3. *Parallel analyses*

When both computation and storage are performed in large distributed data centers, one can leverage technologies such as MapReduce[9, 10] and Dremel[11] for performing analyses and database queries in a much more efficient manner.

2.4. *Sharing data, tools, and algorithms*

One important and oft-overlooked benefit of the cloud model is how it can facilitate sharing. Most obviously, cloud data storage allows easy sharing with access control lists and monitoring of access for sensitive data. However, the ability to package and distribute tools and analyses on cloud platforms offers a new transparency in tool sharing and reproducibility. Furthermore, it helps overcome the problem of web server tools that are often overloaded or impose an undue burden on the host resources. If one imagines an analysis program running on a cloud front-end (such as Google App Engine) where any user can design a job to access a shared or private dataset and be presented with a virtual machine to run on his or her account for a common cloud service provider (Amazon EC2, Google Compute Engine, Microsoft Azure, or other), this allows much greater scalability for any public service and also reduces the cost to an individual researcher of making his or her methods publicly accessible.

3. Challenges in the cloud

Cloud computing offers new computational paradigms to deal with data and analyses at scale. However, simply applying the same algorithms and programming paradigms on 1000x the data often yields poor results. Working at scale generates different limiting factors on performance and cost, both algorithmic and logistical (data locality and transfer speed/cost, network latency, virtual machine start-up). Paradigms and tools such as the aforementioned

MapReduce (available in an open-source implementation from Hadoop) can yield great benefits but require refactoring code and rethinking computational approaches. We will discuss these and other challenges during the workshop session; participants will also share their experiences in overcoming some of these issues. Cloud computing is a classic example of more is different—working on different platforms and at larger scale offers new capabilities but also requires new ways of thinking and generates different performance-limiting problems. Nevertheless, we believe this paradigm offers the possibility for unprecedented insight into biological function.

Acknowledgments

P.K. is a visiting scientist at Google, Inc. and gratefully acknowledges many discussions with colleagues there.

