PRECISION MEDICINE:
DATA AND DISCOVERY FOR IMPROVED HEALTH AND THERAPY

ALEXANDER A. MORGAN

Stanford University School of Medicine
Stanford, CA 94305 USA
Email: alexmo@stanford.edu

DANA C. CRAWFORD

Epidemiology and Biostatistics, Institute for Computational Biology
Case Western Reserve University
Cleveland, OH, 44106 USA
Email: dana.crawford@case.edu

JOSH C. DENNY

Vanderbilt University Medical Center
Nashville, TN 37203 USA
Email: josh.denny@vanderbilt.edu

SEAN D. MOONEY

University of Washington
Seattle, WA 98105 USA
Email: sdmooney@uw.edu

BRUCE J. ARONOW

Center for Computational Medicine
Cincinnati Children's Hospital Medical Center and the University of Cincinnati
Cincinnati, OH 45229 USA
Email: bruce.aronow@cchmc.org

STEVEN E. BRENNER

University of California
Berkeley, CA 94720-3012 USA
Email: brenner@compbio.berkeley.edu

The major goal of precision medicine is to improve human health. A feature that unites much research in the field is the use of large datasets such as genomic data and electronic health records. Research in this field includes examination of variation in the core bases of DNA and

their methylation status, through variations in metabolic and signaling molecules, all the way up to broader systems level changes in physiology and disease presentation. Intermediate goals include understanding the individual drivers of disease that differentiate the cause of disease in each individual. To match this development of approaches to physical and activity-based measurements, computational approaches to using these new streams of data to better understand improve human health are being rapidly developed by the thriving biomedical informatics research community. This session of the 2017 Pacific Symposium of Biocomputing presents some of the latest advances in the capture, analysis and use of diverse biomedical data in precision medicine.

1. Introduction

The major goal of precision medicine is to improve human health. The researchers presenting work in the 2017 PSB conference session on precision medicine represent a wide range of approaches this challenge. The work ranges from examination of variation in the core bases of DNA and their methylation status, through variations in metabolic and signaling molecules, all the way up to broader systems level changes in physiology and disease presentation. Recent advances in areas as diverse as microfluidics, solid phase chemistry, optics, wireless communication, battery technology, and social networking are supporting the collection and analysis of a whole host of highly multiplex biomedical measurements in increasingly fine temporal resolution of sampling. Whether it is understanding the individual drivers of disease that differentiate the cause of disease in each individual, to the creation of customized drug dosing algorithms, the researchers in this session are advancing data-driven medicine from applying to populations down to individuals.

One common thread that unites much of this work is the value of large datasets combining a wide range of features that encompass causal factors, state measures, and differential outcomes. Whether using a large patient registry focused on specific phenotypes and pathologies (such as autism or cancer) or broad spectrum electronic medical record systems, the linking of data collected as part of healthcare delivery combined with molecular and genomic features has provided an invaluable resource to help create data-precision models for disease understanding and improving care.[1-3] Without these data resources, most of the work in this session would essentially be impossible. Although those who make maximal use of these large datasets have been criticized for taking undue advantage of the labor of others,[4] it is clear that making these large datasets available to biomedical informatics researchers is enabling new methodological developments and new insights to advance clinical care. One recently reported study on the genetics of hypertension used samples from over 300,000 people;[5] at this scale, data should be considered a resource of global importance to health and wellbeing, not part of the academic fiefdom of a single researcher. Newborn screening extends this to its largest scale, addressing every member of a population (e.g., nearly 500,000 per year in California) without bias.[6]

The extensive work reported in this session reflects the diversity of activity in precision medicine and the enthusiasm in the field. However, this enthusiasm must be tempered with healthy caution and skepticism. Last year, the PSB session on precision medicine[7] was accompanied by another session focused on aspects and challenges in reproducibility[8] in research, and this continues to be a challenge in our efforts to develop an individualized understanding of physiology and disease as each person is in effect a sample size of one. This is a challenge across science, and much of the research across the psychological sciences has recently been criticized for its poor level of reproducibility.[9] In parallel to the methodological challenges in reproducibility, there continues be healthy skepticism and cautious evaluation of the continuously evolving techniques and approaches to collecting samples, measuring their properties, and evaluating their biomedical significance in isolation or in combination with other data and properties. A recent evaluation of a direct-to-consumer lab testing technology[10] revealed that any claim of technological advance without appropriate controls, comparisons, and supporting evidence must be examined in open formats by external parties before launching its widespread use in clinical care.

The burden of very careful experimental design and reproducibility does not mean that the field of precision medicine is advancing slowly. For example, recent work has shown that it is possible to develop predictive, customized models of blood glucose level in response to different forms of dietary intake, a huge advance in precision, personalized nutrition.[11] Further research will demonstrate whether these models are stable over time.

The recent CAGI (Critical Assessment of Genome Interpretation)[*] evaluations have shown the power of the Common Task Framework to allow researchers to compare techniques that make predictions of phenotype from genotype, a key element of precision medicine. However, one of the trade-offs is that improved prediction accuracy often comes at the cost of human interpretability. For example, in the most recent CAGI of 2016, an approach using deep neural networks to predict psychiatric disease status from exome data performed better than other approaches that used far more interpretable models and those that integrated far more human knowledge. Unfortunately, the maturity of performance of these techniques of machine learning currently exceeds the maturity of the tools the to help interpret their predictions, limiting our ability to correct the apparent biases in our human understanding. Consequently, the significance and application of these findings are unclear. Much work has been done in fields like natural language processing and image processing to help visualize and unpack complex predictive AI models;[12] however, successful approaches and visualizations to fully support this increased understanding of many of the currently "black box" models of genomics and precision medicine are continuing to be developed.[13, 14] The many pieces of work presented in this

[*] https://genomeinterpretation.org

session use a range of visualizations and evaluation metrics, but this continues to be an active area of endeavor in need of new advances.

Concomitant with advances in predictive and analytic approaches, informatics, and machine learning techniques are learning how to perform goal directed tasks, often at better than human levels of performance.[15] It is hoped that we can go beyond simple tasks like playing complex games to guidance of the steps and actions in the delivery of healthcare; however, as noted this will require healthy and active skepticism along with insight into the models developed.

2. Podium presentations

When Hippocrates espoused the idea that physicians should be literate and keep records of patient care and outcomes[16], it was so that these records might be used to improve the understanding of disease and help future patients. It is therefore not a new idea that medical records might be a powerful source of data for advancing biomedicine; the widespread use of electronic medical records systems has allowed several researchers in this session to use these data to deepen our understandings of disease and possible new methods of precision treatment. In particular, **CR Bauer and colleagues** investigate the relationship between genetic variation and 29 common laboratory values. Importantly, they go beyond simply viewing each of the laboratory values as simple quantitative traits, but look at the relationships between those quantitative traits and start to examine compositional quantitative traits derived from those measurements. Although it is common to think about the multiplicity of possible hypotheses derived from examining many genetic variants, little effort is typically spent examining the multiple hypotheses that can be derived from how we partition and divide phenotypes. In the closely related work of **SS Verma and colleagues**, the focus is on the genetic drivers of the variability of common laboratory measurements. Going beyond the conventional central tendency of the laboratory values, they examine genetic associations with heteroscedasticity. This shift in focus from average value or pure prediction accuracy, toward models of higher moments and a focus on understanding what drives dispersion is another theme that runs through several of the papers in this session.

Laboratory values are part of assigning diagnoses, and **MK Beck and colleagues** mine records from 6,923,707 Danish patients to examine issues around the temporal ordering of diagnoses. They focus on the conditions of diabetes and sleep apnea, which often co-occur, but their presence can be hidden from the sufferer for years, and identification of one can lead to ascertainment bias of the other. When mining clinical records, researchers have access to when a disease was diagnosed but little data as to why, which may be impacted by a range of externalities, including differing access to care, but Beck and colleagues investigate patterns of age trajectory and of subsequent disease diagnoses, and data driven methods to stratify patients into subgroups. Moving up from laboratory measurements and diagnoses to directly guiding clinical decision making, but still using data derived from records of clinical care, **LK Wiley and colleagues** evaluate models that determine dosing of a medication with a narrow therapeutic window (warfarin) based on genetic variations in admixed populations, particularly those with African ancestry.

In addition to large datasets of mixed-type patient records, disease registries around specific diseases are a powerful data resource for precision medicine informatics. Three pieces of work in this session focus on techniques for identifying how variants in groups of genes may work together to contribute to phenotype, and much of this work relies on disease specific registry data. **GR Venkataraman and colleagues** use data from an autism patient registry to examine the way *de novo* mutations diffusely spread across sets of genes of shared function and how they may contribute to disease risk. The challenge of polygenic phenotypes is also the focus of the work of **D He and L Parida**, who presently work on disentangling epistasis underlying quantitative traits. **J Gallion and colleagues** have also been working on examining genetic variations in families of genes, in this case families of kinases in cancer, highlighting the shared disease association of variations in homologous locations across genes in a particular family.

Digging in more deeply into cancer, particularly using the data provided by The Cancer Genome Atlas,[17] **JA Thompson and CJ Marsit** present their work combining methylation with gene expression data to predict cancer survival; mixing heterogeneous data with colinearities being a hallmark of much of the cutting edge of precision medicine work. **G Speyer and colleagues** turn focus to drug response in cancer cells. Their work investigates the way expression dependency graphs vary between responsive and non-responsive cells; continuing the theme of mining differences in dispersion, here spread around network connectivity and likelihood, between subgroups. They also make the results of their work available online as a searchable resource.[†]

3. Posters with published papers

The work presented in our poster session with published papers represents a broad range of interests by research groups, with some fairly technical work delving in deeply to new methods of analysis of biomedical data. **A Beck and colleagues** present an approach for using genome uncertainty to modify thresh-holding for tests of Hardy-Weinberg equilibrium; highly relevant to some of the most basic

[†] http://biocomputing.tgen.org/software/EDDY/CTRP/home.html

analysis done in population genomics and often serving as a filter for all the analysis downstream of the variant calling.

The entire *in silico* metabolic modeling of the most simple of single cells is now a reality,[18] and techniques of temporal molecular metabolic flux analysis are advancing dramatically.[19] **A Schultz and colleagues** are working to identify cancer specific metabolic signatures, and we may one day have patient and cancer-specific cellular metabolic models as tools for precision medicine.

As noted, one of the themes of this session has been on the investigation of measures of dispersion as a key biological measures, and **PF Kuan and colleagues** are examining DNA methylation, with methylDMV, a tool that compares not only measures of central tendency but also heteroscedasticity, as a way to highlighting issues like sample bias vs. biological signal.

There is a substantial amount of work in this session delving into methods to better identify cancer subgroups, both as a tool to more precision in individualized prognostic models, but perhaps more importantly to find features that unite these groups that might lead to precisely targeted therapies in those subgroups, or at least provide increased clarity on which existing therapies are likely to more or less efficacious. Cancer driver mutation identification is the focus of the work by **M Ma and colleagues**. **A Durmaz and colleagues** present work on subgraph analysis with a focus on grouping via dysregulated pathways. **H Kabbat and colleagues** use a competitive endogenous RNA based method combining DNA copy number variation, mRNA expression, and microRNA levels.

The interest in pathway analysis and uniting mRNA with microRNA is united in the work of **D Diaz and colleagues**, which focuses on just that topic. Finally, **T Kamp and colleagues** present work on the value of moving to a more Boolean view of gene expression when doing gene set enrichment analysis in improving analytical output.

4. Acknowledgments

We would especially like to thank the many reviewers who contributed their considerable expertise and precious time to help the many paper submitters refine the presentation of their work. We would also like to thank the PSB 2017 chairs and Tiffany Murray of Stanford University for their efforts in organizing the meeting. This work is supported in part by NIH grants U19 HD077627, R01 AI105776, U41 HG007346, and by a Research agreement with Tata Consultancy Services.

5. References

1. Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat Rev Genet*. 2016;17(3):129-45.

2. Denny JC, Bastarache L, Roden DM. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annual Review of Genomics and Human Genetics*. 2016;17(1):353-73.

3. Roden DM, Denny JC. Integrating electronic health record genotype and phenotype datasets to transform patient care. *Clinical Pharmacology & Therapeutics*. 2016;99(3):298-305.

4. Longo DL, Drazen JM. Data Sharing. *New England Journal of Medicine*. 2016;374(3):276-7.

5. Ehret GB, Ferreira T, Chasman DI, Jackson AU, Schmidt EM, Johnson T, et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nat Genet*. 2016;48(10):1171-84.

6. Brenner SE, Kingsmore S, Mooney SD, Nussbaum R, Puck J. Use of genome data in newborns as a starting point for life-long precision medicine. *Pac Symp Biocomput*. 2016;21:568-75.

7. Morgan AA, Mooney SD, Aronow BJ, Brenner SE. Precision medicine: data and discovery for improved health and therapy. *Pac Symp Biocomput*. 2016;21:243-8.

8. Manrai AK, Wang BL, Patel CJ, Kohane IS. Reproducible and shareable quantifications of pathogenicity. *Pac Symp Biocomput*. 2016;21:231-42.

9. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*. 2015;349(6251).

10. Kidd BA, Hoffman G, Zimmerman N, Li L, Morgan JW, Glowe PK, et al. Evaluation of direct-to-consumer low-volume lab tests in healthy adults. *The Journal of Clinical Investigation*.126(7):2773.

11. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized Nutrition by Prediction of Glycemic Responses. *Cell*. 2015;163(5):1079-94.

12. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knolwedge Discovery and Data Mining*. 2016:1135-44.

13. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*. 2013;14(2):178-92.

14. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, et al. Visualization of omics data for systems biology. *Nat Methods*. 2010;7(3):S56-S68.

15. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016;529(7587):484-9.

16. Kassell L. Casebooks in early modern England: medicine, astrology, and written records. *Bull Hist Med*. 2014;88(4):595-625.

17. The Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45(10):1113-20.

18. Karr Jonathan R, Sanghvi Jayodita C, Macklin Derek N, Gutschow Miriam V, Jacobs Jared M, Bolival Jr B, et al. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*. 2012;150(2):389-401.

19. Birch EW, Udell M, Covert MW. Incorporation of flexible objectives and time-linked simulation with flux balance analysis. *Journal of Theoretical Biology*. 2014;345:12-21.

OPENING THE DOOR TO THE LARGE SCALE USE OF CLINICAL LAB MEASURES FOR ASSOCIATION TESTING: EXPLORING DIFFERENT METHODS FOR DEFINING PHENOTYPES

CHRISTOPHER R. BAUER[1], DANIEL LAVAGE[1], JOHN SNYDER[1], JOSEPH LEADER[1], J. MATTHEW MAHONEY[2], SARAH A. PENDERGRASS[1]

Biomedical & Translational Informatics, Geisinger Health System
100 N. Academy Ave. Danville, PA 17821, USA
Email: cbauer@geisinger.com

Department of Neurological Sciences, University of Vermont College of Medicine
149 Beaumont Ave. Burlington, VT 05405, USA

The past decade has seen exponential growth in the numbers of sequenced and genotyped individuals and a corresponding increase in our ability of collect and catalogue phenotypic data for use in the clinic. We now face the challenge of integrating these diverse data in new ways new that can provide useful diagnostics and precise medical interventions for individual patients. One of the first steps in this process is to accurately map the phenotypic consequences of the genetic variation in human populations. The most common approach for this is the genome wide association study (GWAS). While this technique is relatively simple to implement for a given phenotype, *the choice of how to define a phenotype is critical*. It is becoming increasingly common for each individual in a GWAS cohort to have a large profile of quantitative measures. The standard approach is to test for associations with one measure at a time; however, there are many justifiable ways to define a set of phenotypes, and the genetic associations that are revealed will vary based on these definitions. Some phenotypes may only show a significant genetic association signal when considered together, such as through principle components analysis (PCA). Combining correlated measures may increase the power to detect association by reducing the noise present in individual variables and reduce the multiple hypothesis testing burden. Here we show that PCA and k-means clustering are two complimentary methods for identifying novel genotype-phenotype relationships within a set of quantitative human traits derived from the Geisinger Health System electronic health record (EHR). Using a diverse set of approaches for defining phenotype may yield more insights into the genetic architecture of complex traits and the findings presented here highlight a clear need for further investigation into other methods for defining the most relevant phenotypes in a set of variables. As the data of EHR continue to grow, addressing these issues will become increasingly important in our efforts to use genomic data effectively in medicine.

1. Introduction

In the past decade, genome wide association studies (GWAS) have revealed more than ten thousand associations between genetic loci and traits [1]. As GWAS continue to grow in number, sample size, and range of phenotypes, they offer an opportunity to begin to untangle the complex network underlying phenotypic variation. One challenge in this pursuit stems from an asymmetry in the genotype-phenotype map. While the range of genetic variation in humans is fairly well characterized and a given genome can be sequenced to arbitrary depth, there is no obvious way to measure a physiologically complete phenome or even outline how to divide it into separate units [2]. Even subtle choices in how a phenotype is defined can affect which loci associate with it [3, 4]. There is a growing need to analyze these choices and their effects if we wish to build a genotype-phenotype map that captures the relationships most relevant to biology and the clinic.

The first human GWAS defined phenotypes based on clinical case control status [5, 6, 7]. Binary phenotypes such as these are a natural choice if our ultimate goal is to predict disease risk, but diseases are typically diagnosed based on a number of underlying quantitative variables and expert opinions. For example, dozens of loci have been implicated in the risk of multiple sclerosis [8]. However, this condition is heterogeneous in its presentation and is diagnosed based on an accumulation of symptoms, quantitative measures, and subjective categorization, only after ruling out other conditions [9]. There are also subtypes of multiple sclerosis as well as other distinct but related demyelinating syndromes [10, 11]. This complexity makes it exceedingly difficult to understand how each of the associated gene variants might be contributing to the disease.

Recently we have begun to see association studies conducted in cohorts that have been given batteries of quantitative assays [12, 13] and comprehensive electronic health record (EHR) data is being used to construct phenotypic profiles. The availability of these large sets of traits has lead to an approach known as the phenome wide association study (PheWAS) where each variant is tested for associations with a range of phenotypes [14, 15]. Recent applications of PheWAS have revealed many novel genotype-phenotype associations and the potential for a large degree of pleiotropy within disease related traits [14, 16, 17]. Variants that associate with multiple traits could be indicative of genetic modules that underlie multiple diseases but in some cases they may simply represent partially redundant measures that correspond to a single disease state.

Given a profile of quantitative traits, multivariate techniques such as principal component analysis allow us to combine related variables into a set of statistically independent measures. Combining different raw measurements into new metrics can identify new associations that may provide important insights into the biology of complex traits and may provide better predictors of disease risk [18, 19]. Consider for example, four GWAS for height, weight, and body mass index (BMI), and type II diabetes. Even though BMI is simply a function of height and weight, the results of these three associations tests will not identify exactly the same sets of loci. Likewise, many variants associate with both BMI and type II diabetes, but a large part of this overlap stems from BMI being a risk factor for type II diabetes [20]. Metabolomic studies have also demonstrated that some gene variants show much stronger relationships with the the ratios of metabolites than they do with the absolute abundances of either molecule [18].

While an EHR can contain thousands of types of data, such as clinical laboratory measures, similar variables may be collected or reported in different ways. Logical observation identifiers names and codes (LOINC) provides unique numerical identifiers to distinguish relevant differences between laboratory measures [21]. Most analyses that have been conducted to date have involved laborious data harmonization procedures to ensure that grouped lab results measure the same quantity in the the same way [22]. With the large numbers and types of measures in the EHR, it is often not feasible to carefully harmonize each and every phenotype. Thus, it is important to explore approaches that will allow for high throughput use of multiple phenotypes.

Here, we have mined the Geisinger Health System EHR for quantitative measures to produce a high dimensional phenotypic profile for a large population of genotyped patients in the MyCode® Community Health Initiative. Using these data, we outline and compare three general strategies for identifying loci that associate with one or more components of this phenomic profile: PheWAS, PCA, and cluster PCA. Our results show that each of these methods can detect associations that are missed by the others and that the significance of a given association can vary by many orders of magnitude based on how a phenotype is defined. These findings set the stage for further use of EHR data in gene associations studies and highlight important considerations as we attempt to improve the predictive power of medical genomics and clinical phenotyping.

2. Methods

2.1. *Genetic Data*

All of the data described in this paper come from a cohort of patients in the MyCode Community Health Initiative at the Geisinger Health System. Each patient was genotyped for 659,010 SNPs with minor allele frequency greater than 1% using Illumina OMNI Express Exome chips. We excluded any SNPs that had call rates < 99%, sample call rates < 99%, as well as 113 SNPs that show large differences between batches. We restricted our analysis to individuals with greater than 99% likelihood of European ancestry, as defined by quadratic discriminant analysis using the first four principal components of ancestry based on the 1000 genomes project.

2.2. *Phenotypic Data*

For 38,269 patients in the Geisinger Health System that met these criteria, we extracted age, sex, BMI and the median values for the following 29 outpatient laboratory measures as defined by LOINC codes (Table 1). Most of the lab measures showed large deviations from normality at the population level, so we first performed Box-Cox transformations on each variable. Each variable was also centered and scaled by subtracting the mean value and dividing by the standard deviation.

2.3. *Imputation*

Within the set of lab data that we analyzed, 7.1% of patient-lab pairs had no results available. Nearly a third of the missing data came from ~6000, mostly young, individuals that lacked lipid measurements (Figure S1). We used predictive mean matching to impute all missing values. Imputation was performed in R, using the MICE package. Due to multicollinearity, within a subset of the 29 variables, we excluded 11 pairs of variables with correlation coefficients greater than 0.5 as predictors of each other. Aside from this restriction, each variable was modeled as a linear function of all other variable, include age, sex, and BMI. We performed 5 separate imputations, selecting among the 5 closest cases, over 120 iterations. Nearly all chains exhibited convergence with 20 iterations. In the majority of cases, the distribution of imputed values was indistinguishable from the original distribution (Figure S2).

Table 1. Definitions of the LOINC codes extracted from electronic health records.