References

1. Konig, R., Y. Zhou, D. Elleder, T.L. Diamond, G.M. Bonamy, J.T. Irelan, C.Y. Chiang, B.P. Tu, P.D. De Jesus, C.E. Lilley, S. Seidel, A.M. Opaluch, J.S. Caldwell, M.D. Weitzman, K.L. Kuhen, S. Bandyopadhyay, T. Ideker, A.P. Orth, L.J. Miraglia, F.D. Bushman, J.A. Young, and S.K. Chanda, *Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication.* Cell, 2008. **135**(1): p. 49-60.
2. Karlas, A., N. Machuy, Y. Shin, K.P. Pleissner, A. Artarini, D. Heuer, D. Becker, H. Khalil, L.A. Ogilvie, S. Hess, A.P. Maurer, E. Muller, T. Wolff, T. Rudel, and T.F. Meyer, *Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication.* Nature, 2010. **463**(7282): p. 818-22.
3. Hao, L., A. Sakurai, T. Watanabe, E. Sorensen, C.A. Nidom, M.A. Newton, P. Ahlquist, and Y. Kawaoka, *Drosophila RNAi screen identifies host genes important for influenza virus replication.* Nature, 2008. **454**(7206): p. 890-3.
4. Jager, S., P. Cimermancic, N. Gulbahce, J.R. Johnson, K.E. McGovern, S.C. Clarke, M. Shales, G. Mercenne, L. Pache, K. Li, H. Hernandez, G.M. Jang, S.L. Roth, E. Akiva, J. Marlett, M. Stephens, I. D'Orso, J. Fernandes, M. Fahey, C. Mahon, A.J. O'Donoghue, A. Todorovic, J.H. Morris, D.A. Maltby, T. Alber, G. Cagney, F.D. Bushman, J.A. Young, S.K. Chanda, W.I. Sundquist, T. Kortemme, R.D. Hernandez, C.S. Craik, A. Burlingame, A. Sali, A.D. Frankel, and N.J. Krogan, *Global landscape of HIV-human protein complexes.* Nature, 2012. **481**(7381): p. 365-70.
5. Hawrylycz, M.J., E.S. Lein, A.L. Guillozet-Bongaarts, E.H. Shen, L. Ng, J.A. Miller, L.N. van de Lagemaat, K.A. Smith, A. Ebbert, Z.L. Riley, C. Abajian, C.F. Beckmann, A. Bernard, D. Bertagnolli, A.F. Boe, P.M. Cartagena, M.M. Chakravarty, M. Chapin, J. Chong, R.A. Dalley, B.D. Daly, C. Dang, S. Datta, N. Dee, T.A. Dolbeare, V. Faber, D. Feng, D.R. Fowler, J. Goldy, B.W. Gregor, Z. Haradon, D.R. Haynor, J.G. Hohmann, S. Horvath, R.E. Howard, A. Jeromin, J.M. Jochim, M. Kinnunen, C. Lau, E.T. Lazarz, C. Lee, T.A. Lemon, L. Li, Y. Li, J.A. Morris, C.C. Overly, P.D. Parker, S.E. Parry, M. Reding, J.J. Royall, J. Schulkin, P.A. Sequeira, C.R. Slaughterbeck, S.C. Smith, A.J. Sodt, S.M. Sunkin, B.E. Swanson, M.P. Vawter, D. Williams, P. Wohnoutka, H.R. Zielke, D.H. Geschwind, P.R. Hof, S.M. Smith, C. Koch, S.G. Grant, and A.R. Jones, *An anatomically comprehensive atlas of the adult human brain transcriptome.* Nature, 2012. **489**(7416): p. 391-9.
6. Schatz, M.C., B. Langmead, and S.L. Salzberg, *Cloud computing and the DNA data race.* Nat Biotechnol, 2010. **28**(7): p. 691-3.
7. Dudley, J.T. and A.J. Butte, *In silico research in the era of cloud computing.* Nat Biotechnol, 2010. **28**(11): p. 1181-5.
8. Schadt, E.E., M.D. Linderman, J. Sorenson, L. Lee, and G.P. Nolan, *Computational solutions to large-scale data management and analysis.* Nat Rev Genet, 2010. **11**(9): p. 647-57.
9. Dean, J. and S. Ghemawat, *MapReduce: A Flexible Data Processing Tool.* Communications of the Acm, 2010. **53**(1): p. 72-77.
10. Dean, J. and S. Ghemawat, *Mapreduce: Simplified data processing on large clusters.* Communications of the Acm, 2008. **51**(1): p. 107-113.
11. Melnik, S., A. Gubarev, J.J. Long, G. Romer, S. Shivakumar, M. Tolton, and T. Vassilakis, *Dremel: Interactive Analysis of Web-Scale Datasets.* Communications of the Acm, 2011. **54**(6): p. 114-123.

COMPUTATIONAL CHALLENGES OF MASS PHENOTYPING

LAWRENCE HUNTER

Computational Bioscience Program
University of Colorado School of Medicine
Aurora, CO 80045 USA
Email: Larry.Hunter@UCDenver.edu

One of the primary challenges in making sense the dramatic increase in human genotype data is finding suitable phenotype information for correlational analyses. While the price of genotyping has fallen dramatically and promises to continue to decrease, the cost of generating the phenotypes necessary to take advantage of this data has held steady or even increased. Until recently, human phenotype data was primarily derived from assays or measurements made in clinical or research laboratories. However, laboratory phenotyping is expensive and low-throughput. Recently, a variety of promising alternatives have arisen that can provide important new information at greatly reduced costs. However, the nature, extent and complexity of the data produced involve significant new computational challenges.

This workshop will begin with an introduction to some of the new modalities, which include: automated abstraction of information from electronic medical records, data streams from medical instruments (e.g. in an intensive care unit) and implanted devices (e.g. cardiac assist devices), data produces by patient social networks, and data from a new generation of inexpensive wearable sensors measuring everything from physical activity to blood glucose.

Most of these new sources of phenotypic data are secondary to some other purpose. Patient records are generated to support clinical care and payment for medical services. Patient social networking sites support patients emotionally and provide peer counseling. Implantable medical devices produce data streams that meet manufacturers' or caregiver requirements. Wearable sensors satisfy personal curiosity or monitor disease progress. Each of these also produces valuable information for genotype correlations.

We will focus on defining the computational challenges arise in the collection, storage, processing, analysis and, especially, in the useful integration of these many new sources of phenotype data into derivatives that facilitate scientifically or medically valuable correlations with genotype. Computational challenges arise due to the diverse nature of the types of data that characterize human phenotypes, the fact that most phenotyping is a secondary use of data produced for other purposes, and the need to integrate, abstract and summarize data in ways that are likely to show correlations with genotype. There are also bioethical challenges in data sharing, anonymization, openness / privacy, consent, and related topics where computational methods might help address other concerns.

The new sources of phenotypic information produce data at radically different time scales and granularities. Modern medical instruments can produce data streams at 50Hz or greater sampling frequencies for days at a time. Patient social networking users typically update their entries every few days, but can be maintained for many years. Effectively integrating information that is produced at such different resolutions and durations is a difficult task. Sensor fusion approaches from other domains may be relevant, although some problems (and some solutions) may be specific to the biomedical domain. Similarly, signal processing approaches that summarize high frequency data into scalar or categorical values may prove of value in this application.

Many of the new sources of phenotypic information produce unstructured or semi-structured data, such as physician notes in electronic medical records, or postings to patient social networking sites. Biomedical natural language processing (NLP) techniques have shown some value in systematizing and normalizing this kind of textual information, but most research in this area has been for clinical decision support, information retrieval, or information extraction. Performance of NLP tools is too often modest in those applications. Are there aspects of using unstructured information to define phenotype that differ from these other applications? Are there differences that can be exploited to improve performance?

Many interesting precedents for the sort of genetic research that these new sources of phenotype data make possible can be found in traditional epidemiology; so can many of the challenges. One particularly pernicious problem in using observational data is the confounding of covariates. For example, many of the patients taking the drug Metformin have elevated blood glucose levels; that's because Metformin is a front-line drug for diabetes, not because taking the drug increases blood sugar. People who wear activity monitors are more active than those that do not, but just giving everyone an activity monitor is unlikely to increase the level of physical activity in the population. Identifying and normalizing for covariates is a critical task in taking advantage of the growth of phenotype data gathered secondary to some other purpose (such as patient care or finding social support). Integration of new data streams with more traditional epidemiological data types (such as demographics or survey results) are also an interesting area for the development of automated methods.

A different class of computational problems arises from the complex personal, social and bioethical concerns around the collection and use of phenotypic data. Are there computational approaches to anonymization, provenance, de-duplication or other problems in making it possible for patients (and normal controls) to share the data they want to with researchers, to protect their rights, to give research participants access to important conclusions drawn around them? Are there developments in electronic consenting, cryptography, or computer security that can facilitate the flow of useful data to researchers while protecting participants?

These topics are relatively new to the computational biomedicine community. The purpose of the workshop is to bring together experts in diverse areas to: identify specific driving problems, define important research topics, and perhaps to share valuable data sets and other research resources. We hope that the outcome of the workshop is a deeper understanding of the challenges, one that will eventually lead to novel computational approaches to addressing these very important problems.