LOINC	Description
718-7	Hemoglobin [Mass/volume] in Blood
4544-3	Hematocrit [Volume Fraction] of Blood by Automated count
787-2	Erythrocyte mean corpuscular volume [Entitic volume] by Automated count
786-4	Erythrocyte mean corpuscular hemoglobin concentration [Mass/volume] by Automated count
785-6	Erythrocyte mean corpuscular hemoglobin [Entitic mass] by Automated count
6690-2	Leukocytes [#/volume] in Blood by Automated count
789-8	Erythrocytes [#/volume] in Blood by Automated count
788-0	Erythrocyte distribution width [Ratio] by Automated count
32623-1	Platelet mean volume [Entitic volume] in Blood by Automated count
777-3	Platelets [#/volume] in Blood by Automated count
2345-7	Glucose [Mass/volume] in Serum or Plasma
2160-0	Creatinine [Mass/volume] in Serum or Plasma
2823-3	Potassium [Moles/volume] in Serum or Plasma
3094-0	Urea nitrogen [Mass/volume] in Serum or Plasma
2951-2	Sodium [Moles/volume] in Serum or Plasma
2075-0	Chloride [Moles/volume] in Serum or Plasma
2028-9	Carbon dioxide, total [Moles/volume] in Serum or Plasma
17861-6	Calcium [Mass/volume] in Serum or Plasma
1743-4	Alanine aminotransferase [Enzymatic activity/volume] in Serum or Plasma by With P-5'-P
30239-8	Aspartate aminotransferase [Enzymatic activity/volume] in Serum or Plasma by With P-5'-P
1975-2	Bilirubin.total [Mass/volume] in Serum or Plasma
2885-2	Protein [Mass/volume] in Serum or Plasma
10466-1	Anion gap 3 in Serum or Plasma
751-8	Neutrophils [#/volume] in Blood by Automated count
2093-3	Cholesterol [Mass/volume] in Serum or Plasma
2571-8	Triglyceride [Mass/volume] in Serum or Plasma
2085-9	Cholesterol in HDL [Mass/volume] in Serum or Plasma
13457-7	Cholesterol in LDL [Mass/volume] in Serum or Plasma by calculation
2965-2	Specific gravity of Urine

2.4. *Principal Component Analysis*

For each imputed dataset, we performed principal component analysis (PCA) in R, using the *prcomp* function. The PCA results were nearly identical within each imputed dataset. The average angle between all ordered pairs of Eigenvectors for the first 19 components was 4.9° and the only angles greater than 20° were caused by an alternation in the order of components 20 and 21 in some of the analyses (Figure S3). Given the minimal differences between the imputed data sets, we chose the first imputed data set to use in all downstream analyses.

2.5. *K-means clustering*

Using K-means clustering, we divided our 29 variables into 7 clusters based on their pairwise absolute correlations (Figure 1). The distance between two LOINC codes was defined as $1\text{-}R^2$. Clustering was performed in R using the *kmeans* function with 200 random starting clusters. Since sum of squares measures did not indicate an optimal number of clusters, we choose the maximum number of clusters where all clusters contained at least 3 phenotypes.

2.6. *GWAS*

We first performed associations between all 29 phenotypes individually (Figure S4). We also performed associations with 29 principle component scores (Figure S5). Finally, we performed associations with scores of the principal components within each cluster (Figure S6). All association tests were performed using PLATO 2.0 (https://ritchielab.psu.edu/plato). In each case, we modelled the principal component score as an additive function of allele count with age, bmi, sex, and the first four principal components of ancestry included in the model as covariates.

3. Results

Our phenotypic dataset comprised 29 outpatient clinical lab measures extracted from Geisinger Health System EHR. In order to ensure compatibility with other datasets, we choose to include only measures that complied with the LOINC standard of medical laboratory observations [23]. For each of the 29 clinical lab measures, we performed a separate GWAS in PLATO. Using these measures, we identified 6361 statistically significant associations (FDR < 0.01). Every lab measure had multiple SNPs associated with it, ranging from 12 SNPs for chloride concentration in blood to 783 SNPs for the number of leukocytes per unit of blood (Figure 1). Of these associations, 31% involved a SNP that was linked to more than one lab measure.

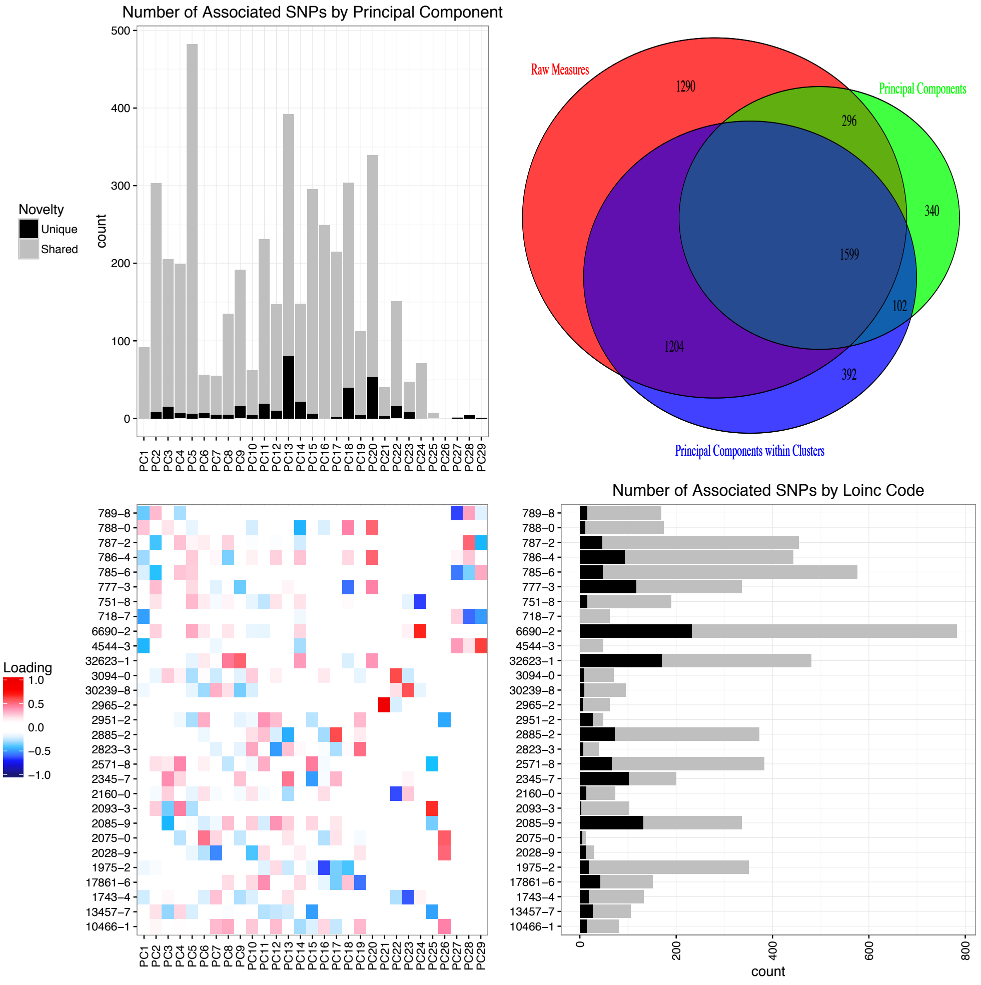


Figure 1. Associations detected with LOINC measures, PCA, and cluster PCA. The Venn diagram in the upper right panel shows the number of unique and shared SNPs that were associated with a phenotypic measure as defined by each of the three methods. The upper left panel shows the number of SNPs that associated with each principal component. The lower right panel shows how the associations were distributed across the LOINC measures. Gray bars represent the total number of SNPs while black shows the number that are unique to that measure. The bottom left panel shows how each of the phenotypes defined by LOINC codes loads onto each of the principal components.

Given that several groups of the lab results had very strong correlations and nearly all showed at least modest correlations with a few other variables (Figure 2) we hypothesized that statistical power might be improved by combining highly correlated measures. To test this, we performed principal component analysis on the combined set of all 29 lab measures. A plot of the cumulative variance explained by each additional component was smooth and increased gradually indicating that even the highest components might be measuring physiologically meaningful traits (Figure S7).

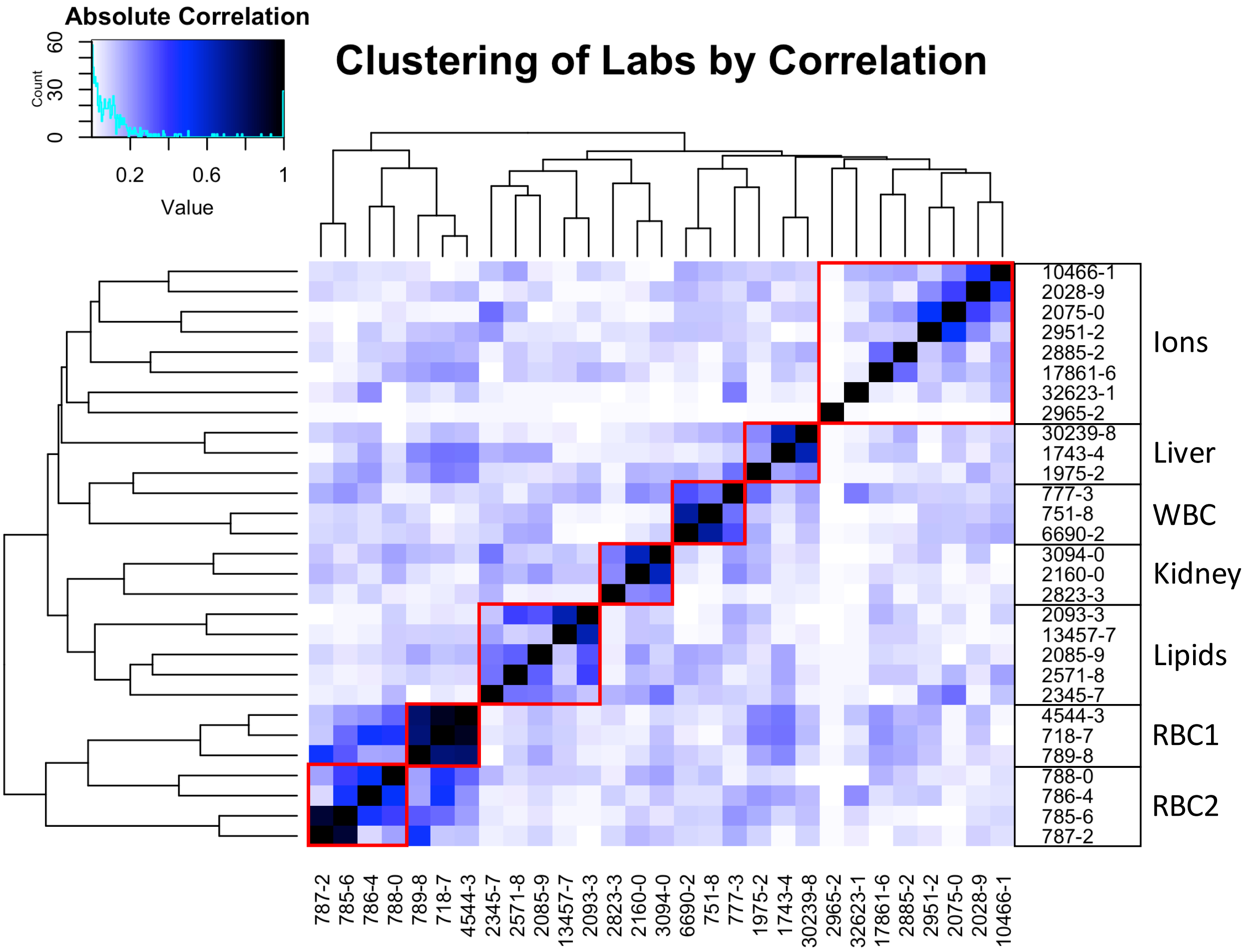


Figure 2.
Clustering of LOINC measures into related groups. The heat map indicates the absolute value of the correlation coefficient between all pairs of LOINC codes. Each cluster, as defined by k-means, is indicated by a red bounding box. The names on right column indicated the functional categories that describe each cluster.

We next performed GWAS for all 29 principal components, just as we did for the original measures (Figure S5). This analysis resulted in 4536 significant associations (FDR = 0.01). We expected to see a reduction in the total number of significant associations as one principal component could capture variation from multiple raw measures. Surprisingly, 48% of these associations involved a SNP that was linked to multiple components. Figure 1 shows a Venn diagram comparing the number of unique and overlapping associations across the various approaches for phenotypes used in this paper. Although 2494 of the SNPs that associated with one or more of the LOINC measures did not show a significant association with any of the principal components we did discover 442 new associations using these scores. 1895 SNPs associated with both a raw measure as well as a principle component (PC).

PC5 had the largest number of significant associations, 482, followed by PC13 with 392 and PC20 with 339. There was no clear pattern in how the significant associations were distributed among the first 24 components, although there were practically no associations with PC25-29 (Figure 1). In PCA, the first few components often capture a large percentage of the variation so it was interesting to see so many SNPs associating with higher components while PC1 only had 92 associated SNPs. Further analyses provide some insights. First, if we include age and BMI in the set of variables prior to PCA, we find that these variables load most strongly onto components 1. This makes sense given that age and BMI contribute to many physiological measures, especially among disease relevant traits. However, since these are both covariates in the GWAS regression model, it would be troubling to see many SNPs associated with PC1.

In PCA, the loadings indicate the magnitudes and directions that the original measures contribute to each component. Analyzing the loadings of some noteworthy components provided some additional clues to the causes of this behavior. PC2 was dominated by a few measures of blood cells: namely, the volume of erythrocytes moving in the opposite direction of cholesterol and the numbers of erythrocytes, white blood cells (WBCs), and platelets (Figure 1). PC5 was similar with WBC counts moving in the opposite directions of platelet volume and cholesterol. These associations may reveal overlap in genetic networks that regulate lipids and and the immune system. A number of studies have previously identified relationships between WBC counts and carotid plaque thickness, body fat percentage, and lipid profiles [24, 25, 26].

While PC2 and PC5 were linked to many loci, these were predominantly the same loci that were linked to one or more of the original measures. More relevant than the total number of associations detected is the number of associations that were unique to a principal component and not detectable using any of the original measures. PC13, PC18, and PC20 were responsible for the majority of these novel associations. PC13 measures a complex relationship among our measures in which serum levels of potassium and glucose vary inversely with total protein and creatinine. This is interesting because potassium and creatinine are highly correlated at the population level and both are diagnostic of kidney function. 89% percent of the associations with PC13 also map to the HLA locus suggesting a relationship between the adaptive immune system and these blood measures. PC18 and PC20 both measure relationships among erythrocyte distribution width, hemoglobin, and platelet measures (Figure 1).

Overall, the principal component approach detected fewer total significant associations than the LOINC measures, but a few components did allow us to identify novel associations. The principal components that proved most useful in this regard seemed to load primarily off of 2-6 measures (Figure 1) and those measures tended to be closely related. Components that were dominated by a single measure or had large number of weak loadings did not yield many novel results. These observations suggested a third approach. If we first divided the original 29 measures into small groups of related traits before performing PCA, we might restrict our range of phenotypes to space that corresponds better to the ways that gene variants actually impact phenotype.

Using K-means clustering, we divided our 29 variables in 7 clusters based on their pairwise absolute correlations. The choice of the number of clusters was somewhat arbitrary as the sums of square both within and between clusters never reached obvious plateaus. The choice of 7 clusters resulted in each group containing 3-8 measures, which corresponded well to our desired range, and it also broke them into groups that made intuitive sense (Figure 2). For example, all of the white blood cell counts formed a single cluster, and all of the lipid measures clustered together with serum glucose. We then performed PCA within each of these clusters and used these principal component scores to run a third GWAS with the same parameters as the previous two (Figure S6).

The genetic variants that associated with the scores of the cluster principal components had much larger overlap with original measures, sharing 2803 SNPs, but it also revealed 392 new SNPs that did not associate with either the original measures or the principal components of the entire data set (Figure 1). The distribution of these new associations varied greatly among each cluster (Figure S8). Within the ions cluster, the majority of the SNPs showed stronger associations with one of original measures than they did with any principal component (Figure S14). Within the three phenotypes that compose the liver cluster (1743-4: alanine aminotransferase, 30239-8: aspartate aminotransferase, and 1975-2: bilirubin), the associations detected for all three principal components correlated almost perfectly with those of one of the original measures (Figure S15). However, within the red blood cell cluster 1 (718-7: hemoglobin, 4544-3: hematocrit, and 789-8: erythrocytes), nearly all of the alleles tested showed their strongest association with one of the principal components (Figure 3).

Within each cluster, the middle components were the most likely to have novel associations. In general, PC1 had associations that were very similar in their significance levels to those found with the original measures. With each successive PC, the p-values would usually become more significant with respect to the LOINC measures, but less significant in absolute terms due to the reduction in total variance with each PC. In the red blood cell 1 (RBC1) cluster, nearly all of the novel significant associations occur with PC2 (Figure 3). A high score in this component corresponds to a low count of erythrocytes per unit volume of blood, but a high hematocrit score, and hemoglobin concentration. Since none of these associations were not found using erythrocyte mean corpuscular volume (787-2) as the phenotype of interest, it seems that there are a large number of gene variants linked to the concentration of hemoglobin within erythrocytes. A Manhattan plot shows that these new associations come from many distinct loci (Figure 4).

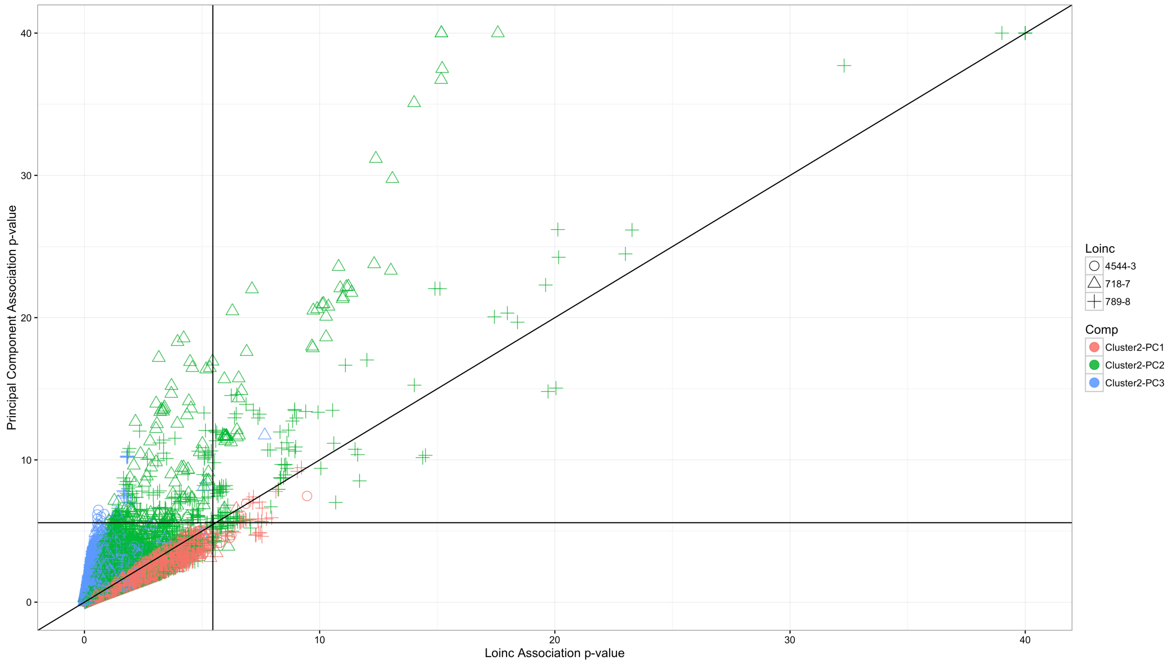


Figure 3.
Comparison of p-values for associations with the principal components and LOINC measures that compose the red blood cell cluster 1. Each point in the scatter plot represent one SNP. Both axes are scaled to the negative log base ten of the p-values. The y-axis indicates the lowest p-value that a given SNP had with any of the principal components. The components are coded by the color or the point. The x-axis indicates the lowest p-value that a given SNP had with any of the LOINC measures. The LOINC measures are coded by the shape of the point.

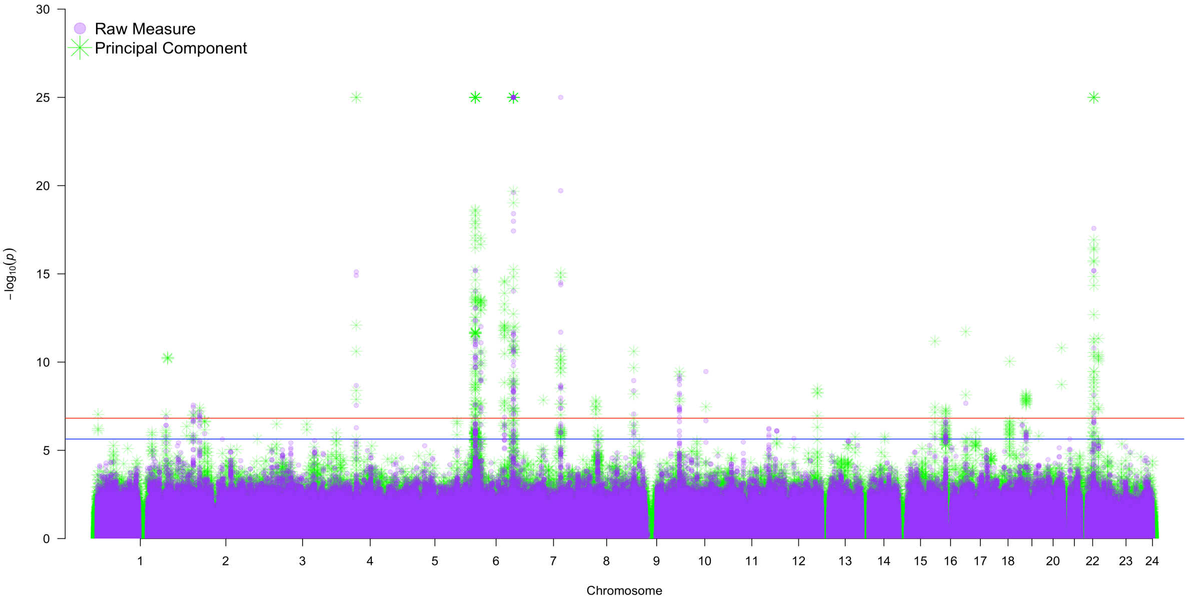


Figure 4.
Manhattan plot of the associations detected for the RBC1 cluster. The x-axis indicates the chromosomal coordinate and the y-axis shows the negative log base ten of the association p-value. Associations with any of principal components and LOINC measures are displayed in green and purple, respectively. The red line indicates a false discovery rate of 0.001 and the blue line indicates a false discovery rate of 0.01.

4. Discussion

Our results demonstrate that the choice of how to define a phenotype can have a large impact on our ability to detect relationships with genetic loci. Given a set of quantitative trait measures, we have outlined three different strategies for defining phenotypes prior to association testing. The standard method is to simply test against whatever phenotypic measures are in hand, without any additional considerations. While some measures of phenotype may be arbitrary or based purely on convenience, this may still be the most reasonable choice in many situations. In this particular case, the original phenotypes come from clinical lab tests that are prescribed because they have proven to be useful diagnostics and we find the greatest number of significant associations using these measures alone.

In spite of this generality, our results also indicate that many genotype-phenotype connections are not apparent when phenotypes are considered individually. Using two different methods based on principal component analysis, we have increased the number of significant associations that we could detect by 19%. Given the extremely large number of hypotheses that are tested in a single GWAS experiment, the p-value threshold for significance must be correspondingly low. Most segregating alleles have relatively small impacts on any given phenotype and we are unlikely to detect a significant association unless the phenotype of interest aligns very well with the effect of the variant. The majority of true positive results will inevitably fail to reach the significance threshold.

Principal component analysis provides one strategy for overcoming some of these obstacles. When performed on the entire dataset, it has the ability to capture relationships between diverse phenotypic measures. In this case, the components with the largest variance did not provide much new information. This is likely because these components capture the covariance of large numbers of measures that relate to the biggest sources of phenotypic variation in populations, such as age. There should not be genetic determinants of age, except in extreme cases, and even if there were, it is common practice to control for the effects of age in regressions.

The greatest utility of this method comes from the middle order components that capture more complex relationships. In our analyses, PC13 was related to a complex interaction between serum concentrations of potassium, creatinine, glucose, and total protein. It is not yet clear how this relates to human physiology, but the fact that 79 SNPs, distributed widely across the HLA locus, associate more strongly with this principal component than they do with any of the measures that contribute to it suggests some underlying mechanistic connection between this combination of variables and the function of the immune system. Perhaps phenotypic profiles such as this will also prove to be useful indicators of disease risk or progression.

This approach also allows us to observe effects that are orthogonal to the primary axis of variation. For example, creatinine and urea levels are both indicative of kidney function and they are very highly correlated at the population level. However, urea is a byproduct of all protein metabolism while creatinine is produced only by muscles so it is reasonable to assume that various genes could influence these traits independently. Indeed, principal component 22 corresponds to an inverse relationship between these two variables and several variants associate with the ratio of creatinine to urea in the blood without a detectable relationship to either variable in isolation.

One of the weaknesses of using principal component scores from a large dataset is that the eigenvectors correspond to the maximal variance within a set of measures which may not have any relationship to how traits are influenced at the gene regulatory level. A SNP that might correspond to elevated total cholesterol is unlikely to affect every other trait that correlates with cholesterol in a population. It can also become difficult to extract meaning from a principal component that is influenced by many, potentially disparate measures. If we hope to translate research findings into clinically relevant information, it can be useful to limit our search space to a number of dimensions that a human can understand. In order to strike balance between exploring the full range of complex interactions in biology and maintaining the ability to interpret our results, we also investigated a third approach that involved clustering our data based on the correlation structure of the variables prior to performing PCA.

While this did not improve our power in all cases, several groups of related measures yielded many more genetic associations, and at least a few new associations were discovered within each cluster. In particular, assays of blood cells and kidney function seem to benefit the most from this technique. The first 3 components of the RBC2 cluster collectively associate with 134 SNPs that do not show significant associations with any other measure that we tested. These components each measure different ways that the variance in erythrocyte size relates to hemoglobin concentration and mean erythrocyte volume. It is interesting to note that PC3 from this cluster had the most unique associations and is related to PC20 from the global PCA, which also identified new SNPs. Within the kidney cluster, PC3 measures the difference between urea and creatinine levels and associates with 41 unique variants. Again, this is related to PC22 from the global analysis. The fact that both clustered and global PCA identify associations with complex interactions between multiple blood cell and kidney function measurements indicates that the genetic regulation of these traits is not captured well by any single measure. It will be interesting to test if these same interactions are linked to the prevalence or prognosis for any disease states.

It is likely that numerous other combinations of the underlying measures would yield even more connections between gene variants and phenotypes but there is no way to exhaustively explore them. As the number of phenotypic measures that we can collect for a GWAS cohort continues to grow, it will be increasingly important to develop better strategies for specifying exactly which measures to choose test for associations. Further investigation into this topic will be critical to gaining insight into gene function and has deep implications for how we think about concepts such as pleiotropy.

Acknowledgements

We would like to acknowledge Marylyn Ritchie, Marta Byrska-Bishop, Anna Basile, Anurag Verma, and H. Lester Kirchner for helpful discussions.

References

1. D. Welter et al., *Nucleic Acids Res.* **1;42**, 1001 (2014).
2. M. Samuels, *Curr Genomics*. **11(7)**, 482 (2010).
3. W. Bush and J. Moore, *PLoS Comput Biol*. **8(12)** (2012).
4. E. Stergiakouli et al., *Obesity.* **10**, 2252 (2014).
5. R. Klein et al., *Science*. **308**, 385 (2005).
6. A. Dewan et al., *Science*. **314**, 989 (2006).
7. The Wellcome Trust Case Control Consortium, *Nature*. **447**, 661 (2007).
8. International Multiple Sclerosis Genetics Consortium et al., *Nat Genet*. **45(11)**, 1353 (2013).
9. S. Katz, *Curr Opin Neurol*. **28(3)**, 193 (2015).
10. T. Avsar et al., *PLoS One*. **5;10(5)**, e0122045 (2015).
11. D. Karussis, *J Autoimmun*. **48-49**, 134 (2014).
12. Y. Kamatani et al., *Nat Genet.* **42(3)**, 210 (2010).
13. K. Suhre et al., *Nat Genet.* **43(6)**, 565 (2011).
14. J. Denny et al., *Nat Biotechnol*. **31**, 1102 (2013).
15. S. Pendergrass et al., *Hum Hered.* **79(3-4)**, 111 (2015).
16. S. Pendergrass et al., *PLoS Genet*. **9(1)**, e1003087 (2013).
17. M. Hall et al., *PLoS Genet*. **4;10(12)**, e1004678 (2014).
18. C. Geiger et al., *PLoS Genet*. **4(11)**, e1000282 (2008).
19. E. Stergiakouli et al., *Obesity*. **22(10)**, 2252 (2014).
20. P. Visscher et al., *Am J Hum Genet*. **13;90(1)**, 7 (2012).
21. A. Forrey et al., *Clin Chem*. **42(1)**, 81 (1996).
22. S Bennett et al., *Genet Epidemiol*. **35(3)**, 159 (2011).
23. J. Deckard et al., *J Am Med Inform Assoc.* **22(3)**, 621 (2015).
24. S. Mitchell et al., *Stroke.* **32**, 842 (2001).
25. M. Farhangi et al., *J Health Popul Nutr.* **31(1)**, 58 (2013).
26. L. Ferreira et al., *Rev. Bras. Hematol. Hemoter*. **35**, 3 (2013).

A POWERFUL METHOD FOR INCLUDING GENOTYPE UNCERTAINTY IN TESTS OF HARDY-WEINBERG EQUILIBRIUM

ANDREW BECK

Department of Biostatistics, University of Michigan
Ann Arbor, MI, 48109 USA
Email: beckandy@umich.edu

ALEXANDER LUEDTKE

Department of Biostatistics, University of California- Berkeley
Berkeley, CA, 94720 USA
Email: aluedtke@berkeley.edu

KELI LIU

Department of Statistics, Harvard University
Cambridge, MI, 02138 USA
Email: kliu@college.harvard.edu

NATHAN TINTLE

Department of Mathematics, Statistics, and Computer Science, Dordt College
Sioux Center, IA 51250, USA
Email: Nathan.Tintle@dordt.edu

The use of posterior probabilities to summarize genotype uncertainty is pervasive across genotype, sequencing and imputation platforms. Prior work in many contexts has shown the utility of incorporating genotype uncertainty (posterior probabilities) in downstream statistical tests. Typical approaches to incorporating genotype uncertainty when testing Hardy-Weinberg equilibrium tend to lack calibration in the type I error rate, especially as genotype uncertainty increases. We propose a new approach in the spirit of genomic control that properly calibrates the type I error rate, while yielding improved power to detect deviations from Hardy-Weinberg Equilibrium. We demonstrate the improved performance of our method on both simulated and real genotypes.

1. Introduction

With recent advances in high-throughput gene sequencing technologies, it is now possible to obtain measurements on millions of single nucleotide variants (SNVs) throughout the human genome. Large scale genetic data sets, whether from microarray, sequencing or imputation, contain genotype uncertainty which, if unaccounted for in downstream analyses, can significantly decrease power to detect disease-variant associations [1,2] if the uncertainty is not associated with the phenotype, or affect the corresponding type I error rate [3,4] if the uncertainty is associated with the phenotype. To minimize the impact of genotype uncertainty, a standard pre-processing step in most studies is to remove markers that are not in Hardy-Weinberg Equilibrium (HWE), since genotyping errors due to factors like DNA contamination and allelic dropout can cause deviation from HWE [5,6].

The standard approach to testing HWE uses a χ^2_{GOF} test whereby observed genotype frequencies at a variant site are used to obtain maximum likelihood estimates (MLEs) of the minor allele frequency (MAF; *f*) at the site. A one degree of freedom χ^2_{GOF} statistic is then computed to test the null hypothesis that the observed genotype frequencies follow HWE, namely $(1-f)^2, 2f(1-f)$ and f^2 for the major homozygote, heterozygote and minor homozygote, respectively. While this version of the test is the most straightforward and widely used, alternatives exist including methods for testing HWE in datasets with excess correlation between subject genotypes [7,8], missing genotypes [9] and those that account for covariates [10].

Recently, another alternative HWE testing approach was proposed, $\chi^2_{Posterior}$ [6], which extends the standard χ^2_{GOF} approach to allow for the incorporation of genotype uncertainty. The method has widespread application since for all common genotyping technologies (SNP microarray technology [11], imputation [12] and next-generation sequencing technology [13,14]), probabilistic genotypes are obtained as part of the standard genotype calling pipeline. Such probabilistic genotypes typically take the form of a vector of three posterior probabilities for each individual at each variant site, representing the posterior probability that the individual is actually each of the three possible genotypes. While standard analysis techniques typically "call" genotypes by summarizing the posterior probability by a single discrete genotype (e.g., mode posterior probability), researchers are increasingly using alternative approaches. For example, researchers may use of the entire vector of posterior probabilities or they may use the expected genotype (dosage) [15]. The simulation results of Zheng et al. [15], which were recently made rigorous [16], demonstrate substantial power loss from the use of the modal genotype in many realistic situations and approximately equivalent power from use of the dosage or the entire vector of posterior probabilities in case-control tests of genetic association. These results underscore the importance of considering HWE testing methods, which incorporate genotype uncertainty via the underlying posterior probabilities.

The traditional χ^2_{GOF} makes the key assumption that genotype counts are non-negative integers at each variant site, an assumption that is violated with the inclusion of probabilistic calls. A recently proposed alternative approach, $\chi^2_{Posterior}$, allows for the incorporation of probabilistic genotypes. However, $\chi^2_{Posterior}$ has been shown to be overly conservative (empirical type I error rate less than nominal) as uncertainty at the variant site increases [6]. In this manuscript, we explore reasons for the conservative nature of $\chi^2_{Posterior}$ and propose an alternative approach to HWE testing which incorporates genotype uncertainty while maintaining the type I error rate at nominal levels. We then evaluate the type I error and power of the new approach across a variety of realistic HWE and non-HWE settings to identify powerful and robust HWE tests for probabilistic genotypes. Finally, we implement the new method on a real data set illustrating its improved ability to maintain the type I error rate, while improving power to detect variants not in HWE.

2. Methods

2.1. Notation

To facilitate the presentation and evaluation of existing and novel approaches to testing for HWE while incorporating genotype uncertainty, we start by defining some basic notation we will use throughout the manuscript. Genotypes for a given individual i can be represented as a vector of three posterior probabilities, $\alpha_i \triangleq (\alpha_{i0}, \alpha_{i1}, \alpha_{i2})$, where α_{ik} is the posterior probability that individual i has k minor alleles, $k = 0,1,2$ at a variant site of interest. The vector of posterior probabilities, α_i, suggests that the true minor allele count for individual i, denoted $x_i \in 0,1,2$, can be modeled as being a single random draw from a multinomial distribution with probabilities indicated by α_i. We assume that α_i is available for each individual.

2.2. Existing approaches to incorporating genotype uncertainty

The most straightforward and widely used approach to manage genotype uncertainty is to summarize the vector of posterior probabilities α_i with the modal genotype, namely, $m_i \triangleq \arg\max_k(\alpha_i)$ in place of the individuals true genotype. When the modal genotype is used as the true genotype, a standard χ^2 goodness of fit test can be used to test for HWE (χ^2_{Mode}). However, when using such a method we expect an increase in the type I error rate and/or decrease in power due to the introduction of genotype errors caused by ignoring the genotype uncertainty represented in the posterior probabilities vector [2,6]. For example, if $\alpha_{i0} = 0.95$ (the mode), we "call" the individual as having no rare alleles and, thus, there is a 5% chance we are incorrect.

A recently proposed test for HWE, $\chi^2_{Posterior}$, utilizes the entire vector of posterior probabilities [6]. This method starts by computing three, non-discrete, genotype counts based on α_i: $A^*_0 = \sum_{i=1}^{N} \alpha_{i0}$, $A^*_1 = \sum_{i=1}^{N} \alpha_{i1}$, and $A^*_2 = \sum_{i=1}^{N} \alpha_{i2}$, where N is the total sample size and we use A^* to represent genotype counts computed by summing the posterior probabilities across the sample. This approach applies a standard χ^2 goodness of fit test as follows

$$\chi^2_{Posterior} = \chi^2_{GOF}(A^*) = N\left[\frac{\left|\frac{A^*_0}{N}-(1-\hat{f})^2\right|-c/N}{(1-\hat{f})^2} + \frac{\left|\frac{A^*_1}{N}-2(1-\hat{f})\hat{f}\right|-c/N}{2(1-\hat{f})\hat{f}} + \frac{\left|\frac{A^*_2}{N}-(\hat{f})^2\right|-c/N}{(\hat{f})^2}\right] \quad (1)$$

where c is a continuity correction, e.g. 0.5 [17], and where the maximum likelihood estimate (MLE) of the minor allele frequency (MAF), $\hat{f}$, at the site is estimated as $\frac{A^*_1 + \mathbf{2}A^*_2}{\mathbf{2N}}$. The test uses as its null hypothesis that the variant site is in HWE, and as the alternative hypothesis that the variant site is not in HWE. This approach uses a central χ^2 distribution with a single degree of freedom as the null distribution for $\chi^2_{Posterior}$.

2.3. Direct likelihood approach

As shown via simulation in prior work [6], and confirmed in our simulations (see *Results*), the $\chi^2_{Posterior}$ test has an overly conservative type I error rate, which becomes more pronounced as genotype uncertainty increases. We now argue that the reason for this overly conservative type I error rate is due to a change in the covariance structure of the genotypes when using probabilistic genotypes (α_i). In particular, the $\chi^2_{Posterior}$ test assumes that each individual genotype occurs according to a multinomial distribution. However, this is no longer the case when observed genotype counts are obtained by summing over the posterior probability vectors [18]. Thus, the covariance structure assumed by the $\chi^2_{Posterior}$ test is not true in practice when using probabilistic genotypes. In situations where the alternative covariance structure due to probabilistic genotypes can be explicitly modeled or otherwise controlled for, likelihood based approaches to testing with uncertain genotypes are possible [15,18]. However, that is not the case for HWE testing, as we explain in the following paragraph.

In particular, in order to develop a likelihood ratio test you must have an explicit expression for the likelihood function of the population genotype frequencies, $G_0, G_1, and\ G_2$. Here the likelihood function can be written as $L(G_0, G_1, G_2; \alpha_1, \dots, \alpha_N) = P(\alpha_1, \dots, \alpha_N | G_0, G_1, G_2) = P(\alpha_1, \dots, \alpha_N | g_1, \dots, g_N, G_0, G_1, G_2) P(g_1, \dots, g_N | G_0, G_1, G_2)$, where g_i indicates the true genotype of individual i. Thus, you must have knowledge of the true uncertainty mechanism, $P(\alpha_1, \dots, \alpha_N | g_1, \dots, g_N, G_0, G_1, G_2)$ in order to develop a likelihood ratio test based on the posterior probabilities alone. Because explicit knowledge of the true uncertainty mechanism is unlikely, a likelihood approach to HWE testing using $\alpha_1, \dots, \alpha_N$ will not be possible without making unwarranted assumptions.

2.4. Alternative approach

Because of the overly conservative nature of existing approaches and the limitations we describe above when deriving an explicit likelihood approach, we present an alternative strategy: a post-hoc empirical correction in the spirit of genomic-control. Genomic control [19] is a widely-utilized post-hoc correction factor in genome-wide association studies. When systematic inflation of SNP-association statistics occurs in the data, which can occur due to population stratification or differential genotyping errors, dividing the distribution of observed chi-squared statistics by the median observed chi-squared statistic properly controls the empirical type I error rate. Essentially, this approach assumes that when testing thousands of variant sites for association with the phenotype, the vast majority of sites will not be associated with the phenotype. Thus, the observed distribution of test statistics, aside from the extreme upper-tail, can, in essence, be used as its own null distribution.

To extend the notion of genomic control to HWE testing, we argue that in most real testing situations, the majority of variant sites in a sample of many thousands of variants will be in HWE. Thus, we propose computing $\chi^2_{Posterior,j}$ from A^* as shown above for all variants of interest, *j*=1,…,*m*, where *m* is large. Then the measure of inflation/deflation in the null distribution of test statistics is computed as $\hat{\lambda} = \frac{median(\chi^2_{GOF,1}, \chi^2_{GOF,2}, \ldots, \chi^2_{GOF,m})}{median(\chi^2_1)}$, where $median(\chi^2_1) = 0.455$ [19]. The genomic control-like test statistic for HWE is then computed as $\chi^2_{GC,j} = \frac{\chi^2_{GOF,j}}{\hat{\lambda}}$ for all j=1,…,*m*. We consider four different versions of χ^2_{GC}: $\chi^2_{GC,overall}$, $\chi^2_{GC,MAF}$, χ^2_{GC,r^2} and χ^2_{GC,MAF,r^2}, where $\hat{\lambda}$ is computed on different subsets of the data. Overall indicates that $\hat{\lambda}$ is computed across all *m* SNPs in the set. MAF indicates that $\hat{\lambda}$ is computed separately by MAF group (0.05-0.10, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5). r^2 indicates that $\hat{\lambda}$ is computed separately by r^2 group (0-0.5, 0.5-0.75, 0.75-0.85, 0.85-0.95 and 0.95-1), where r^2 is a measure of genotype uncertainty- see next section for details. And, MAF, r^2 computes $\hat{\lambda}$ in groups defined by both MAF and r^2 (25 separate groups).

2.5. Simulation

We simulated genotype data in order to explore the performance of our proposed new approach under a wide variety of situations. We simulated approximately 850,000 SNPs where HWE was maintained (HWE SNPs). To ensure that the characteristics of this simulation reflected both a realistic allele frequency distribution as well as genotype uncertainty, we randomly sampled (*f*, r^2) pairs with replacement from a large dataset of genotypes from the FUSION study [20] that were imputed using MaCH [12]. For each (*f*, r^2) pair, we then simulated the ‘real’ genotypes of 10,000 individuals according to the specified allele frequency, *f*, assuming the population was in Hardy-Weinberg Equilibrium (HWE) $((1-f)^2, 2f(1-f), f^2)$. To model genotype uncertainty at the appropriate level, r^2, we drew from one of the following Dirichlet distributions conditional on the true genotype [16].

If g_i = 2 then $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2}) \sim Dirichlet(aq^2, 2aq(1-q), a(1-q)^2)$
If g_i = 1 then $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2}) \sim Dirichlet(aq(1-q), a(1-q)^2 + aq^2, aq(1-q))$
If g_i = 0 then $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2}) \sim Dirichlet(a(1-q)^2, 2aq(1-q), aq^2)$
for $a > 0$ and $0 < q < 1$, where a and q are chosen to yield a desired r^2 value. This model is chosen to simulate symmetric noise in posterior probabilities while maintaining HWE. Further details are available in *Appendix #1* and elsewhere [16]. In short, parameter q is the “average” amount of error. For example, if *q*=0.05 (5% noise/error level in posterior probabilities), then for the major homozygote, g_i = 2, $E(\alpha_i) = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2})$ = (0.9025, 0.0.95, 0.0025) and, likewise, if there is no noise/error (*q*=0), then $\alpha_i = (0,0,1)$. Parameter a is the variation in the error from person to person. For example, as a increases, then $Var(\alpha_i)$ also increases, and so for very small values of a (e.g., *a*=0.01), there is virtually no variation in the values of α_i from person to person.

We also simulated three sets, each with approximately 75,000 SNPs, that were not in HWE (non-HWE SNPs). To do this we randomly sampled two SNPs (*i* and *j*) that were in HWE from the set of 850,000 SNPs described above, keeping track of the difference in the allele frequencies of the two SNPs, $d_{i,j}=f_i-f_j$. We then randomly sampled *n*(1-*k*) individuals from SNP *i* and *nk* individuals from SNP *j*, combining the individuals into a single sample of *n* individuals. We used values of *k*=0.1, 0.3 and 0.5, and continued to use a total sample size of 10,000. Thus, the resulting sample is not in HWE because the observed genotype frequencies were generated from two subpopulations with different allele frequencies.

The three resulting sets of 75,000 simulated SNP genotypes were analyzed using (a) a standard HWE test on the simulated 'real' genotypes (χ^2_{True}), (b) chi-squared on the modal genotype χ^2_{Mode}, (c) the approach utilizing posterior probabilities ($\chi^2_{Posterior}$) and (d) four different GC-like approaches (χ^2_{GC}; see previous section for details). For the purposes of the GC-like approach we combined random subsets of 25,000 non-HWE SNPs with the 850,000 HWE SNPs and applied the adjustment, keeping the total proportion of non-HWE SNPs in the set below 3%.

Type I error rates were computed on the 850,000 HWE SNPs as the proportion of SNPs that were detected to be 'not in HWE' at a particular significance level and for a particular combination of MAF and r^2 levels. Power was computed as the fraction of non-HWE SNPs with a p-value less than the significance level in 300 separate groups created by values of *k (*0.1, 0.2, 0.5*),* difference in MAF between the two SNPs being mixed together (0.1, 0.1-0.2, 0.2-0.3 or >0.3), observed MAF of the combined variant (0.05-0.10, 0.10-0.20, 0.20-0.30, 0.30-0.40 and 0.40-0.50) and observed r^2 of the combined variant (0-0.50, 0.50-0.75, 0.75-0.85, 0.85-0.95 and 0.95-1.0). We examined significance levels of 0.01, 1x10^{-3}, and 1x10^{-5}. We computed power and type I error rates across a variety of subsets of the variants including minor allele frequency, genotype uncertainty (r^2), and deviation from HWE.

2.6. Real data analysis - FUSION

As a proof of concept, we ran χ^2_{Mode}, $\chi^2_{Posterior}$ and χ^2_{GC,MAF,r^2} on 29,361 SNPs imputed with MaCH from chromosome 21 of the FUSION study (n=2456) [20]. We also created 2,377 new variants based on the 29,361 imputed variants, which were out of Hardy-Weinberg equilibrium. These 2,377 new variants were created by first randomly selecting two variants with differences in minor allele frequency of between 0.1 and 0.2 and r-squared values between 0.75 and 0.85. A new variant is created by randomly selecting 10% of the genotypes from one of the variants and 90% from the other. All three Hardy-Weinberg equilibrium tests (χ^2_{Mode}, $\chi^2_{Posterior}$ and χ^2_{GC,MAF,r^2}) were also applied to the 2,377 new non-HWE variants as well. We used a significance level of 1x10^{-5} on the 29,361 real and 2,377 new FUSION variants.