THE FUTURE OF GENOME-BASED MEDICINE

QUAID MORRIS
University of Toronto, Donnelly Centre,160 College Street, Toronto, ON M5S 3E1, Canada
Email: QUAID.MORRIS@UTORONTO.CA

STEVEN E. BRENNER
Department of Plant & Microbial Biology, 111 Koshland Hall, University of California, Berkeley 94720
Email: brenner@compbio.berkeley.edu

JENNIFER LISTGARTEN
MICROSOFT RESEARCH, 110 GLENDON AVENUE, SUITE PH1, LOS ANGELES, CA
EMAIL: JENNL@MICROSOFT.COM

OLIVER STEGLE
MAX PLANCK INSTITUTES TÜBINGEN, 72076 TÜBINGEN, GERMANY
EMAIL: *OLIVER.STEGLE@TUEBINGEN.MPG.DE*

There has been unprecedented public investment in sequencing human and cancer genomes in the hopes of understanding disease [1, 2]. At the same time, large genome-wide association studies have helped elucidating the genetic underpinning of common diseases, identifying thousands of putative disease relevant loci [7, 8]. Complementary molecular profiling studies have revealed that several of these loci are co-associated with individual mRNA levels, suggesting candidate pathways that are putative mediators of genetic signals [7]. Coupled with the public investment, there is considerable personal investment in genetic profiling, being offered by companies such as 23andMe, deCODEme and others. Although this work has led to amazing discoveries, such as the surprising genetic, subclonal diversity within tumor populations (e.g., [3-5]), it's not clear how much how these insights will improve personalization of medicine.

In this workshop, we hope to address questions about how much genome sequencing has helped our understanding of the causal factors in disease and how much will these data change the way we treat disease in the clinic. Are genome clinics even realistic? If not, what other data will we need on individual patients before genome-based personalized medicine is possible?

Establishing causal mutations
Genetic variants cause disease through their impact on cell function. Despite promising recent efforts to predict phenotype from genotype in single cells [6], we are still far from being able to connect each mutation with its molecular and, ultimately, gross phenotypic consequences in humans. Establishing the causal mutation(s) is often a prerequisite to subsequent therapeutic and preventive actions especially for disease caused by a rare or somatic variant. So, a major challenge in genome-based medicine is distinguishing disease-causing polymorphisms among a large number of candidates, many of which are spuriously correlated.

In cancer, the causal mutations are called driver mutations because they contribute to the progression of cancer; these variants must be distinguished from passenger mutations that have little effect on cell function but appear as a result of greatly increased somatic mutation rates in cancer cells. In our workshop, we have two invited speakers who will present different strategies to solving the problem of identifying driver mutations. One approach is to identify those regions that are aberrant more often than expected across different tumors of the same type. This approach assumes that driver mutations appear in regions that are under positive selection during cancer biogenesis. But limits on the resolution of this analysis make it difficult to identify driver regions that are small enough to contain a single gene or single functional element. **Dr Dana Pe'er** will introduce Helios, a new method that incorporates multiple layers of cancer profiling data to increase the effective resolution of the analysis, allowing to identifying driver mutations. Another approach to find driver mutations is to infer the evolutionary history of subclonal populations of cancer cells by genetically profiling multiple sites within primary and metastatic tumors and deconvolving the subclonal evolutionary structure from these data. **Dr Sohrab Shah** will discuss analysis approaches of ovarian tumors that embody this strategy.

Connecting genetic variants to molecular and disease phenotypes
Even if the causal mutation(s) can be identified, their phenotypic impact often remains unclear, especially if they do not directly affect the encoded protein sequence. Fortunately, genomes are being sequenced along with their products on the level of RNA or other molecular layers, allowing connections to be made between genotype and molecular phenotypes. In some cases, the RNA profile alone can be used to select therapy. **Dr Lars Steinmetz** will discuss efforts to use yeast as a model system for identifying molecular signals that suggest targets for therapeutic intervention. Ultimately, as we learn more and more about genome function in coding and non-coding regions [9], we should be able to design algorithms that predict the functional consequences of individual mutations. **Dr Steven Brenner** will discuss the Critical Assessment of Genome Interpreation (CAGI) project, a community experiment to objectively assess computational methods for predicting the phenotypic impact of genome variation.