Table 1. Overall type I error rates

Method	**Significance level**		
	0.01	**0.001**	**1x10^{-5}**
$\chi^2_{Posterior}$	0.0067	0.00057	3.5x10^{-6}
χ^2_{Mode}	0.0134	0.00166	2.6x10^{-5}
$\chi^2_{GC,overall}$	0.0112	0.00127	2.2x10^{-5}
$\chi^2_{GC,MAF}$	0.0112	0.00128	2.3x10^{-5}
χ^2_{GC,r^2}	0.0104	0.0011	1.3x10^{-5}
χ^2_{GC,MAF,r^2}	0.0101	0.00105	1.2x10^{-5}
χ^2_{True}	0.0099	0.00097	8.1x10^{-6}

3. Results

3.1. Type I error simulation

Table 1 gives the overall type I error rates at three different significance levels for each of the six methods applied to posterior probabilities on SNPs in HWE, along with the significance level when using the true genotypes. As expected, use of the true genotypes yields type I error rates at the significance level. Overall, $\chi^2_{Posterior}$ yielded the most conservative type I error rates, while χ^2_{Mode} yielded anti-conservative type I error rates. The χ^2_{GC} corrected approaches tended to yield approximately correct type I error rates, with the version which adjusts statistics both within *MAF* and r^2 (χ^2_{GC,MAF,r^2}) bins providing the best Type I error control. A logistic regression model predicting the type I error rate $\chi^2_{Posterior}$ test across all

850,000 SNPs indicates that both MAF and r^2, as well as an interaction term between MAF and r^2, are significant predictors of the type I error rate, which further supports the necessity to use both bins for both MAF and r^2 when correcting statistics as is done by χ^2_{GC,MAF,r^2}.

The patterns observed in Table 1 remain true across all MAF and r^2 subgroups as shown in Supplemental Table 1. In particular we also see that $\chi^2_{Posterior}$ is the most conservative for less well imputed SNPs, though even well imputed SNPs are treated anti-conservatively by $\chi^2_{Posterior}$ ($8.5x10^{-3}$ for r^2>0.95). In contrast, χ^2_{Mode} is the most anti-conservative for less well imputed SNPs, with some inflation of the type I error rate for moderately well imputed SNPs (e.g., 0.85<r^2<0.95). χ^2_{Mode} only controls the type I error rate for extremely well imputed SNPs (r^2>0.95). χ^2_{GC,MAF,r^2} controls the Type I error rate across MAF and r^2 strata. While Supplemental Table 1 only shows results for a significance level of 0.01, patterns remain the same across other more stringent significance levels (e.g., 0.001, $1x10^{-5}$, detailed results not shown). Figure 1 illustrates the anti-conservative performance of χ^2_{Mode}, the conservative performance of $\chi^2_{Posterior}$ and good control of the type I error rate by χ^2_{GC,MAF,r^2}

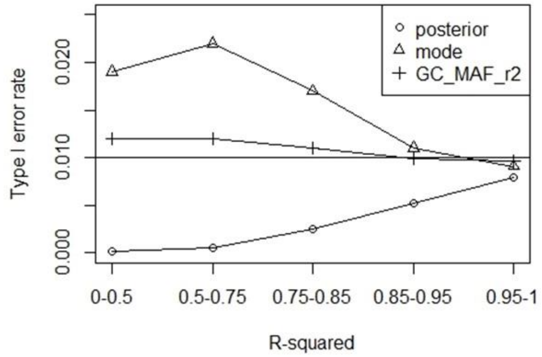


Figure 1. Type I error rate for three different HWE testing methods across different uncertainty levels. Type I error rate is shown across different r^2 settings for three different HWE testing approaches at the 1% significance level. SNPs in the low minor allele frequency range are depicted (MAF between 0.05 and 0.1)

Power simulation

To understand the power of the different approaches for HWE testing, we considered 300 combinations of average minor allele frequency across SNPs *i* and *j*, observed r^2, difference in minor allele frequency and *k* (proportion of individuals from SNP *i*; where *1-k* is the proportion of individuals are from SNP *j*) across 225,000 SNPs which are a mixture of two different allele frequencies. One-hundred twenty-two of the settings yielded 100% power when using all methods, and another 40 combinations yielded no SNPs, and so these 162 settings are eliminated from further consideration. Due to the fact that χ^2_{Mode} has an inflated Type I error rate, we do not consider it in the following comparative analysis of the power of the different methods. Across these 162 settings the median number of SNPs per group was 490 (Min=5; Q1=182; Q3=1708; Max=4353), with only three settings having less than 20 SNPs.

Across the 138 remaining combinations of settings, χ^2_{GC,MAF,r^2}had higher power than $\chi^2_{Posterior}$ 122 times, by an average of 0.038 (SD=0.039). Across the 16 times that$\chi^2_{Posterior}$ yielded higher power than χ^2_{GC,MAF,r^2}, the average power gain was only 0.0029 (SD=0.0024). Table 2 illustrates a subset of 138 simulation settings, illustrating that χ^2_{GC,MAF,r^2} consistently yields higher power than $\chi^2_{Posterior}$ for all but the most certain SNPs, when performance is comparable. Largest gains in power were for the least certain

Table 2. Power[1] by MAF and r^2

MAF	r^2	Number of variants	$\chi^2_{Posterior}$	χ^2_{GC,MAF,r^2}	χ^2_{True}
0.05-0.1	0-0.50	122	0.91	0.98	1
	0.5-0.75	265	0.82	0.94	0.98
	0.75-0.85	166	0.79	0.84	0.98
	0.85-0.95	480	0.86	0.87	0.99
	0.95-1.0	695	0.85	0.85	0.99
0.1-0.2	0-0.50	123	0.67	0.84	0.86
	0.5-0.75	382	0.65	0.78	0.85
	0.75-0.85	441	0.66	0.76	0.82
	0.85-0.95	1411	0.62	0.65	0.82
	0.95-1.0	2561	0.62	0.61	0.81
0.2-0.3	0-0.50	152	0.53	0.66	0.7
	0.5-0.75	365	0.52	0.58	0.7
	0.75-0.85	489	0.56	0.67	0.74
	0.85-0.95	2029	0.53	0.57	0.72
	0.95-1.0	4217	0.52	0.51	0.7
0.3-0.4	0-0.50	81	0.43	0.51	0.57
	0.5-0.75	209	0.39	0.46	0.52
	0.75-0.85	277	0.4	0.52	0.58
	0.85-0.95	1324	0.36	0.41	0.54
	0.95-1.0	3321	0.38	0.38	0.53
MAF>0.4	0-0.50	25	0.32	0.32	0.44
	0.5-0.75	87	0.29	0.36	0.55
	0.75-0.85	160	0.33	0.43	0.48
	0.85-0.95	629	0.37	0.41	0.51
	0.95-1.0	1649	0.38	0.38	0.52

1. At the 1% significance level and when the observed SNP is a mix of two subgroups of individuals with a difference of between 0.10 and 0.20 in minor allele frequency between the two subgroups, and 10% of the individual are from one subgroup and 90% from the other (*k*=0.1).

SNPs, with overall higher power for all methods with lower MAF. Figure 2 illustrates this relative gain in power. Supplementary Table 1 gives the full power results for all 300 settings.

Real data example

When applying the three HWE testing methods to the 29,361 imputed FUSION SNPs, 237 variants were determined to be out of HWE by χ^2_{Mode}, none by $\chi^2_{Posterior}$ and two by χ^2_{GC,MAF,r^2} at a significance level of 1x10^{-5}. While true HWE status for these variants is unknown, these results suggest an inflated type I error rate for the χ^2_{Mode} test. When we applied the $\chi^2_{Posterior}$ and χ^2_{GC,MAF,r^2} tests to the 2,377 non-HWE variants, the power was always higher for the χ^2_{GC,MAF,r^2} test (see Table 3).

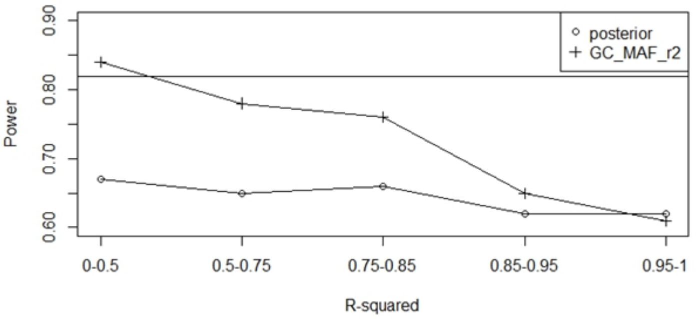


Figure 2 Power for two different approaches to HWE testing across different uncertainty levels
Power is illustrated across different r^2 settings for two different HWE testing approaches at the 1% significance level, with a horizontal line at the power of a test using the real genotypes. Power for SNPs with MAF between 0.1 and 0.2 are depicted, when the observed SNP is a mix of two subgroups of individuals where the difference in MAF between the two subgroups is between 0.1 and 0.2, and the 10% of the individuals are from one subgroup and 90% from the other.

Table 3. Power to detect pseudo variants not in Hardy-Weinberg Equilibrium from the FUSION study

Observed MAF	Number of variants	$\chi^2_{Posterior}$	χ^2_{GC,MAF,r^2}
0.05-0.10	375	5.3%	8.0%
0.10-0.20	731	28.6%	29.7%
0.20-0.3	412	28.9%	30.8%
0.3-0.4	385	26.8%	32.2%
0.4-0.5	374	2.9%	5.9%
Overall	2277	20.3%	22.8%

4. Discussion

We have proposed a new way to incorporate posterior probabilities in tests of HWE that provides a well-calibrated and more powerful way to incorporate genotype uncertainty. While it is common to use the modal posterior genotype, this approach inflates the type I error rate by failing to incorporate genotype uncertainty---treating uncertain genotypes as if they are error-free. Furthermore, another recent approach which explicitly incorporates posterior probabilities yields an overly conservative test (deflated type I error rate), due to an overestimation of the covariance of the posterior probability genotypes. Our approach applies a post-hoc correction to adjust the test statistic, yielding a calibrated type I error rate and improved power.

The proposed approach is approximately the same as other approaches when genotype uncertainty is low, but shows increasing benefit as genotype uncertainty increases. This result is in line with the fact that the genotype covariance estimates are increasingly biased when using $\chi^2_{Posterior}$ as genotype uncertainty increases. While it is common practice to simply drop markers with very high genotype uncertainty from analyses we've demonstrated that this may not be necessary when using our approach. Furthermore, even if practitioners wish to drop markers with high genotype uncertainty (e.g., r^2<0.5), we've demonstrated that our approach to HWE testing still outperforms other HWE testing procedures for markers with modest genotype uncertainty (0.5<r^2<0.95). Importantly, recent work has shown that simply screening for HWE using r^2 is not sufficient, and that HWE testing is still necessary [21].

While the proposed approach performs well relative to the existing approaches by applying a post-hoc correction, a more explicit approach may also be possible. Preliminary exploration of such methods by our group has taken two separate paths to date. First, we considered multiple imputation by creating many, equally likely, versions of each individual's genotype according to the vector of calibrated posterior genotype probabilities and then computing the standard chi-squared GOF test on each multiply-imputed dataset. Methods for computing significance from a set of multiply-imputed datasets are standard [22–24], but may not be well-calibrated [25]. A lack of calibration was our experience for this application (detailed results not shown). A second approach is a Bayesian approach using the posterior probabilities for each individual's genotype explicitly. Evaluation of this method across a wide-range of simulation settings showed performance comparable to the $\chi^2_{Posterior}$ method and, thus, not as good as χ^2_{GC,MAF,r^2} in many cases (detailed results not shown).

We now make some important notes and comments on limitations of the χ^2_{GC,MAF,r^2}approach. While not considered here, the authors of the $\chi^2_{Posterior}$ approach also considered an exact test for small sample sizes. Future work is needed to evaluate the performance of the post-hoc correction strategy for small sample size situations (e.g., rare variants), though, in principle, there is no reason to believe that an approach in this same spirit is likely to perform well. A key assumption of χ^2_{GC,MAF,r^2} is that a relative small proportion of all markers overall will not be in HWE. In rare cases where a very large proportion of markers are out of HWE, the χ^2_{GC,MAF,r^2} approach may, in fact, be overly conservative by applying a correction factor based on markers not in HWE. However, these cases should be rare as a substantial portion of the markers in the correction set would need to be out of HWE in order to impact the median observed statistic and, hence, the lambda, in a practically significant way. However, since χ^2_{GC,MAF,r^2} computes a separate adjustment for many different MAF, r^2 'bins,' an aggregation of markers not in HWE in any bin could impact results. Finally, the size and quantity of MAF, r^2 bins selected in this study showed good performance, but may need adjustment in practice based on the MAF distribution, r^2 (or other uncertainty metric) distribution and number of variants. Care should be taken to ensure all bins have sufficient markers (generally recommended to be at least 100, but less may be fine) and examination of $\hat{\lambda}$ values within each bin is recommended. Future work may wish to explore the potential for a robust, continuous correction strategy.

Supplemental Files

All supplemental and appendix files are available online at the following URL: http://homepages.dordt.edu/ntintle/hwe.zip

Acknowledgments

This work was funded by the National Human Genome Research Institute (R15HG006915). We acknowledge the use of the Hope College parallel computing cluster for assistance in data analysis. We are grateful for the generosity of the FUSION investigators for access to imputed genotypes on which we based our simulations and demonstrated a proof-of-concept of our proposed approach. The FUSION study is funded by the National Institute of Diabetes and Digestive and Kidney Diseases (U01DK062370).

References

1. Powers S, Gopalakrishnan S, Tintle N. Assessing the impact of non-differential genotyping errors on rare variant tests of association. Hum Hered. 2011;72: 153–60.

2. Gordon D, Finch SJ. Factors affecting statistical power in the detection of genetic association. J Clin Investig. 2005;115.

3. Mayer-Jochimsen M, Fast S, Tintle NL. Assessing the impact of differential genotyping errors on rare variant tests of association. PLoS One. 2013;8: e56626.

4. Moskvina V, Craddock N, Holmans P, Owen MJ, O'Donovan MC. Effects of differential genotyping error rate on the type I error probability of case-control studies. Hum Hered. 2006;61: 55–64.

5. Wang J, Shete S. Testing Hardy-Weinberg proportions in a frequency-matched case-control genetic association study. PLoS One. 2011;6: e27642.

6. Shriner D. Approximate and exact tests of Hardy-Weinberg equilibrium using uncertain genotypes. Genet Epidemiol. 2011;35: 632–7.

7. Li Y. A comparison of tests for Hardy-Weinberg Equilibrium in national genetic household surveys. BMC Genet. 2013;14: 14.

8. She D, Zhang H, Li Z. Testing Hardy-Weinberg equilibrium using family data from complex surveys. Ann Hum Genet. 2009;73: 449–55.

9. Graffelman J, Nelson S, Gogarten SM, Weir BS. Exact Inference for Hardy-Weinberg Proportions with Missing Genotypes: Single and Multiple Imputation. G3 Genes|Genomes|Genetics. 2015;5: g3.115.022111.

10. Schaid DJ, Sinnwell JP, Jenkins GD. Regression Modeling of Allele Frequencies and Testing Hardy Weinberg Equilibrium. Hum Hered. 2013;74: 71–82.

11. Hong H, Su Z, Ge W, Shi L, Perkins R, Fang H, et al. Evaluating variations of genotype calling: a potential source of spurious associations in genome-wide association studies. J Genet. 2010;89: 55–64.

12. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: Using sequence and genotype data to esimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34: 816–834.

13. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18: 1851–1858.

14. Nielson R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12: 443–451.

15. Zheng J, Li Y, Abecasis GR, Scheet P. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genet Epidemiol. 2011;35: 102–10.

16. Liu K, Luedtke A, Tintle NL. optimal methods for using posterior probabilities in association testing. Hum Hered. 2013;75: 2–11.

17. Yates F. Contingency table involving small numbers and the X2 test. Suppl to J Roayl Stat Soc. 1934;1: 217–235.

18. Tintle N, Gordon D, McMahon F, Finch SJ. Using Duplicate Genotyped Data in Genetic Analyses : Testing Association and Estimating Error Rates. Stat Appl Genet Mol Biol. 2007;6.

19. Devlin B, Roeder K. Genomic Control for Association Studies. Biometrics. 1999;55: 997–1004.

20. Scott L, Mohlke K, Bonnycastle L, Willer C, Li Y, Duren W, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science (80-). 2007;316: 1341–5.

21. Shriner D. Impact of Hardy-Weinberg disequilibrium on post-imputation quality control. Hum Genet. 2013;132: 1073–5.

22. Li K-H, Meng X-L, Raghunathan TE, Rubin DB. Signifiacnce levels from repeated p-values with multiply imputed data. Stat Sin. 1991;1: 65–92.

23. Meng X-L, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. Biometrika. 1992;79: 103–111.

24. Harel O, Zhou X-H. Multiple imputation: review of theory, implementation and software. Stat Med. 2007;26: 3057–3077.

25. Licht C. New methods for generating significance levels from multiply-imputed data Ph.D. Dissertation. University of Bamberg, Germany. 2010.

TEMPORAL ORDER OF DISEASE PAIRS AFFECTS SUBSEQUENT DISEASE TRAJECTORIES: THE CASE OF DIABETES AND SLEEP APNEA

METTE K. BECK
Novo Nordisk Foundation Center for Protein Research,
University of Copenhagen, Blegdamsvej 3B
Copenhagen, DK-2200, Denmark
Email: mette.beck@cpr.ku.dk

DAVID WESTERGAARD
Novo Nordisk Foundation Center for Protein Research,
University of Copenhagen, Blegdamsvej 3B
Copenhagen, DK-2200, Denmark
Email: david.westergaard@cpr.ku.dk

ANDERS BOECK JENSEN
Novo Nordisk Foundation Center for Protein Research,
University of Copenhagen, Blegdamsvej 3B
Copenhagen, DK-2200, Denmark
Email: anders.b.jensen@cpr.ku.dk

LEIF GROOP
Lund University Diabetes Centre, Department of Clinical Sciences,
Jan Waldenströms gata 35, SE-205 02 Malmö, Sweden
Email: Leif.Groop@med.lu.se

SØREN BRUNAK
Novo Nordisk Foundation Center for Protein Research,
University of Copenhagen, Blegdamsvej 3B
Copenhagen, DK-2200, Denmark
Email: soren.brunak@cpr.ku.dk

Most studies of disease etiologies focus on one disease only and not the full spectrum of multimorbidities that many patients have. Some disease pairs have shared causal origins, others represent common follow-on diseases, while yet other co-occurring diseases may manifest themselves in random order of appearance. We discuss these different types of disease co-occurrences, and use the two diseases “sleep apnea” and “diabetes” to showcase the approach which otherwise can be applied to any disease pair. We benefit from seven million electronic medical records covering the entire population of Denmark for more than 20 years. Sleep apnea is the most common sleep-related breathing disorder and it has previously been shown to be bidirectionally linked to diabetes, meaning that each disease increases the risk of acquiring the other. We confirm that there is no significant temporal relationship, as approximately half of patients with both diseases are diagnosed with diabetes first. However, we also show that patients diagnosed with diabetes before sleep apnea have a higher disease burden compared to patients diagnosed with sleep apnea before diabetes. The study clearly demonstrates that it is not only the diagnoses in the patient’s disease history that are important, but also the specific order in which these diagnosis are given that matters in terms of outcome. We suggest that this should be considered for patient stratification.

1. Introduction

Much epidemiological research has focused on simple associations between two diseases. Temporal approaches have been suggested to uncover both causal and genetic links among statistically associated diseases [1-4]. Many recent studies have analyzed more complicated relations between several diseases and have found bidirectional relationships, where one disease increases the risk or severity of the other or vice versa [1–4]. This type of relationship is mostly found for pairs of common diseases such as depression, cardiovascular diseases and diabetes [2,4]. In one example Mezuk et al. reported a 15% increased risk of depression in patients with type 2 diabetes (T2D), but 60% increased risk of developing type 2 diabetes in patients with depression [5]. Since then several papers have confirmed this particular bidirectional observation [6,7]. Similarly, diabetes has been bidirectionally linked with both periodontitis and sleep apnea [1,8,9].

Until now there has not been general studies investigating the effect of the temporal order in which bidirectionally linked diseases are diagnosed, and how the order affects the further disease progression and the general health status of the patients. In this study we highlight the importance of the temporal order using the bidirectionally linked disease pair: diabetes and sleep apnea. Subsequently we generalize this method to a disease-spectrum wide approach for T2D patients.

Sleep apnea is the most common sleep-related breathing disorder, affecting up to 10% of middle-aged women and up to 20% of middle-aged men in high-income and Asian countries [10–12]. It is traditionally stratified into obstructive sleep apnea and central sleep apnea, where obstructive sleep apnea is the most prevalent subgroup that accounts for up to 85% of sleep apnea patients [13–15]. Furthermore, sleep apnea can occur in both children and adults, although these are treated as two different diseases [16–19]. Untreated, sleep apnea increases the risk for cardiovascular, metabolic, and neurocognitive complications and it is therefore a prototypical example of a disease involved in comorbidities [20,21]. Specifically, it is associated with T2D [1,9,22,23].

Although obesity is a predictor of both obstructive sleep apnea and T2D, the bidirectional link between these diseases appears to be independent of weight [1,9,20]. T2D contributes to sleep apnea, by causing neuromyopathy, which impairs reflexes of the upper airway [9,20]. Sleep apnea contributes to the development of T2D by increased activation of the sympathetic nervous system leading to increased insulin resistance [22,24,25]. It has even been suggested that successful treatment of sleep apnea may reduce the risk of T2D, although this is still controversial [9].

To investigate the effect of the order of the diagnoses we combined the Danish National Patient Registry (NPR), which covers all hospital encounters, both public and private, in Denmark from 1994 to 2015, a patient population of nearly seven million individuals with prescription data from the Danish diabetes registry. NPR records diseases using the International Classifications of Diseases, 10th revision (ICD-10), which organizes diseases hierarchically.

Using this unbiased, country-wide data set we describe the comorbidity map of sleep apnea patients in a data driven manner, and show that the diagnostic order of sleep apnea and T2D is close to 50:50. Interestingly, while the order overall appears to be random we show that the order is associated with significantly different frequencies of comorbid diseases, implying two distinct patient groups.

T2D is a chronic disease with a high risk of many servere complications, including cardiovascular, neurological and infectious complications [5,6,26–30]. Consequently, we generalized our approach to all diseases appearing together with T2D. We showed that the disease burden was dependent on the diagnosis order for twelve T2D comorbidities, of which ten show an increase in comorbidities if T2D was diagnosed first.

2. Materials and methods

In this retrospective cohort study we investigated the association between sleep apnea and T2D. We used the NPR, covering all hospital encounters in Denmark from 1994 to 2015, from where we could include 6,923,849 Danish subjects. Specifically, this registry contained 218,750 T2D patients and 95,853 sleep apnea patients.

To define T2D patients we combined the NPR with the Danish Diabetes registry, which contains medical prescription data. We defined T2D patients, as patients diagnosed at least two times with NIDDM but not with IDDM, if oral hypoglycemic agents were prescribed at least two times and they were diagnosed with NIDDM, or if oral hypoglycemic agents and insulin were prescribed at least two times and they were diagnosed with NIDDM and/or IDDM.

Adult sleep apnea patients were defined as patients first diagnosed with sleep apnea at the age of 16 years or older.

2.1. *Comorbidity calculations*

We tested for significant associations between all level three diagnoses in the ICD-10 terminology. The relative risk of a particular disease was calculated using the Cochran–Mantel–Haenszel method, where each bin corresponds to patients of a particular gender and born in a particular decade. We included patients born from 1900 until 2015, giving rise to up to 24 bins per test. We used the Cochran–Mantel–Haenszel test to identify the p-value and accepted results with a Benjamini-Hochberg corrected p-value of 0.05 or below. This method was used both for time-independent and time-dependent analyses.

2.2. *From temporal diagnosis pairs through disease trajectories to disease network*

The method for identifying the trajectories has been described previously in detail [33]. The method consists of three steps: First temporal directed pairs of co-morbid diseases were tested to identify pairs where one disease is associated with an increase in the occurrences of the other. In the second step, the pairs found are tested for directionality (one disease primarily occurring before the other) using a binomial test. Third, the pairs with significant temporality were combined into disease trajectories of three consecutive diseases. Trajectories were only included if at least 100 sleep apnea patients followed them.

2.3. *Difference in mean number of comorbidities*

The difference in mean number of comorbidities was modeled by a Poisson regression using the covariates: years between the two diagnoses, which disease was diagnosed first, age and gender. All four covariates significantly contributed to the model. This Poisson regression was subsequently used to predict the number of comorbidities for all patients to avoid age and/or gender bias. The difference in mean predicted number of comorbidities was tested using student's t-test, stratified by the order of the diagnoses. This was done twenty times, requiring a minimum from zero years up to nineteen years in between the two diagnoses.

2.4. *Diabetes comorbidities selection criteria*

We tested if any level three ICD-10 diagnoses were significantly correlated with T2D using the method for comorbidity calculations. For the diagnosis with a significant association and a relative risk above one, we used a binomial test to ensure lack of directionality. We required the 95% confidence interval to be within 45% - 55% (making the diagnostic order close to 50:50). Lastly, we required a minimum of 1,000 T2D patients to have the disease. For the remaining diseases we performed the method described in "Difference in mean number of comorbidities". We required ten time points to be significant.

3. Results

In the Danish population of 6,923,707 patients, we found 117,913 patients diagnosed with sleep disorders (G47), of these 95,853 patients were diagnosed with sleep apnea (G47.3). The age distribution at which these patients were first diagnosed with sleep apnea is shown in Figure 1A. It has two clearly distinct peaks, the first at age three, and the second major peak just after 50 years, supporting that this diagnosis could cover two distinct disease progression patterns. We computed the relative risk (RR) for both the adult onset of sleep apnea (aged 16 or above) and the childhood onset, compared to all other level three ICD-10 diagnoses. Even though both groups of patients are diagnosed with the same diagnosis, their repertoire of comorbidities is very different (Figure 1B), in part due to the difference in age. We therefore excluded childhood onset of sleep apnea, and investigated sleep apnea in the adult population further. Of the 95,853 sleep apnea patients 90,157 were diagnosed in adult patients, 75% of these were males.

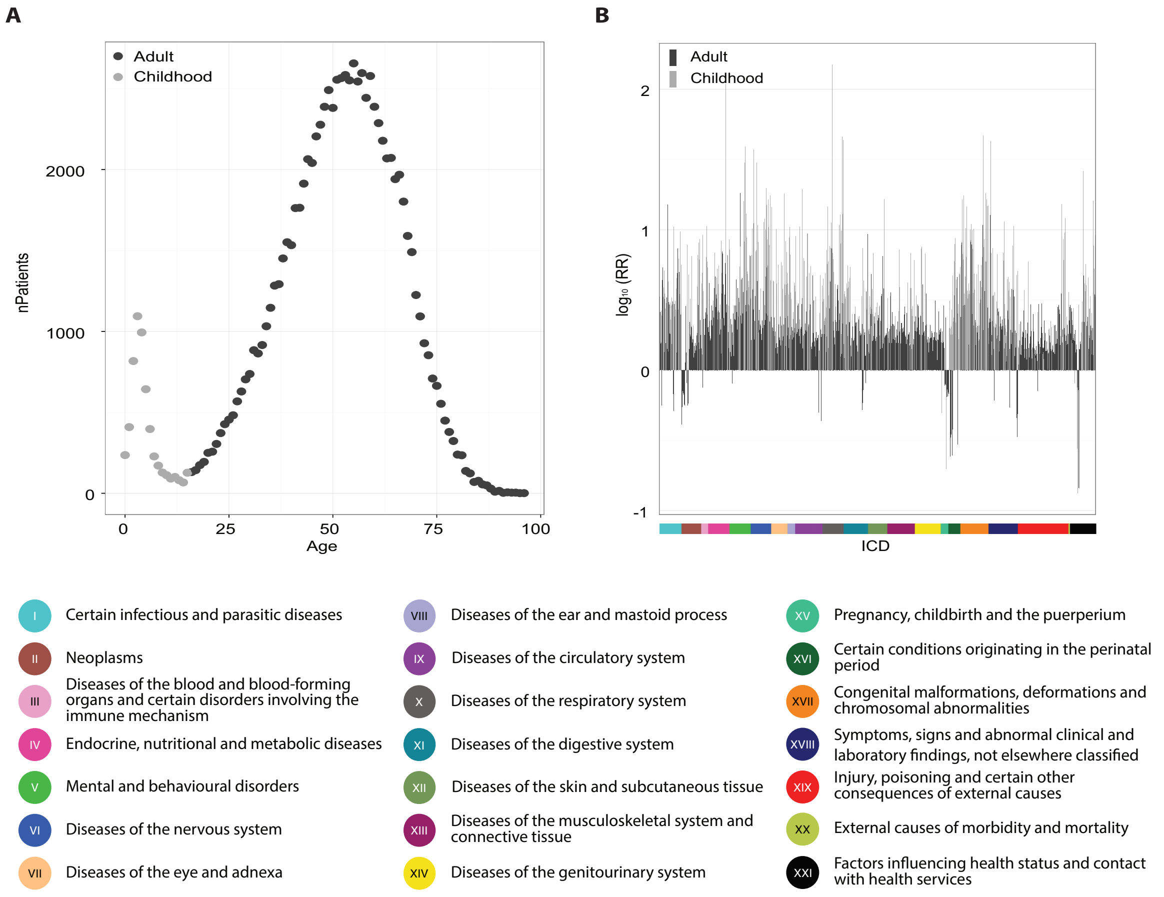


Fig. 3. The increased comorbidity burden for patients diagnosed with T2D before sleep apnea. (A) Distribution of years between T2D and sleep apnea for patients diagnosed with T2D first (pink) and for patients diagnosed with sleep apnea first (blue). (B) The excess number of comorbidities for patients diagnosed with T2D first compared to those diagnosed with sleep apnea first (black line) with the 95% confidence interval (grey area). The x-axis indicates the minimum number of whole years between diagnoses (e.g. 0 years means more than one day but less than a year). The dots indicate the number of patients having minimum x years between the two diagnoses.

3.1. *Temporal disease network reveals no direct connection between diabetes and adult sleep apnea*

We identified all diseases that co-occurred more often than we would expect from their individual frequencies in the patients with adult sleep apnea. For each such disease pair, we testedif one of the diseases occurred significantly more often before the other. This led to the identification of a pool of significant, directed disease-pairs (see Methods). These pairs were combined into linear, temporal disease trajectories of which we found 103 where 100 sleep apnea

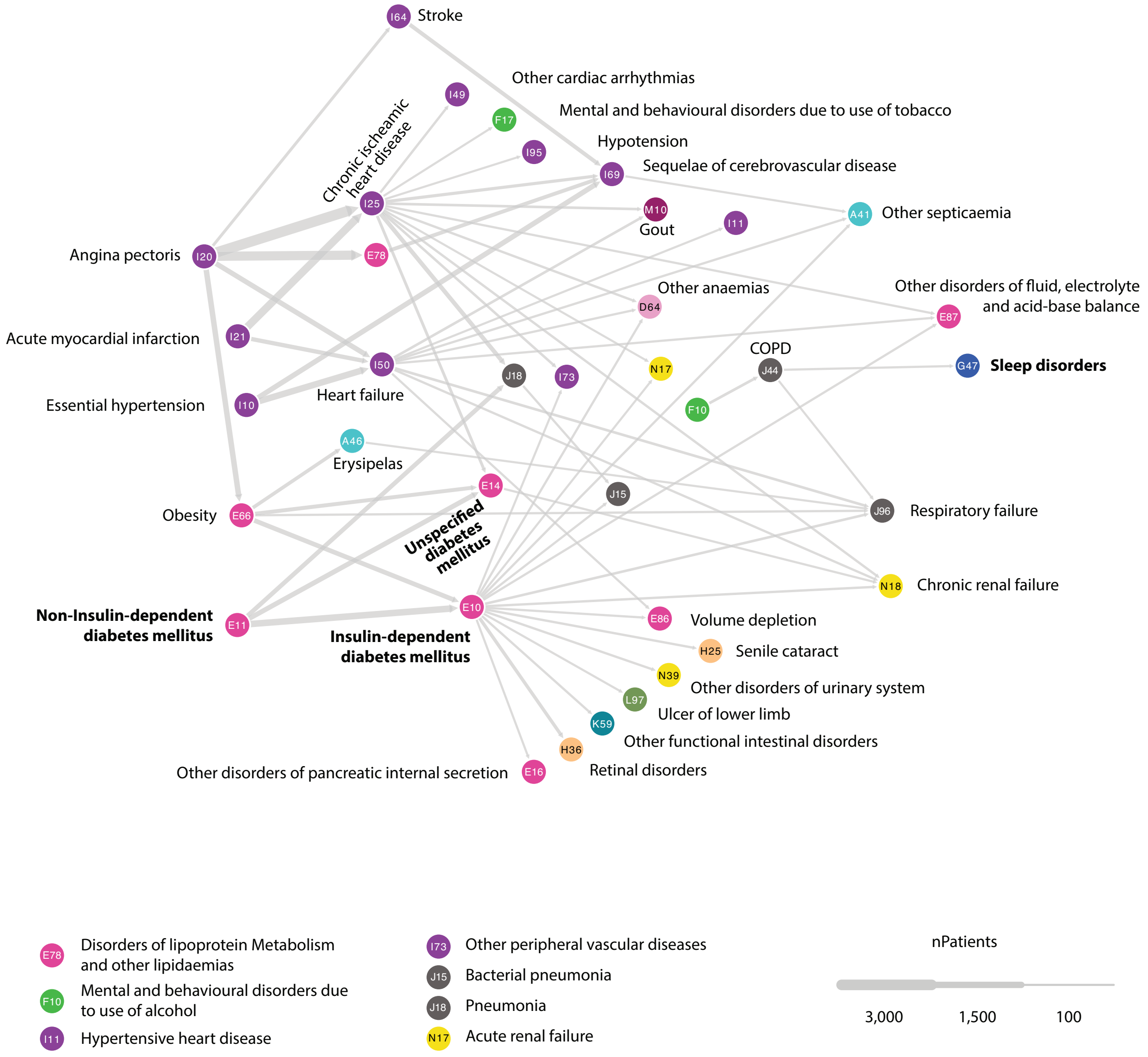

I64 Stroke
Other cardiac arrhythmias
I49
Mental and behavioural disorders due to use of tobacco
F17
Chronic ischeamic heart disease
I95
Hypotension
I69 Sequelae of cerebrovascular disease
I25
M10 Gout
I11
A41 Other septicaemia
Angina pectoris I20
E78
Other anaemias
D64
Other disorders of fluid, electrolyte and acid-base balance
E87
Acute myocardial infarction I21
I50 Heart failure
J18
I73
N17
COPD
J44
G47 Sleep disorders
Essential hypertension I10
F10
A46 Erysipelas
E14
Unspecified diabetes mellitus
J15
Obesity E66
J96 Respiratory failure
N18 Chronic renal failure
Non-Insulin-dependent diabetes mellitus E11
E10 Insulin-dependent diabetes mellitus
E86 Volume depletion
H25 Senile cataract
N39 Other disorders of urinary system
L97 Ulcer of lower limb
K59 Other functional intestinal disorders
H36 Retinal disorders
Other disorders of pancreatic internal secretion E16
E78 Disorders of lipoprotein Metabolism and other lipidaemias
F10 Mental and behavioural disorders due to use of alcohol
I11 Hypertensive heart disease
I73 Other peripheral vascular diseases
J15 Bacterial pneumonia
J18 Pneumonia
N17 Acute renal failure
nPatients
3,000
1,500
100

Fig. 2. Temporal disease network based on sleep apnea patients. The network was constructed from 103 sleep apnea trajectories and illustrates the number of patients taking a particular step in the disease network (width of arrow). The nodes are colored based on their ICD-10 chapter relationships. The names are written next to their node in the network or mentioned in the legend in alphabetical order.

patients followed three consecutive steps of diseases. Subsequently, the 103 linear trajectories found in the adult sleep apnea patient group were combined into a temporal disease network providing a concerted overview of the comorbidity spectrum (Figure 2). As expected, obesity, a known risk factor for sleep apnea, appears as a statistically significant component in this overview network (present in 20 of the 103 temporal trajectories as either starting or midpoint). Several cardio-vascular complications are also prominent in the network. Additionally, both insulin-dependent diabetes mellitus (IDDM) and non-insulin-dependent diabetes mellitus (NIDDM) are part of the disease network along with several diabetes complications. There is no direct path connecting diabetes and sleep disorders in the disease network, due to the lack of temporality between these diagnoses.

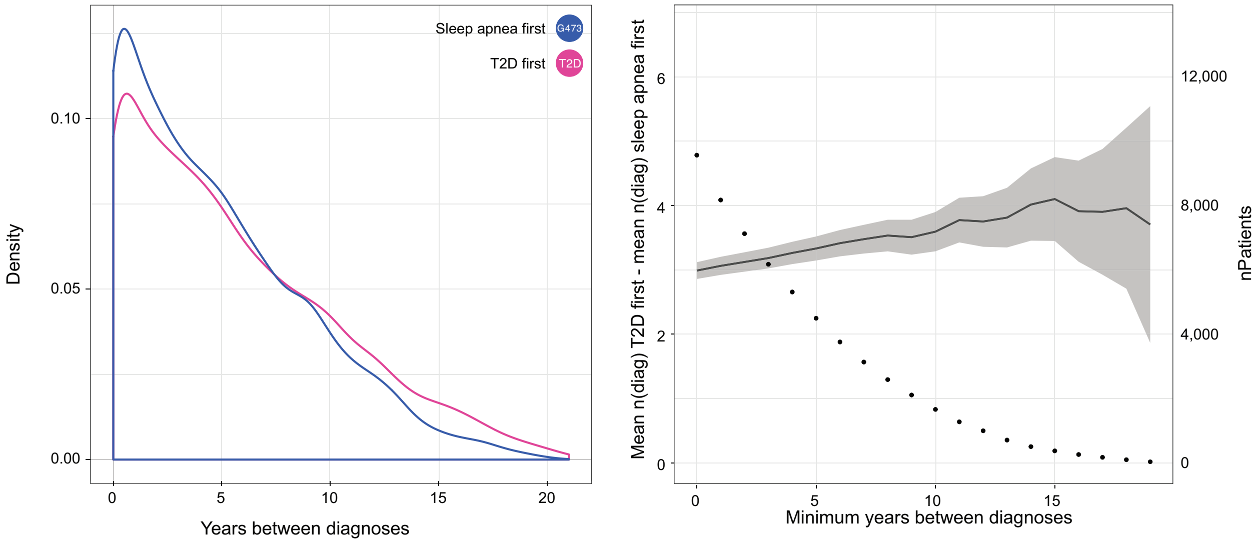


Fig. 3. The increased comorbidity burden for patients diagnosed with T2D before sleep apnea. (A) Distribution of years between T2D and sleep apnea for patients diagnosed with T2D first (pink) and for patients diagnosed with sleep apnea first (blue). (B) The excess number of comorbidities for patients diagnosed with T2D first compared to those diagnosed with sleep apnea first (black line) with the 95% confidence interval (grey area). The x-axis indicates the minimum number of whole years between diagnoses (e.g. 0 years means more than one day but less than a year). The dots indicate the number of patients having minimum x years between the two diagnoses.

3.2. *Diabetes before sleep apnea is associated with an increased amount of comorbidities*

To further investigate the temporal association between sleep apnea and T2D, we defined T2D patients based on the method presented by Lind at al [29,31], using a combination of prescribed drugs and disease codes (see Methods). We found that 11,054 T2D patients have been diagnosed with sleep apnea. A total of 6,061 patients (54,8%) were diagnosed with T2D before sleep apnea, and 4,752 patients were diagnosed with sleep apnea before T2D. In addition, 241 patients were diagnosed with T2D and sleep apnea on the same day. These 241 patients are disregarded in this study, since there is no reliable way to determine which disease came first. Consequently, even though sleep apnea was significantly associated with T2D (RR = 2.87, $p < 2.3E-308$), NIDDM and sleep apnea does not appear as a temporal pair, due to the lack of a significant temporal order in which these diseases are diagnosed.

To investigate if the patients diagnosed with T2D before adult sleep apnea and patients diagnosed with adult sleep apnea before T2D are two distinct patient groups, we examined the RR for all level three ICD-10 diagnoses for patients with adult sleep apnea and T2D. Those first diagnosed with T2D had on average 3.0 (95% CI: 2.9-3.1) comorbidities more than those diagnosed with adult sleep apnea first. We interpret this as an indicator that the patients first diagnosed with T2D have, on average, a higher disease burden.

The time of diagnosis can be imprecise since neither sleep apnea nor T2D are acute diseases. Consequently, it could be arbitrary which disease was diagnosed first. For some patients the two diagnoses are acquired relatively close to each other, but for many patients there are several years or even decades between the diagnoses (Figure 3A).

We tested if there was a significant difference in the number of diagnoses between these two groups using a Poisson regression model. Covariates include years between the two diagnoses, which disease was diagnosed first, age and gender. We used the fitted model to calculate a point estimate of the number of comorbidities for each patient, given the minimum number of years between sleep apnea and T2D (Figure 3B). The overall difference was 3.0 comorbidities, with patients first diagnosed with T2D being most sick. This difference increases as the number of years between T2D and sleep apnea increases (Figure 3B). Collectively, this clearly illustrates a difference in the general health status of these patients groups.

3.3. *Diabetes before other diseases tends to increase the comorbidity burden*

We applied the same method to investigate if other diabetes comorbidities showed a different comorbidity burden depending on the diagnosis order. We found seventeen diseases positively associated with T2D, and where the diagnostic order for each disease and T2D was close to 50:50 (see Methods). To remove rare disorders we required a minimum of 1,000 T2D patients to have the diagnosis, reducing the number down to sixteen diagnoses of interest. Lastly, we performed an analysis calculating the difference in mean number of comorbidities for patients diagnosed with T2D first compared to patients diagnosed with the other particular diagnosis first. This resulted in twelve diagnoses with a minimum of ten significant time points (Figure 4). Ten out of the twelve diagnoses were associated with a higher comorbidity burden if they were diagnosed with T2D before the other diagnosis, with the two exceptions: Migraine and “Poisoning by psychotropic drugs, not elsewhere classified”.

4. Discussion

In this study we examined the complex issue of temporal directionality of disease co-occurrences and used temporal disease trajectories to present a model for stratification of patient groups according to longitudinal patterns.

Using one example analyzed in detail we illustrated the complexities and rediscovered that age of sleep apnea diagnoses follow a bimodal distribution, illustrating two distinct diseases: childhood sleep apnea and adult sleep apnea – a distinction well known in the literature [11,16–18,32]. By investigating the detailed time-ordered relationships between sleep apnea and T2D we confirmed that sleep apnea in the adult population is significantly associated with T2D in the time-dependent analysis. Surprisingly, there was no direct edge between any of the diabetes diagnoses in our temporal disease network, showing that there was no directionality of the T2D and adult sleep apnea diagnoses, in fact we showed that 4,752 patients acquire adult sleep apnea before T2D, 6,061 acquire T2D first while 241 patients acquired the diagnoses on the same day.

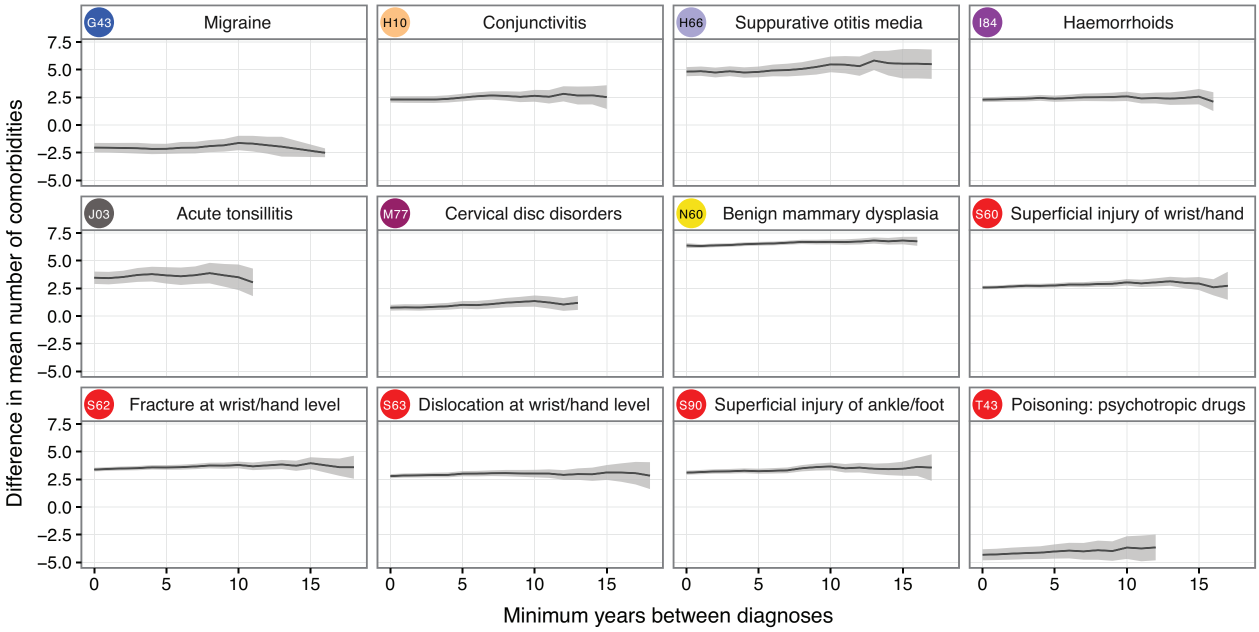


Fig. 4. Comorbidity burden levels as function of time span between diagnoses. Together the panels show that the change in comorbidity burden depends on the disease order. Each disease is indicated by an ICD-10 code colored according to the ICD-10 chapter followed by the name of the disease. The excess number of comorbidities for patients diagnosed with T2D first compared to those diagnosed with the other particular diagnosis (black line) with the 95% confidence interval (grey area). The x-axis indicates the minimum number of whole years between diagnoses.