Acting on mutations
Finally, once the causal mutations and their phenotypic consequence are identified, their remains the problem of how to design targeted treatment. This remains a significant, unsolved problem but some progress has been made in identifying actionable mutations based on known (or assumed) targets of current drugs [10].

Workshop contributions
The workshop includes invited talks from researchers active in genome-based clinical research.

Dana Pe'er is an Associate Professor in Biology and Systems Biology at Columbia University, New York. She is the director of the laboratory on Computational Systems Biology and is an active researcher in both the systems biology and machine learning communities. Dana pioneered the application of Bayesian networks and Bayesian modeling techniques to molecular profiling data. She will discuss new computational methodology to integrate multiple cancer genome profiling layers to identify copy number-based driver mutations.

Sohrab Shah is Assistant Professor in the Dept. of Pathology at the University of British Columbia and a research scientist at the BC Cancer Agency. He heads a computational biology laboratory that combines deep expertise in genome sequencing and analysis with machine learning methodology development to make discoveries in breast and ovarian cancer, including two recent Nature papers. In this talk, he will discuss recent approaches to study the variation between spatially and temporally distinct tumor specimens in ovarian cancer. Understanding the variation between tumor specimens within individual patients can be used to detect driver mutations and the evolution of mutational accumulation.

Lars Steinmetz is co-chair of the genome biology unit at the European Molecular Biology Laboratory in Heidelberg, which consists of over 100 scientists and 9 research teams. In parallel, Dr. Steinmetz leads a focused research team at the Stanford Genome Technology Centre in the USA. Lars is a leading scientists at the forefront of genetic and genomics research. His lab pioneered the development and the application of high-throughput approaches to functionally profile genetic and molecular systems at a genome-wide scale. His lecture will address how observational molecular and genetic profiling information can be used to identify causal molecular mediators that confer genetic signals to phenotype. The hypotheses generated by computational modeling are systematically validated in a yeast model.

Steven E. Brenner is Professor in the Department of Plant and Microbial Biology at the University of California, Berkeley with adjunct appointments in Bioengineering and Therapeutic Sciences. He has won numerous awards for his research including the prestigious ICSB Overton Prize reflecting his broad, seminal contributions to computational biology in diverse areas including alternative splicing, protein evolution, critical assessment of bioinformatics methodology, and, most recently, genome-based medicine. His presentation will give an overview of recent community efforts to assess genome interpretation. The Critical Assessment of Genome Interpreation (CAGI) is a community experiment to objectively assess computational methods for predicting the phenotypic impact of genome variation. CAGI has revealed the relative strengths of different prediction approaches, showing some that worked consistently well, while other classes worked only on special types of problems. Even with the simplest dataset, involving nonsynonymous mutations in a human metabolic enzyme, yielded great variability of the result. Overall, CAGI revealed very significant biomedical insights into the implications of genetic variation are embodied in current algorithms, but that the ability of generic methods to make clinically important decisions is presently limited.

References

1. International Cancer Genome, C., et al., *International network of cancer genome projects.* Nature, 2010. **464**(7291): p. 993-8.
2. Siva, N., *1000 Genomes project.* Nat Biotechnol, 2008. **26**(3): p. 256.
3. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-92.
4. Shah, S.P., et al., *The clonal and mutational evolution spectrum of primary triple-negative breast cancers.* Nature, 2012. **486**(7403): p. 395-9.
5. Schuh, A., et al., *Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns.* Blood, 2012.
6. Karr, J.R., et al., *A whole-cell computational model predicts phenotype from genotype.* Cell, 2012. **150**(2): p. 389-401.
7. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.
8. Hu, X. and M. Daly, *What have we learned from six years of GWAS in autoimmune diseases, and what is next?* Curr Opin Immunol, 2012.
9. Consortium, E.P., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
10. Dancey, J.E., et al., *The genetic basis for cancer treatment decisions.* Cell, 2012. **148**(3): p. 409-20.