Interestingly, we found that this order significantly influenced the amount of comorbidities acquired, indicating that patients diagnosed with diabetes before adult sleep apnea have a worse general health status than patients first diagnosed with adult sleep apnea. This is, to our knowledge, the first time this temporal effect of sleep apnea and T2D has been described. To further illustrate the importance of the order, we showed that the difference in the quantity of comorbidities slightly increased with increased time between the diagnoses. Based on these observations we suggest that there is a synergetic effect of T2D and adult sleep apnea, which is dependent on the order of the diagnoses.

We further underlined the importance of order of diagnoses by applying this method to all T2D comorbidities. This resulted in twelve diagnoses with a significant different number of comorbidities depending on the diagnosis order.

Precision medicine attempts to subdivide patients into groups that will benefit from tailor-made treatment. We show in this paper that disease progression patterns can be highly complex even in cases where disease co-occurrence orders appear to be random. The identification of genomic biomarkers could most likely to a higher degree benefit from taking this type of stratification into account in contrast to current models that mostly are based on the case/control paradigm where diseases are investigated individually.

5. Acknowledgements

We would like to acknowledge funding from the Novo Nordisk Foundation (grant agreement NNF14CC0001) as well as the H2020 project MedBioInformatics.

References

1. Aurora, R. N. & Punjabi, N. M. *Lancet Respir. Med.* **1,** 329–338 (2013).
2. Golden, S. H. *et al. JAMA* **299,** 2751–2759 (2008).
3. Hesdorffer, D. C. *et al. Ann. Neurol.* **72,** 184–191 (2012).
4. Lippi, G., Montagnana, M., Favaloro, E. J. & Franchini, M. *Seminars in Thrombosis and Hemostasis* **35,** 325–336 (2009).
5. Mezuk, B., Eaton, W. W., Albrecht, S. & Golden, S. H. *Diabetes Care* **31,** 2383–2390 (2008).
6. Pan, A. *et al. Arch. Intern. Med.* **170,** 1884–91 (2010).
7. Pan, A. *et al. Diabetes Care* **35,** 1171–1180 (2012).
8. Lalla, E. & Papapanou, P. N. *Nat. Rev. Endocrinol.* **7,** 738–48 (2011).
9. Rajan, P. & Greenberg, H. *Nat. Sci. Sleep* **7,** 113–25 (2015).
10. Peppard, P. E. *et al. Am. J. Epidemiol.* **177,** 1006–1014 (2013).
11. Sharma, S. K. & Ahluwalia, G. *Indian J. of Med. Res.* **131,** 171–175 (2010).
12. Ip, M. S. M. *et al. Chest* **119,** 62–69 (2001).
13. Javaheri, S. *Clinics in Chest Medicine* **31,** 235–248 (2010).
14. Morgenthaler, T. I., Kagramanov, V., Hanak, V. & Decker, P. A. *Sleep* **29,** 1203–1209 (2006).
15. Khan, M. T. & Franco, R. A. *Sleep Disord.* 798487 (2014).
16. Tan, H.-L., Gozal, D. & Kheirandish-Gozal, L. *Nat. Sci. Sleep* **5,** 109–23 (2013).
17. Marcus, C. L. *et al. Pediatrics* **130,** 576–84 (2012).
18. Marcus, C. L. *et al. N. Engl. J. Med.* **368,** 2366–76 (2013).
19. Bixler, E. O. *et al. Sleep* **32,** 731–6 (2009).
20. Malhotra, A. & White, D. P. *The Lancet* **360,** 237–245 (2002).
21. Parati, G. *et al. J. Hypertens.* **30,** 633–46 (2012).
22. Cappuccio, F. P., D'Elia, L., Strazzullo, P. & Miller, M. A. *Diabetes Care* **33,** 414–420 (2010).
23. Malhotra, A. *et al. Am. J. Respir. Crit. Care Med.* **166,** 1388–1395 (2002).
24. Chervin, R. D. *Chest* **118,** 372–379 (2000).
25. Barceló, A. *et al. Thorax* **63,** 946–50 (2008).
26. Fowler, M. J. *Clin. Diabetes* **29,** 116–122 (2011).
27. Alves, C., Casqueiro, J. & Casqueiro, J. *Indian J. Endocrinol. Metab.* **16,** 27 (2012).
28. DeFronzo, R. A. *et al. Nat. Rev. Dis. Prim.* **1,** 15019 (2015).
29. Lind, M. *et al. Diabetologia* **55,** 2946–2953 (2012).
30. Bertoni, A. G., Saydah, S. & Brancati, F. L. *Diabetes Care* **24,** 1044–1049 (2001).
31. Lind, M. *et al. N. Engl. J. Med.* **371,** 1972–1982 (2014).
32. Jordan, A. S., McSharry, D. G. & Malhotra, A. *Lancet* **383,** 736–47 (2014).
33. Jensen, A. B. *et al. Nat. Commun.* **5,** 4022 (2014).

MICRORNA-AUGMENTED PATHWAYS (mirAP) AND THEIR APPLICATIONS TO PATHWAY ANALYSIS AND DISEASE SUBTYPING

DIANA DIAZ[1], MICHELE DONATO[3], TIN NGUYEN[1], SORIN DRAGHICI[1,2]

[1] *Department of Computer Science, Wayne State University,*
Detroit, MI 48202, U.S.A.
[2] *Department of Obstetrics and Gynecology, Wayne State University,*
Detroit, MI 48202, U.S.A.
[3] *Institute for Immunity, Transplantation and Infection, Stanford University Medical Center,*
Stanford, CA 94305, U.S.A.
E-mail: sorin@wayne.edu

MicroRNAs play important roles in the development of many complex diseases. Because of their importance, the analysis of signaling pathways including miRNA interactions holds the potential for unveiling the mechanisms underlying such diseases. However, current signaling pathway databases are limited to interactions between genes and ignore miRNAs. Here, we use the information on miRNA targets to build a database of miRNA-augmented pathways (mirAP), and we show its application in the contexts of integrative pathway analysis and disease subtyping. Our miRNA-mRNA integrative pathway analysis pipeline incorporates a topology-aware approach that we previously implemented. Our integrative disease subtyping pipeline takes into account survival data, gene and miRNA expression, and knowledge of the interactions among genes. We demonstrate the advantages of our approach by analyzing nine sample-matched datasets that provide both miRNA and mRNA expression. We show that integrating miRNAs into pathway analysis results in greater statistical power, and provides a more comprehensive view of the underlying phenomena. We also compare our disease subtyping method with the state-of-the-art integrative analysis by analyzing a colorectal cancer database from TCGA. The colorectal cancer subtypes identified by our approach are significantly different in terms of their survival expectation. These miRNA-augmented pathways offer a more comprehensive view and a deeper understanding of biological pathways. A better understanding of the molecular processes associated with patients' survival can help to a better prognosis and an appropriate treatment for each subtype.

1. Introduction

The identification of biological processes underlying conditions is crucial for disease prognosis and treatment programs. As gene signaling pathways are capable of representing complex interactions between genes, pathway databases have become essential for several gene expression analyses. Signaling pathway databases are remarkably important because they allow researchers to analyze high-throughput data in a functional context, reducing complexity and increasing the explanatory power. However, there are other molecules that play important roles in gene regulation, such as microRNAs, which are not included into current pathway databases. MicroRNAs (miRNAs) are small RNA molecules capable of suppressing protein production by binding to gene transcripts. In fact, more than 30% of the protein-coding genes in humans are miRNA-regulated. Additionally, miRNAs have been shown to play an important role in diagnosis and prognosis for different types of diseases[1].

The integration of miRNA into signaling pathways have multiple applications, such as pathway analysis and disease subtyping. Pathway analysis techniques and methods aim to analyze high-throughput data with the goal of identifying pathways that are significantly impacted by a given condition. The typical input of pathway analysis includes gene expression data from two different phenotypes (e.g., condition vs. control) and a set of signaling pathways. Although current pathway analysis methods using gene expression (mRNA) have achieved excellent results[2–4], mRNA expression alone is unable to capture the complete picture of biological processes, as other entities also play important roles. Relevant work has been done to elucidate the important interplay between miRNAs and biological pathways[5–9]. The state-of-the-art approach for miRNA-mRNA pathway analysis is microGraphite[8] which uses an empirical gene set approach. microGraphite's main goal is the identification of signal transduction paths correlated with the condition under study[10].

A second crucial process in the understanding of complex diseases is disease subtyping. Identifying clinically meaningful subtypes in complex diseases is crucial for improving prognosis, treatment, and precision medicine[11]. A typical input of disease subtyping consists of various clinical variables and gene expression data from patients affected by a particular disease. The expected output consists of well-identified groups of patients that highly correlate with one or more variables, such as observed survival (e.g., long-term vs. short-term survival patients). Disease subtyping is typically expressed as a clustering problem with the goal of partition patients in groups based on their genetic similarities with the additional complexity that the number of clusters is unknown. Several methods for disease subtyping using gene expression data have been developed[11–15]. Integrative analysis using clinical data, multi-'omics' data, and prior biological knowledge can leverage current disease subtyping methods.

In this paper, we present a tool for integrating miRNA into signaling pathways (mirIntegrator), a publicly available miRNA-augmented pathway database (mirAP), and we show the applications of such augmentation to pathway analysis and disease subtyping. We have used mirIntegrator previously as a part of our orthogonal meta-analysis approach[16].

Our pathway analysis pipeline uses mirAP and Impact Analysis[3,4], a topology-aware pathway analysis method previously developed by our group. To demonstrate the advantage of our method, we analyze 9 datasets studying 7 different diseases with mRNA and miRNA expression. We show that the proposed approach is able to identify the pathways that describe the underlying diseases as significant. The p-values and rankings of these pathways are significantly smaller than those obtained without data integration as well as when using microGraphite[8].

Our disease subtyping pipeline uses miRNA and mRNA expression data, available clinical variables, and prior biological knowledge. This method includes a feature selection approach based on mirAP to reduce the effective dimensionality of the unsupervised clustering problem. We analyze colorectal cancer miRNA, gene expression data, and clinical records downloaded from the Cancer Genome Atlas (TCGA) with our pipeline and SNF[15], a recently proposed integrative disease subtyping method. The colorectal cancer-relevant pathways and subgroups identified with our approach are significantly different in terms of their survival expectation, outperforming the approach that does not use miRNA, and providing information on biological mechanisms relevant to the difference in survival.

2. Methods

In this section, we propose an algorithm for integrating miRNA into signaling pathways. We also describe two pipelines using miRNA-augmented pathways (mirAP). The first pipeline is for pathway analysis (PA) and the second one is for disease subtyping (DS). The scenarios for these analyses are different. PA is used in biological studies comparing genetic samples from two different phenotypes (e.g., disease vs. control samples), and DS is used in studies with samples of patients undergoing the same disease for which the clinical subtypes are unknown. Our PA pipeline is able to integrate miRNA and mRNA expression data and identify pathways that are related to the disease under study. Our DS pipeline is able of incorporate biological pathways to partition patients into groups with very different survival patterns.

2.1. *Pathway augmentation*

This method augments the graphical representation of original signaling pathways with interactions between miRNAs and their target genes. The input of this method includes a set of signaling pathways and known miRNA-mRNA interactions (Fig. 1a,b). The output is a set of augmented pathways that consists of the original genes, the miRNAs that target those genes and their interactions. Let $P = (V, E)$ denote the graphical representation of the original gene-gene pathway, and $T : M \to V$ a function that identifies the target genes of miRNAs in M. An edge $e \in E$ can be represented as a 3-tuple $e = (g_1, g_2, interaction)$. We augment the nodes and edges of the original pathway as follows:

$$\bar{V} = V \cup \{m \in M | T(m) \cap V \neq \varnothing\}$$
$$\bar{E} = E \cup \{(m, g, inhibition) | m \in V \cap M \wedge g \in T(m)\}$$

We implemented this algorithm in R and published it as the Bioconductor package named mirIntegrator (`http://bit.ly/mirIntegrator`). mirIntegrator is flexible and allows users to integrate user-specific pathway databases with user-specific miRNA-mRNA target databases. Additionally, it generates graphical representations of the augmented pathways (see Fig. 5). We integrated pathways from Kyoto Encyclopedia of Genes and Genomes[17] (KEGG) (version 73) with miRNA targets from miRTarBase[18] (version 4.5) to generate mirAP, a database of miRNA-augmented pathways (`http://www.cs.wayne.edu/dmd/mirAP`).

2.2. *Integrative pathway analysis*

Our pathway analysis pipeline consists of two main steps. In the first step, we augment the signaling pathways with interactions between miRNAs and their targets. Once this is done, the data integration problem is mapped to the original pathway analysis problem for which existing methods can be applied. The difference is that here both miRNA and mRNA expression can be taken into consideration. In the second step, we apply any pathway analysis that uses fold change and p-value as input, e.g., Over-representation analysis[19] (ORA) and Impact Analysis[3,4]. ORA and Impact Analysis are well-known methods developed by our group to identify signaling pathways that are impacted by the effects of diseases. Fig. 1 displays the overall pipeline of our approach.

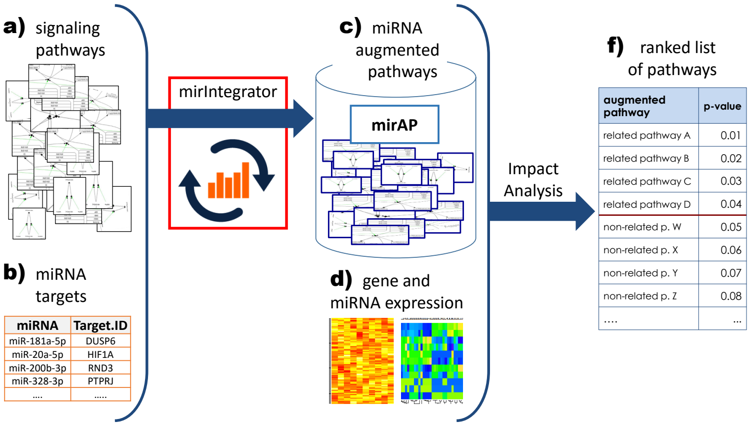


Fig. 1. Workflow of pathway analysis using augmented pathways.

Impact Analysis[3,4] is a widely used topology-aware method that combines two types of evidence: i) the over-representation (ORA) of differentially expressed (DE) genes in a pathway[19], and ii) the perturbation (PERT) of such a pathway, as measured by propagating expression changes through the pathway topology. These two types of evidence are captured by two independent p-values[4]: p_{ORA} and p_{PERT}. These p-values are combined using Fisher's method to obtain a global p-value per pathway. Each global p-value represents the probability of having the observed number of DE genes, as well as the observed amount of impact just by chance (i.e. when the null hypothesis is true)[4]. To calculate p_{ORA} on mirAP, we assumed that the number of DE entities (genes and miRNAs) on the given pathway follows a hypergeometric distribution. The following information is needed to compute p_{ORA}: i) the total number of measured entities, ii) the number of entities belonging to the given augmented pathway, iii) the total number of DE entities, and iv) the number of DE entities in the given augmented pathway. To calculate p_{PERT} on mirAP, we perform a bootstrap procedure using the following input: i) the log-fold change of DE entities, and ii) the given augmented pathway.

2.3. *Integrative disease subtyping*

Our disease subtyping pipeline is presented on Fig. 2. The input includes: i) mRNA and miRNA sample-matched expression data, ii) survival records, iii) a database of miRNA-target gene interactions, and iv) a database of signaling pathways (see Fig. 2a). The output is a set of selected pathways (Fig. 2f) yielding to subtypes with significantly distinct survival patterns.

First, we obtain the miRNA-augmented pathways from mirAP (Fig. 2b). Second, we partition the patients using the genes and miRNAs provided by each augmented pathway (Fig. 2c). e.g., let us say that we want to analyze gene and miRNA expression from $\mathcal{N}$ number of patients and we obtained $\mathcal{P}$ number of augmented pathways from mirAP. Taking one pathway at the time, we filter the gene expression data by selecting only genes that belong to the pathway. Similarly, we filter the miRNA expression data by selecting only miRNAs that belong to the pathway. Now, we need to combine the filtered gene expression and miRNA data and then perform clustering on the combined data. So, we use Similarity Network Fusion method[15] (SNF) in conjunction with spectral clustering[20] for this purpose. We repeat this process with each pathway to obtain $\mathcal{P}$ different pathway-based clusterings, one per each pathway.

Third, we perform survival analysis on each of the pathway-based clusterings (Fig. 2d). In order to do this, we compute the log-rank test p-value (Cp) of Cox proportional hazards regression analysis by using the input survival information.This p-value represents how significant the difference between the survival curves is. For instance, a Cox log-rank test p-value close to zero may indicate that these groups have well-differentiated survival patterns. Now the question is whether we could obtain the same clustering just by chance[21]. To answer this question we use the random sampling technique. For example, if the pathway has G number of genes and m number of miRNAs, we randomly select G genes and m miRNAs from the measured values. Then, we partition the patients using this randomly selected set of entities and then compute its Cox p-value (rCp). We repeat this random selection a large number of times (e.g., 2,000 times) to construct an empirical distribution of Cox p-values (Fig. 2d). Next, we compare the observed Cox p-value Cp with the distribution of rCp, calculated from randomly selected genes and miRNAs. We estimate the probability of obtaining this Cp by computing the proportion of resampling p-values less than or equal to the observed Cp (e.g., In Fig. 2d the vertical red line indicates the observed Cp). For each pathway, we estimate this probability in order to quantify how likely it is to observe by chance a Cox p-value less than or equal to the one observed with the actual genes and miRNAs in the pathway.

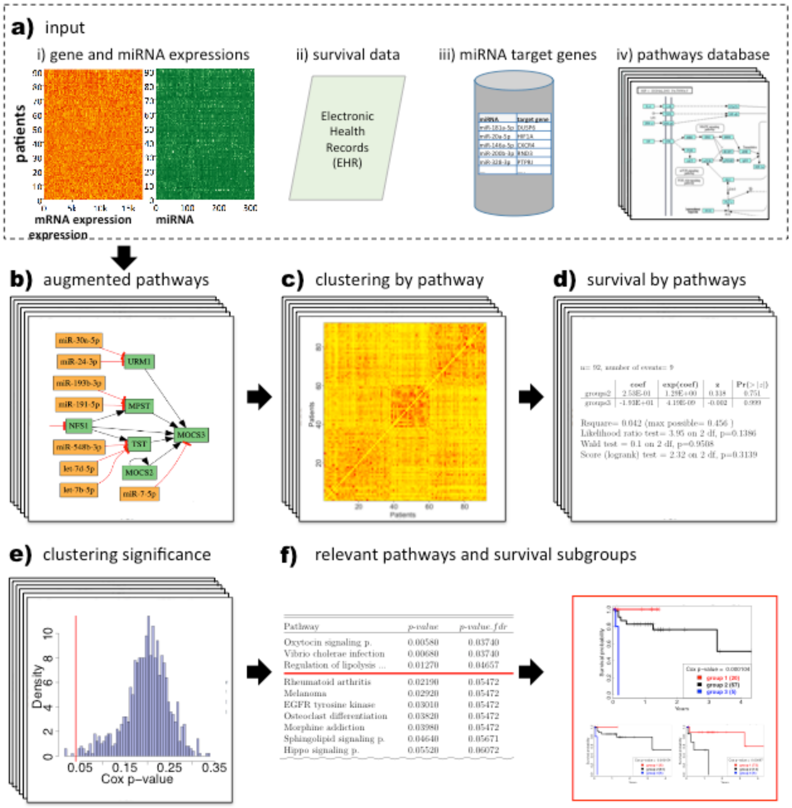

a) input
i) gene and miRNA expressions
ii) survival data
iii) miRNA target genes
iv) pathways database
patients
mRNA expression
miRNA
expression
Electronic Health Records (EHR)
b) augmented pathways
c) clustering by pathway
d) survival by pathways
e) clustering significance
Density
Cox p-value
0.05 0.15 0.25 0.35
f) relevant pathways and survival subgroups
Pathway p-value p-value.fdr
Oxytocin signaling p. 0.00580 0.03740
Vibrio cholerae infection 0.00680 0.03740
Regulation of lipolysis ... 0.01270 0.04657
Rheumatoid arthritis 0.02190 0.05472
Melanoma 0.02920 0.05472
EGFR tyrosine kinase 0.03010 0.05472
Osteoclast differentiation 0.03820 0.05472
Morphine addiction 0.03980 0.05472
Sphingolipid signaling p. 0.04640 0.05671
Hippo signaling p. 0.05520 0.06072

Fig. 2. The proposed pipeline for disease subtyping.

The final step is to select the pathways that are relevant to survival, i.e., pathways yielding to significantly distinct survival curves. To do this, we adjust the p_i p-values for multiple comparisons using False Discovery Rate (FDR). We then rank the pathways by FDR.p-value and select those less than or equal to the significance threshold of 5% as *relevant pathways*. We note that this pipeline can be used in conjunction with other integrative clustering methods.

3. Results

In this section, we present the results of our pathway analysis and disease subtyping pipelines using the miRNA-augmented pathways (mirAP). First, we perform pathway analysis of 9 mRNA/miRNA sample-matched datasets using two different methods (Impact Analysis and ORA) and show that mirAP offers a significant improvement over analyzing mRNA data alone. We also compare the obtained results with the state-of-the-art method (microGraphite)[8]. Second, we perform disease subtyping of a colorectal cancer dataset from TCGA using our subtyping pipeline and compare with the traditional pipeline for subtyping.

3.1. *Validation of our pathway analysis pipeline*

We analyze nine sample-matched datasets from seven different diseases: GSE43592 (multiple sclerosis, 10 controls, 10 cases), GSE35389 (melanoma, 4 controls, 4 cases), GSE35982 (colorectal cancer, 8 controls, 8 cases), GSE26168 (type II diabetes, 8 controls, 9 cases), GSE62699 (alcoholism, 18 controls, 18 cases), GSE35834 (colorectal cancer, 23 controls, 55 cases), GSE43797 (pancreatic cancer, 5 controls, 7 cases), GSE29250 (non-small cell lung cancer, 6 controls, 6 cases), and GSE32688 (pancreatic cancer, 7 controls, 25 cases). For each of these datasets, we used the normalized expression values as found in GEO.[22] The microarray probes were annotated according to their corresponding platform's metadata using GEOquery.[23] Next, we estimated log-fold-change between disease and control groups by fitting to a gene-wise linear model using the R package limma[24]. We use the following two criteria to identify differentially expressed (DE) genes: i) genes with adjusted p-value lower than 5%, and ii) among the genes that satisfy the first criterion, we choose the genes with the highest log-fold change, up to 10% of measured genes. We use the same criteria to identify DE miRNAs.

The nine datasets were selected due to two important reasons. First, these datasets have both mRNA and miRNA measurements for the same set of patients. Second, for each of the underlying diseases, there is a KEGG pathway, henceforth *target pathway*, that was created to describe the underlying mechanisms of the disease. To demonstrate the advantage of the miRNA data integration, we compare the use of the original KEGG pathways with the use of our miRNA augmented pathways (mirAP) by performing two pathway analysis methods that use p-value and fold-change: Impact Analysis (IA)[4] and over-representation analysis (ORA)[19]. The input for IA and ORA using KEGG is mRNA expression data. The input for IA and ORA using mirAP includes both mRNA and miRNA expression data. The output of each method is a list of p-values – one per pathway. These p-values are adjusted for multiple comparisons using False Discovery Rate (FDR)[25].

We also analyze the nine GEO datasets using microGraphite[8] after quantile normalization to compare with our pipeline. The main goal of microGraphite is the identification of signal transduction paths correlated with the condition under study. It is implemented in a four-steps recursive procedure as follows: (i) selection of pathways, (ii) best path identification, (iii) metapathway construction, and (iv) metapathway analysis. Here we only consider the first step of the approach, which is the selection of significant pathways. This selection is based on the significance levels obtained from the test on the mean of the pathways (alpha-mean). The input is the mRNA and miRNA expression data and it does not take in account fold-changes nor differentially expressed entities.

For each dataset, we expect a good method to identify the target pathway as significant, as well as to rank it on top. For instance, in the colorectal cancer dataset which compares colorectal cancer tissue vs. normal, the *Colorectal cancer pathway* must be shown as significant and should be as close to the top of the ranking as possible since this is the pathway that describes the phenomena involved in colorectal cancer. Based on this, we compare the rank and p-value of the target pathway in each disease using the five methods: i) mRNA expression alone using standard KEGG pathways with ORA and ii) IA, iii) mRNA and miRNA expression data using the augmented pathways (mirAP) with iii) ORA and iv) IA, and v) mRNA and miRNA expression data analyzed with microGraphite.

Table 1. Results of target pathway identification using traditional ORA (column 3), traditional IA (col. 4), ORA on mirAP (col. 5), IA on mirAP (col. 6), microGraphite (col. 7)

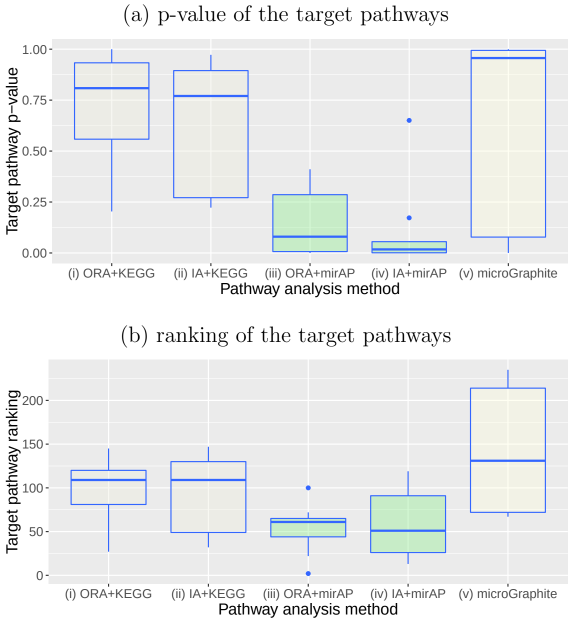
GEO ID	Target pathway	ORA	IA	ORAmir	IAmir	microGraphite
GSE26168	Type II diabetes mellitus	no	no	no	no	yes
GSE29250	Non-small cell lung cancer	no	no	yes	no	no
GSE35982	Colorectal cancer	no	no	no	no	no
GSE32688	Pancreatic cancer	no	no	yes	yes	no
GSE35389	Melanoma	no	no	yes	yes	no
GSE35834	Colorectal cancer	no	no	yes	yes	no
GSE43592	Amyotrophic lateral scle.	no	no	no	yes	no
GSE43797	Pancreatic cancer	no	no	yes	yes	yes
GSE62699	Alcoholism	no	no	no	yes	no

Table 1 shows the target pathways and their significance for the 9 datasets. The first and second columns display the datasets and their corresponding target pathways while the other five columns indicate whether the target pathways are identified as significant using the five methods: ORA of mRNA expression on KEGG pathways (ORA+KEGG), IA of mRNA expression on KEGG (IA+KEGG), ORA of miRNA and mRNA expression data on mirRNA-augmented pathways (ORA+mirAP), our approach IA of miRNA and mRNA expression on mirAP (IA+mirAP), and miRNA and mRNA expression analysis using microGraphite, respectively. The significance threshold is 5% for FDR p-values. IA and ORA fail to identify any target pathway as significant when using only mRNA whereas our approach (IA+mirAP) correctly identify the target in 6 out of 9 datasets (GSE32688, GSE35389, GSE35834, GSE43592, GSE43797, GSE62699) and ORA+mirAP correctly identify the target pathway as significant in 5 out of 9 datasets (GSE29250, GSE32688, GSE35389, GSE35834, GSE43797). microGraphite correctly identifies the target pathway as significant in only 2 out of 9 datasets (GSE26168, GSE43797). The results demonstrate that our integration of mRNA and miRNA lifts the statistical power for both pathway analysis techniques (ORA and IA) and outperforms microGraphite in target pathway identification.

Fig. 3 shows the p-values and rankings of the target pathways using the five methods. The panel (a) shows the FDR corrected p-values of the target pathways. We compare the lists of p-values using Wilcoxon test. The FDR p-values produced by IA+mirAP are significantly smaller than by IA+KEGG (p=0.007), ORA+KEGG (p=0.005), and microGraphite (p=0.009).

The panel (b) shows the rankings of the target pathways. Again, the rankings produced by IA+mirAP are significantly smaller than those of IA+KEGG (p=0.03 using t-test, and p=0.04 using Wilcoxon test), ORA+KEGG (p=0.03 using t-test and p=0.04 using Wilcoxon test), and microGraphite (p=0.0051 using t-test and p=0.0058 using Wilcoxon test). This confirms that our augmented pathways, mirAP, improve the performance of traditional Impact Analysis and ORA. Also, the results show that the proposed integrative pathway analysis also outperforms microGraphite in terms of both p-values and rankings for target pathway identification.

Furthermore, our pathway database (mirAP) is generated with validated miRNA-mRNA interactions, while microGraphite uses predicted interactions, which increases the number of false positive miRNA-target interactions. Another drawback of microGraphite is it execution time. A typical analysis with microGraphite takes approximately 22 hours while our approach takes only a few minutes. We ran these experiments on a typical desktop workstation with a 2.6 GHz Intel Core i5, 8GB of RAM, on a single thread, and the OS X 10.11 operative system.


(a) p-value of the target pathways

(b) ranking of the target pathways

Fig. 3. Corrected p-values and rankings of the target pathways using different methods.

3.2. *Validation of our disease subtyping pipeline*

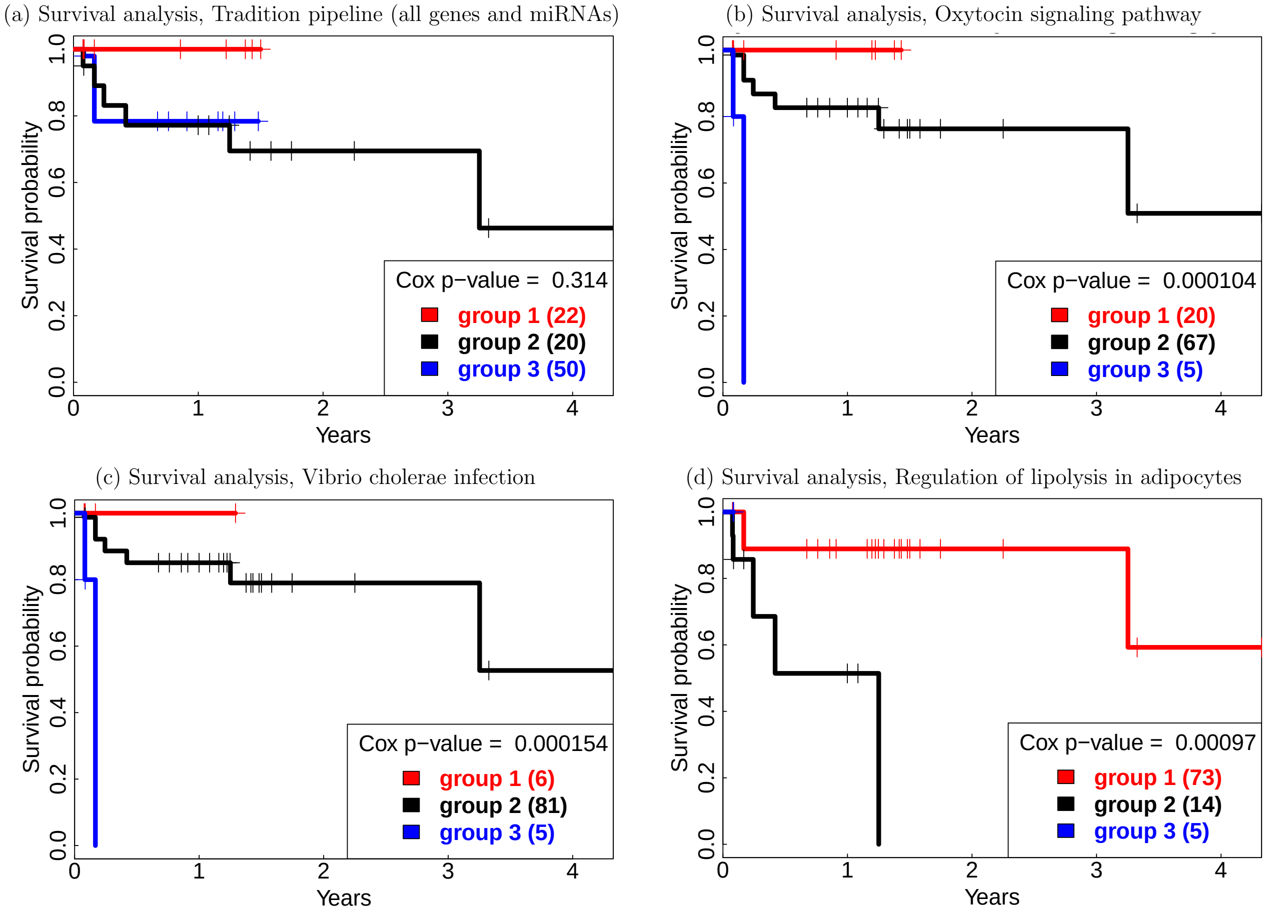
To assess our disease subtyping pipeline we use matched-sample gene and miRNA expression data (level 3 from platforms Agilent G4502A-07 and Illumina GASeq miRNASeq, respectively) of colorectal cancer patients (COAD) downloaded from the Cancer Genome Atlas (TCGA) (`cancergenome.nih.gov`). We selected the largest set of patients with miRNA-mRNA matched samples and available survival records, as were selected in SNF[15]. The number of patients is $M = 92$, the number of genes is $N_g = 17,814$, and the number of miRNAs is $N_m = 705$. We performed unsupervised clustering with the number of clusters set as $k = 3$ according to prior knowledge of the number of subtypes of COAD[15]. We use SNF[15] in conjunction with spectral clustering[20] as integrative clustering method. To perform SNF clustering, we used the SNFtool package with the suggested parameters.

For each miRNA-augmented pathway, our method partitions the patients using the genes and miRNAs in the pathway as clustering features, resulting in a total of 184 clusterings. Then for each pathway-based clustering, we construct the empirical distribution and then estimated the *p-value* of how likely the pathway helps to improve disease subtyping. The *p-values* of the relevant pathways are shown in Table 2. We select the pathways with a FDR-corrected *p-value* ≤ 0.05 as *relevant pathways*. The horizontal red line represents the significance cutoff at 5%. For TCGA-COAD, we identify three relevant pathways: *Oxytocin signaling pathway*, *Vibrio cholerae infection*, and *Regulation of lipolysis in adipocytes*.

We also cluster the 92 patients using SNF with the traditional pipeline, i.e., using all the measured genes and miRNAs. We compare these partitions with those obtained by our pipeline. To assess the correlation between the obtained groups and survival patterns (e.g., long-term vs. short-term survival), we performed survival analysis for all the cases using Kaplan-Meier analysis.

Table 2. List of relevant pathways for colorectal subtyping.

Pathway	*p-value*	*p-value.fdr*
Oxytocin signaling pathway	0.00580	0.0374
Vibrio cholerae infection	0.00680	0.0374
Regulation of lipolysis in adipocytes	0.01270	0.0466
Rheumatoid arthritis	0.02190	0.0547
...	...	...



(a) Survival analysis, Tradition pipeline (all genes and miRNAs)

(b) Survival analysis, Oxytocin signaling pathway

(c) Survival analysis, Vibrio cholerae infection

(d) Survival analysis, Regulation of lipolysis in adipocytes

Fig. 4. Kaplan-Meier survival analysis of the obtained COAD subtypes. a) Survival curves using all genes and miRNAs. b), c), and d) Survival curves using relevant pathways.

Fig. 4 shows the Kaplan-Meier plots, each one represents the association of the obtained groups with the observed patient survival. Fig. 4a shows the subtypes obtained with the traditional pipeline using all 17, 814 genes and 705 miRNAs. In a Cox proportional hazards regression analysis, we find that there is no statistically significant difference between survival groups obtained with the traditional pipeline (log rank test p-value = 0.314). Fig. 4b, c, and d. shows the resultant clustering on the relevant pathways identified with our approach (Table 2). Clustering based on *Oxytocin signaling pathway* entities gives a log rank test p-value of 0.000104, which indicates a significant difference between the survival curves (Fig. 4b). Similarly, clusterings based on *Vibrio cholerae infection* and *Regulation of lipolysis in adipocytes* augmented pathways indicate significant differences between the survival curves with p-values of $p = 0.000154$ and $p = 0.00097$, respectively (Fig. 4c and d). As we can see, integrative clustering based on relevant mirAP pathways produce subtypes significantly more related to survival data than the traditional subtyping pipeline (approximately 1000 times lower p-values).

Given that our approach requires resampling for computing the pathways' significance (*p-values*), our pipeline is more time consuming than the traditional pipeline. For the computational experiments presented here, we generated 2, 000 random clusterings per each pathway. Our pipeline took some hours to subtype the set of patients (approximately 4 hours) while running SNF alone takes only some minutes (less than 3 minutes).

3.2.1. *Biological Significance of relevant Signaling Pathways*

Our pipeline identifies the *Oxytocin signaling pathway* to be related to the survival subtyping of colorectal cancer patients ($p = 0.000104$). Oxytocin (OXT) is a hormone with a well-known effect on uterine smooth muscles and myoepithelial cells. Additionally, it has been shown that oxytocin is expressed along the entire human gastrointestinal (GI) tract, including colon, and it contributes to the control of the GI motility[26]. Moreover, studies have shown that exposure to OXT leads to a significant decrease in cell proliferation for some epithelial cancer cells (e.g., breast and prostate cancer)[27]. In contrast, OXT has a growth-stimulating effect in other types of cancer cells (e.g., small-cell lung cancer, endothelial cancer, and Kaposiâs sarcoma)[28,29]. We think that the evidence of OXT expression on colon and the dual role that OXT has in some cancer cells (as inhibitor and promoter of cancer cells proliferation) may indicate that OXT could also play an important role in differentiating short and long-term survival COAD patients. In addition, OXT is also known to be capable of mitigating symptoms caused by stress, OXT levels increase in acute(short-lived) stress and decrease during chronic stress. Also, it is well-known that chronic stress has an outstanding role in cancer growth and metastasis.[30] From this, we also hypothesize that patients in the short term survival group (Fig. 4b, gr. 3) may have been in a metastatic stage with chronic stress and different OXT expression than patients in the other groups (Fig. 4b,1-2).

Similarly, we identify *Vibrio cholerae infection* pathway as relevant. This pathway describes the colonization of the intestine by Vibrio cholerae bacteria (VC). The main factor involved in this process is Cholera toxin (CTX). Several studies have exhibit relations between gastrointestinal tract bacteria and colon cancer progression. In particular, it has been shown that CTX suppresses carcinogenesis of inflammation-driven sporadic colon cancer[31].

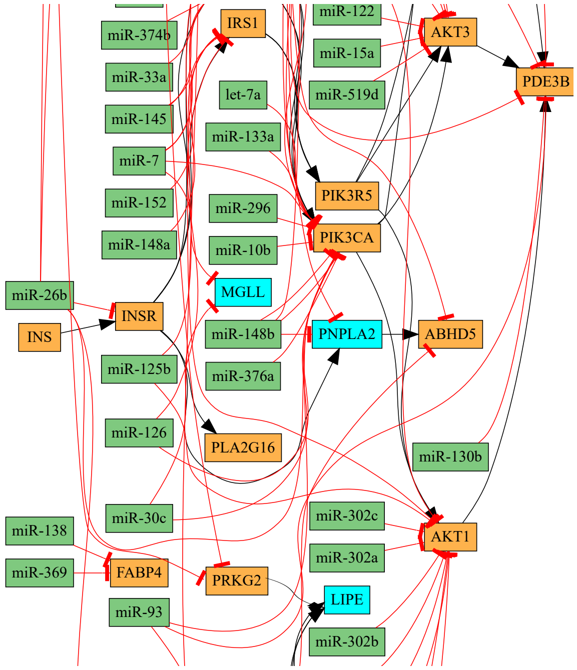


Fig. 5. Portion of the miRNA-augmented *Regulation of lipolysis in adipocytes* pathway.

Ultimately, the *Regulation of lipolysis in adipocytes* pathway describes a unique function of white adipose tissue in which triacylglycerols (TAGs) are broken down into fatty acids and glycerol. Fatty acid (FA) pathways play an important role in cancer[32]. In particular, increased gene expression of AGPAT9(PNPLA2), MAGL(MGLL), and HSL(LIPE), FA metabolism regulators, is associated with increased cancer cells proliferation in colorectal cancer[32] (see blue boxes in Fig. 5). By instance, MAGL pharmacological inhibition attenuated aggressiveness of colorectal cancer cells. On the other hand, decreased gene expression of CD36/FAT regulator has been implicated in contributing to colorectal cancer progression, a higher metastasis grade, and low relapse-free survival[33].

Fig. 5 shows a portion of the *Regulation of lipolysis in adipocytes* augmented pathway obtained from our database (see the complete pathway at `http://bit.ly/hsa04923`).The green boxes show the protein coding genes while the orange boxes display the miRNAs. The black arrows denote activation and the red bar-headed arrows denote repression.

4. Discussion

In this article, we present a method to augment signaling pathways with miRNA-target interactions. The miRNA-augmented pathways (mirAP) offer a more comprehensive view and a deeper understanding of complex diseases. We also present two pipelines that use mirAP to integrate miRNA and mRNA expression data for the purpose of pathway analysis and disease subtyping. As miRNA expression data are becoming freely accessible, miRNA-mRNA integrative analyses are likely to become a routine.

Our pathway analysis pipeline augments gene-gene signaling pathways with miRNA-target interactions. Then we perform a topology-based pathway analysis that takes into consideration both types of molecular data. We analyze 9 sample-matched datasets that were assayed in independent labs. Our pipeline outperforms traditional methods in identifying target pathways (smaller p-values and rankings of the target pathways). We plan to explore methods for augmenting the pathways using only the process(es) described by each given pathway.

Our disease subtyping pipeline combines gene and miRNA expression data, clinical records, and mirAP. The contribution of our disease subtyping pipeline is two-folds. First, this framework introduces a way to exploit the additional information available in biological databases and integrates clinical data, miRNA and gene expression data for disease subtyping. Second, it identifies pathways associated with survival differentiated subgroups of diseases, which bring us closer to the identification of causal pathways associated with survival. We analyze a colorectal cancer data downloaded from TCGA. Our framework provides pathways relevant to survival patterns and subtypes significantly difference between the survival curves. It greatly improves the former approach with p-values $1,000$ times lower than the former. This pipeline is limited by the availability of datasets containing survival records, miRNA, and mRNA expression matched-samples. We plan to extend this study by investigating more diseases and larger datasets.

Acknowledgments

This work was supported by the National Institutes of Health [R01 DK089167, R42 GM087013]; National Science Foundation [DBI-0965741]; and the Robert J. Sokol Endowment in Systems Biology. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

References

1. Y. S. Lee and A. Dutta, *Annual Review of Pathology* **4** (2009).
2. P. Khatri, M. Sirota and A. J. Butte, *PLoS Computational Biology* **8**, p. e1002375 (2012).
3. S. Drăghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichiţa, C. Georgescu and R. Romero, *Genome Research* **17**, 1537 (2007).
4. A. L. Tarca, S. Drăghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic and R. Romero, *Bioinformatics* **25**, 75 (2009).

5. C. Backes, E. Meese, H.-P. Lenhof and A. Keller, *Nucleic Acids Research* **38**, 4476 (July 2010).
6. J. B.-K. Hsu, C.-M. Chiu, S.-D. Hsu, W.-Y. Huang, C.-H. Chien, T.-Y. Lee and H.-D. Huang, *BMC Bioinformatics* **12**, p. 300 (July 2011).
7. I. S. Vlachos, N. Kostoulas, T. Vergoulis, G. Georgakilas, M. Reczko, M. Maragkakis, M. D. Paraskevopoulou, K. Prionidis, T. Dalamagas and A. G. Hatzigeorgiou, *Nucleic Acids Research* **40**, W498 (July 2012).
8. E. Calura, P. Martini, G. Sales, L. Beltrame, G. Chiorino, M. D'Incalci, S. Marchini and C. Romualdi, *Nucleic Acids Research* **42**, p. e96 (2014).
9. S. Nam, M. Li, K. Choi, C. Balch, S. Kim and K. P. Nephew, *Nucleic Acids Research* **37**, W356 (May 2009).
10. P. Martini, G. Sales, M. S. Massa, M. Chiogna and C. Romualdi, *Nucleic Acids Research* **41**, e19 (2013).
11. S. Saria and A. Goldenberg, *IEEE Intelligent Systems* **30**, 70 (2015).
12. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, *Science* **286**, 531 (October 1999).
13. T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler *et al.*, *Proceedings of the National Academy of Sciences* **100**, 8418 (2003).
14. P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag *et al.*, *Breast Cancer Research* **10**, p. R65 (2008).
15. B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains and A. Goldenberg, *Nature Methods* **11**, 333 (2014).
16. T. Nguyen, D. Diaz, R. Tagett and S. Draghici, *Nature Scientific Reports* **6**, p. 29251 (2016).
17. M. Kanehisa and S. Goto, *Nucleic acids research* **28**, 27 (2000).
18. S.-D. Hsu, Y.-T. Tseng, S. Shrestha, Y.-L. Lin, A. Khaleel, C.-H. Chou, C.-F. Chu, H.-Y. Huang, C.-M. Lin, S.-Y. Ho *et al.*, *Nucleic Acids Research* **42**, D78 (January 2014).
19. S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier and S. A. Krawetz, *Genomics* **81**, 98 (2003).
20. U. Von Luxburg, *Statistics and Computing* **17**, 395 (2007).
21. E. Czwan, B. Brors and D. Kipling, *BMC Bioinformatics* **11**, p. 19 (2010).
22. T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W. C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi and R. Edgar, *Nucleic Acids Research* **33**, D562 (2005).
23. S. Davis and P. Meltzer, *Bioinformatics* **14**, 1846 (2007).
24. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth, *Nucleic Acids Research* **43**, e47 (April 2015).
25. Y. Benjamini and D. Yekutieli, *Annals of Statistics* **29**, 1165 (August 2001).
26. B. Ohlsson, M. Truedsson, P. Djerf and F. Sundler, *Regulatory Peptides* **135**, 7 (July 2006).
27. A. Reversi, V. Rimoldi, T. Marrocco, P. Cassoni, G. Bussolati, M. Parenti and B. Chini, *Journal of Biological Chemistry* **280**, 16311 (April 2005).
28. P. Cassoni, T. Marrocco, S. Deaglio, A. Sapino and G. Bussolati, *Annals of Oncology* **12**, S37 (January 2001).
29. C. Pqueux, B. P. Keegan, M.-T. Hagelstein, V. Geenen, J.-J. Legros and W. G. North, *Endocrine-Related Cancer* **11**, 871 (December 2004).
30. M. Moreno-Smith, S. K. Lutgendorf and A. K. Sood, *Future Oncology* **6**, 1863 (December 2010).
31. M. Doulberis, K. Angelopoulou, E. Kaldrymidou, A. Tsingotjidou, Z. Abas, S. E. Erdman and T. Poutahidis, *Carcinogenesis* **36**, p. bgu325 (December 2014).
32. S. Balaban, L. S. Lee, M. Schreuder and A. J. Hoy, *BioMed Research International* **2015**, p. 274585 (2015).
33. S. M. Rachidi, T. Qin, S. Sun, W. J. Zheng and Z. Li, *PLOS ONE* **8**, p. e57911 (March 2013).

FREQUENT SUBGRAPH MINING OF PERSONALIZED SIGNALING PATHWAY NETWORKS GROUPS PATIENTS WITH FREQUENTLY DYSREGULATED DISEASE PATHWAYS AND PREDICTS PROGNOSIS

ARDA DURMAZ*

Systems Biology and Bioinformatics Graduate Program,
Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA
Email: arda.durmaz@case.edu

TIM A. D. HENDERSON*

Department of Electrical Engineering and Computer Science,
Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA
Email: tadh@case.edu

DOUGLAS BRUBAKER

Department of Biological Engineering,
Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139
Email: dkb50@mit.edu

GURKAN BEBEK†

Center for Proteomics and Bioinformatics, Department of Nutrition,
Department of Electrical Engineering and Computer Science,
Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA
Email: gurkan.bebek@case.edu

Motivation: Large scale genomics studies have generated comprehensive molecular characterization of numerous cancer types. Subtypes for many tumor types have been established; however, these classifications are based on molecular characteristics of a small gene sets with limited power to detect dysregulation at the patient level. We hypothesize that frequent graph mining of pathways to gather pathways functionally relevant to tumors can characterize tumor types and provide opportunities for personalized therapies.
Results: In this study we present an integrative omics approach to group patients based on their altered pathway characteristics and show prognostic differences within breast cancer ($p < 9.57E-10$) and glioblastoma multiforme ($p < 0.05$) patients. We were able validate this approach in secondary RNA-Seq datasets with $p < 0.05$ and $p < 0.01$ respectively. We also performed pathway enrichment analysis to further investigate the biological relevance of dysregulated pathways. We compared our approach with network-based classifier algorithms and showed that our unsupervised approach generates more robust and biologically relevant clustering whereas previous approaches failed to report specific functions for similar patient groups or classify patients into prognostic groups.
Conclusions: These results could serve as a means to improve prognosis for future cancer patients, and to provide opportunities for improved treatment options and personalized interventions. The proposed novel graph mining approach is able to integrate PPI networks with gene expression in a biologically sound approach and cluster patients in to clinically distinct groups. We have utilized breast cancer and glioblastoma multiforme datasets from microarray and RNA-Seq platforms and identified disease mechanisms differentiating samples.
Supplementary information: Supplementary methods, figures, tables and code are available at *https://github.com/bebeklab/dysprog.*

*Co-first Author
†Corresponding Author

1. Introduction

Personalized medicine aims to tailor treatment options for patients based on the makeup of their diseases. In the case of cancer, the genetic makeup of tumors is characterized to identify unique tendencies and exploit vulnerabilities of these tumors. However, identifying genomic alterations and molecular signatures that better describe or classify cancer to accomplish this goal has been challenging. Furthermore complex disease phenotypes, such as cancer, cannot be fully explained by individual genes and mutations. Recent studies have explored various approaches to uncover the molecular network signatures of cancers including multivariate linear regression[1] or factor graphs[2] to combine information flow based approaches with copy numbers and DNA methylation data. These techniques identified patient loci with high risk of disease along with genes that are dysregulated for various cancers.[3,4] Gene expression profiles and (in some cases) DNA methylation or metabolomics data have also been used to identify subtypes of the disease.[3–7] However prognostic classification of tumors still requires attention and it is an important step toward identifying most effective approaches in precision medicine.

Glioblastoma multiforme (GBM) is the most common form of malignant brain tumor in adults. GBM is characterized by a median survival of one year and an overall poor prognosis.[8] There have been numerous attempts to classify GBM by differential gene expression to identify clinically and prognostically relevant subtypes.[9,10] Previously methylation status of the *MGMT* promoter is suggested to be associated with tumor response of gliomas to alkylating agents and later associated with increased survival.[11,12] More recently The Cancer Genome Atlas (TCGA) project also provided supporting findings of the methylation status of the *MGMT* promoter as a prognostic marker through analysis of high dimensional data for 206 GBM tumors.[13] Further work utilizing the TCGA data classified GBM by aberrations and gene expression of *EGFR, NF1, and PDGFRA/IDH1* into four subtypes, Classical, Mesenchymal, Neural, and Proneural.[14] These classifications implied strong relationships between subtypes and neural lineages as well as response to aggressive therapy. Though these studies introduced GBM classification, there remained a need to classify dysregulations in tumors more specifically by survivability. While earlier approaches have focused on identifying gene sets,[10,15–18] these had little impact on finding dysregulated pathway segments. For instance, using nearest shrunken centroid classification method,[18,19] or clustering algorithms,[14] gene sets that stratify samples were identified, yet functionally these were not strongly related. Hence, they present little potential for improved treatment opportunities for patients.

Breast Invasive Carcinoma (BRCA) is the most diagnosed cancer among woman consisting of multiple sub classes with distinct clinical outcomes. Previously, 5 subtypes were identified using expression profiles of and later applied to develop predictors by manually selected genes.[6,20,21] Consecutive studies identified differing number of subtypes similar to inital identification. For instance using expression profiles Sotiriou *et al.* identified 6 subtypes further separating luminal-like and basal-like groups.[22,23] Furthermore a comprehensive study integrating multiple omics data to identify unified classification of the breast cancer samples provided strong evidence for 4 subtypes; *Basal, Her-2 enriched, Luminal-A, Luminal-B*.[4] However studies incorporating network or pathway information either used manual selection of pathways or produced limited results. For instance Gatza *et al.* identified 17 subgroups using pathway based classification with mixed intrinsic subtype signatures.[24]

We describe an integrative omics approach based on frequent subgraph mining (FSM) that brings Protein-Protein Interaction (PPI) networks and gene expression data together to infer molecular networks that are dysregulated in patient samples. We tested our approach using gene expression data for both glioblastoma and breast cancer datasets collected with microarray and next generation sequencing (NGS) approaches. The networks inferred from FSM not only stratify patients into clinically-relevant subtypes, but also provides significant prognostic differences. Our results suggest that a network-based stratification of patients is more informative than using gene-level or feature-based data integration. Identifying personalized dysregulated signaling networks will offer effective means to diagnose and treat patients.

2. Methods

The proposed method uses a novel approach to integrate mRNA expression profiles and PPI networks to identify personalized dysregulated signaling pathways. We hypothesize that dysregulated sub-pathways observed in cancer can discriminate between tumors types which lead to different patient outcomes. We utilized publicly available datasets to develop and validate a method to detect altered molecular signatures in canonical pathways. Our classifications better distinguish patient prognosis in biologically relevant terms than previous studies.[14,25,26]

Our approach is to construct personalized networks of PPIs for cancerous tumors based on mRNA expression data. Section 2.1 details the construction of these networks called *dysregulated signaling pathways*. A network is constructed for each of the patients in each of the datasets used in Section 3. Personalized networks are mined using a new algorithm called QSPLOR (queue explorer) to identify a subset of frequently occurring subgraphs with 4 to 8 proteins as detailed in Sections 2.2 and 2.3. Finally, Non-Negative Matrix Factorization is used to cluster the patients via the frequently occurring subgraphs (Section 2.4 and 2.5).

In Section 3 the clusters are shown to separate patients into short-term and long-term survival groups. The methodology presented has the potential to stratify patients based on their molecular signatures, improve delivery of therapies and assist clinicians and researchers alike to better assess patient prognosis.

2.1. *Dysregulated Signaling Pathways*

Dysregulated Signaling Pathways are labeled graphs (Section 2.2) where vertices represent proteins and edges represent dysregulated activation/inhibition interactions. They are constructed from mRNA expression data (Section 3) and known PPI data.[27,28]

Dysregulation is computed by constructing a matrix $\mathbf{P}$, where $\mathbf{P}_{i,a}$ is the standard score of expression level of gene a for patient i. Then an *interaction matrix* $\mathbf{S}$ constructed from $\mathbf{P}$ in Equation 1. In Equation 1 (ab) represents two genes a and b such that the protein encoded by a interacts with the protein encoded by b. The variable i represents a particular patient.

$$\mathbf{S}_{(ab),i} = \sqrt{\mathbf{P}_{i,a}^2 + \mathbf{P}_{i,b}^2} \tag{1}$$

To determine if the relationship between two genes a and b is dysregulated for patient i the *z-score* for each interaction is computed. In Equation 2, $\mu(\mathbf{S}_{(ab),\cdot})$ and $\sigma(\mathbf{S}_{(ab),\cdot})$ respectively refer to the mean and standard deviation of the dysregulation scores for genes a and b.

$$Z(\mathbf{S})_{(ab),i} = \frac{\mathbf{S}_{(ab),i} - \mu(\mathbf{S}_{(ab),\cdot})}{\sigma(\mathbf{S}_{(ab),\cdot})} \tag{2}$$

If $Z(\mathbf{S})_{(ab),i} > c$ then an edge $a \rightarrow b$ is included in the graph for patient i indicating a and b are dysregulated. In Section 3 the constant c, the z-score threshold, was set to 2 to mine for dysregulation.

2.2. *Frequent Subgraph Mining*

Frequent Subgraph Mining (FSM) is a data mining technique which looks for repeated subgraphs in a graph database. As in Inokuchi *et al.*,[29] the database $\mathcal{D}$ is a set of transactions where each "transaction" is the dysregulated signaling pathways for a patient. FSM detects signaling sub-pathways which are dysregulated in multiple patients.

A dysregulated signaling pathway is a directed labeled graph G consisting of a set of vertices V, a set of edges $E = V \times V$, a set of labels L, and a labeling function which maps vertices (or edges) to labels $l : V|E \rightarrow L$. A graph $H = (V_H, E_H, L, l)$ is a subgraph of $G = (V_G, E_G, L, l)$ if $V_H \subseteq V_G$ and $E_H \subseteq E_G$.

A graph H is a subgraph of G ($H \sqsubseteq G$) if there is an injective mapping $m : V_H \rightarrow V_G$ s.t.

(1) All vertices in H map vertices in G with the same label: $\forall\ v \in V_H\ [l(v) = l(m(v))]$

(2) All edges match: $\forall\ (u, v) \in E_H\ [(m(u), m(v)) \in E_G]$

(3) All edge labels match: $\forall\ (u, v) \in E_H\ [l(u, v) = l(m(u), m(v))]$

Such a mapping m is known as an *embedding.* The problem of determining if a graph H is a subgraph of G is called the *subgraph isomorphism problem* and is NP-Complete.[30] The *frequency* of a subgraph H is the number of graphs (transactions) in $\mathcal{D}$ which H *embeds* into.

The subgraph relationship $\cdot \sqsubseteq \cdot$ induces a *partial order* on the subgraphs of the graphs in $\mathcal{D}$. That partial order is referred to as the *subgraph lattice.* If the subgraphs in the lattice are all *connected* it is known as the *connected subgraph lattice.* The connected subgraph lattice of $\mathcal{D}$ can be viewed as a graph $\mathcal{L}_{\mathcal{D}} = (V_{\mathcal{L}}, E_{\mathcal{L}})$. The vertices $V_{\mathcal{L}}$ are all of the connected subgraphs of G. If u and v are both vertices of $\mathcal{L}_{\mathcal{D}}$ then there is an edge between u and v if and only if $u \sqsubseteq v$ and v and be constructed from u by adding one edge and at most one vertex. The k *frequent connected subgraph lattice* k-$\mathcal{L}_{\mathcal{D}}$ contains only those subgraphs of graphs in $\mathcal{D}$ which are present in at least k graphs in the graph database $\mathcal{D}$. The leaf nodes of the k-$\mathcal{L}_{\mathcal{D}}$ are the *maximal frequent subgraphs.*

The objective of frequent subgraph mining is to discover the vertices of k-$\mathcal{L}_{\mathcal{D}}$. If a subgraph does have at least k transactions it is embedded in, it is known as a *frequent subgraph.* Since finding a frequent subgraph requires repeated subgraph isomorphism queries the problem complexity of FSM is exponential. The number of steps in frequent subgraph mining is bounded from above by $\mathcal{O}(2^g g^h)$ where g is the size of the graph and h is the size of the largest frequent subgraph. The term 2^g is an upper bound on the number of subgraphs of g. Tighter bounds can be obtained if one has more specific knowledge of the graph. The term g^h is an upper bound on number of steps to check if a graph of size h is a subgraph of g.

We present QSPLOR, a new algorithm to find a subset of frequent subgraphs in Section 2.3. It is used to find frequently dysregulated signaling sub-pathways. QSPLOR uses a fixed

```
1  # param start: frequent single vertex subgraphs
2  # param score: a function to score queue items
3  # param max_size: the max size of the queue
4  # param min_sup: int, amount of support
5  # returns: a generator of frequent subgraphs
6  def qsplor(start, score, min_sup):
7      while not start.empty():
8          queue = [ start.pop() ]
9          while not queue.empty()
10             lattice_node = take(queue, score)
11             kids = lattice_node.extend(min_sup)
12             for ext in kids: add(queue, score, ext, max_size)
13             yield subgraph
14 def add(queue, score, item, max_size):
15     queue.append(item)
16     while len(queue) >= max_size:
17         i = argmin(score(idx, queue) for idx in sample(10, len(queue)))
18         queue.drop(i)
19 def take(queue, score):
20     i = argmax(score(idx, queue) for idx in sample(10, len(queue)))
21     return queue.take(i)
```

Fig. 1. QSPLOR: a new algorithm for mining a subset of frequent subgraphs.

amount of memory and a user defined scoring heuristic to guide the search. The algorithm only reports the maximal frequent subgraphs found for compactness. We report only a subset, and not all of frequently dysregulated signaling pathways because (i) it is much faster to report only some of the frequent subgraphs and (ii) using a greater number of frequent subgraphs does not necessarily lead to a more discriminating clustering of samples in our analysis.

There have been a variety of FSM algorithms developed over the last two decades and there are several recent surveys available.[31,32] In recent years interest in collecting representative subsets of frequent subgraphs has emerged.[33,34] Both studies employ random walks on the frequent connected subgraph lattice to collect a sample of the frequent subgraphs. Finally, Leap Search[35] was proposed to find interesting patterns as defined by an objective function.

2.3. *QSPLOR: Mining a Subset of Frequent Subgraphs*

Figure 1 shows pseudo code for QSPLOR a new algorithm to mine a subset of frequent subgraphs. It proceeds as a graph traversal of k-$\mathcal{L}_{\mathcal{D}}$ (the k frequent connected subgraph lattice of the graph database). It begins the traversal at each lattice node representing a frequent subgraph containing only one vertex. At each outer step it initializes a queue with one of the starting lattice nodes. Then in each inner step it removes an item of the queue. The `take` function removes one item from a uniform sample of the queue such that a user supplied scoring function is maximized.

On line 11, the lattice node is extended. This involves finding all possible one edge extensions to the subgraph represented by the lattice node. The ones that are frequent are returned by the extend method. After the extensions are found they are added to the queue with the `add` method. If the queue is at the maximal size after the addition, one item from the queue is dropped. The dropped item is from a uniform sample of the queue and minimizes the user supplied score function. After all extensions have been processed the subgraph is output.

The key to our algorithm is the user supplied scoring function which guides the traversal. The simplest scoring function simply returns a uniform random number. This will cause the traversal to be unguided. Complex scoring functions can prioritize certain labels or structures.

The best general scoring functions are those that prioritize *queue diversity* such that traversal is encouraged to explore as much of the lattice as possible. We use a distance function which captures both structural and labeling differences between graphs as the scoring function for this paper. See the supplementary methods for more details on QSPLOR.

2.4. *Non-Negative Matrix Factorization*

Clustering via Non-Negative Matrix Factorization (NMF) is used to partition patients into subgroups. Section 3 shows that the partitions are prognostically discriminative between the patient subgroups. NMF method was first proposed by Lee and Seung[36] with the aim of decomposing images into explanatory basis vectors. NMF has also been used on gene expression data.[37] For a description of our usage of NMF see the supplementary methods.

2.5. *Clustering Metrics*

Use of NMF requires careful evaluation of the results. Since NMF is based on random initialization of the initial stratification we have applied consensus clustering approach. Using R package NMF[38] we have applied method *'nsNMF'* and random seed with 150 runs. To identify best clustering rank k cophenetic correlation coefficient, silhouette values, residual metrics are evaluated. Cophenetic correlation coefficient is first suggested by Brunet *et al.*[37] to quantify the stability of the clusters. It is calculated as the correlation between sample distances obtained from consensus matrix and the cophenetic distances obtained from hierarchical clustering of the consensus matrix. Brunet *et al.* suggested to choose the ranks where cophenetic correlation coefficient starts to decrease. Silhouette is another method for quantifying cluster stability.[39] The values range between -1 and 1. Intuitively the average silhouette value represents how similar each sample is to the cluster the sample belongs to and how distant from neighbor clusters. Clustering with silhouette values > 0.7 are considered strong as patterns. Residual is the error of the NMF method. Since the method produces an approximation of the original matrix, the residuals represent how close the factorization is to the original data. Note that the residuals decrease naturally as the rank of factorization increases since more variables are added to represent the original matrix.

2.6. *Data Sources*

PPI networks were downloaded from Reactome(v56). Reactome is an expert curated publicly available repository which stores multiple types of relations including reactions, indirect and direct complexes.[27,28] Gene expression data was obtained from previously published studies and TCGA using UCSC Cancer Browser.[40] Clinical data is obtained from both TCGA and corresponding publications (See Figure 2).

3. Results

3.1. *Breast Cancer (Microarray)*

Curtis *et al.*[41] used genomic variations to identify novel subgroups in breast cancer and validated on a sample of 995 patients. Using the same discovery dataset we were able to identify 5 groups with significant differences in survival. QSPLOR mined 145 sub-pathways, with 4-8 proteins each, dysregulated in at least 25 patients.

Fig. 2. Summary of Data including sample and network numbers, median days and interquartile range, sample count of alive and dead event status. In this study both microarray (MA) and RNA-Seq data for breast cancer (BRCA) (MA: [41] and RNA-Seq:[4]) and late stage brain tumors (GBM) (MA:[14] and RNA-Seq:[42]) was utilized.

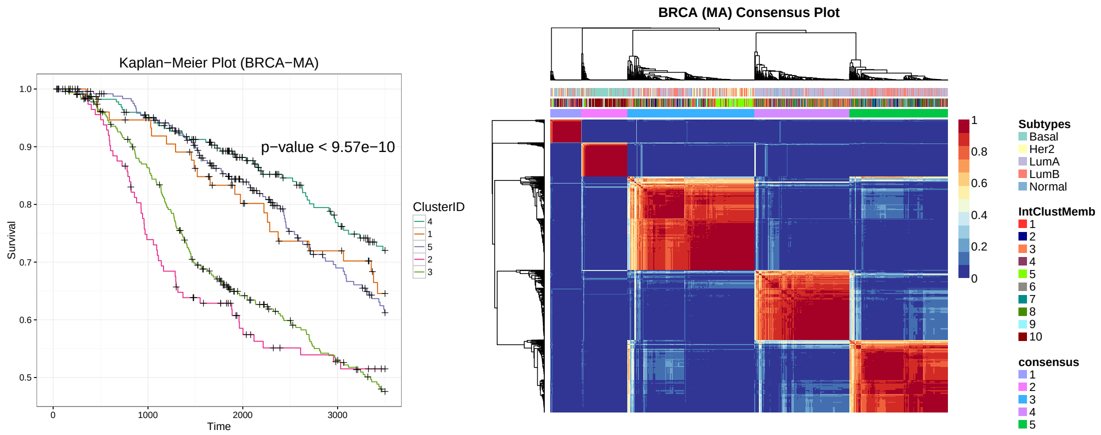
DataSet	Patients	Sub-Pathways	Median Days	Alive/Dead
BRCA MA	995	145	1449	645/350
BRCA RNA-Seq	200	200	1230	685/106
GBM MA	197	553	375	22/175
GBM RNA-Seq	163	548	335	50/113

Consensus clustering and utilization of clustering metrics identified 5 patient groups. The clustering results are similar to clustering of patient samples reported in Curtis *et al.*[41] Identified clusters 1 and 2 matched with clusters 10 and 5 respectively in Curtis *et al.* study as shown in Figure 3b. Furthermore given clusters also match with Basal and Luminal B intrinsic subtypes with further stratification. Compared to previously established subtypes based on the PAM50 classifier, identified clusters are significantly separated in terms of survival(Figure 3a). Enrichment analysis for Reactome pathways in short survivor group revealed pathways that are functionally relevant or predictor of poor survival, i.e. Nonsense-Mediated Decay (NMD),[43] SRP Dependent cotranslational protein targeting to membrane,[44] Selenocysteine synthesis,[45] Signaling by WNT.[46] In contrast, long survivor group was enriched in Neuronal System,[1,45] GABA receptor activation,[47] Signaling by GPCR[48] (See Supplementary Tables S1-S5).

3.2. *Breast Cancer (RNA-Seq)*

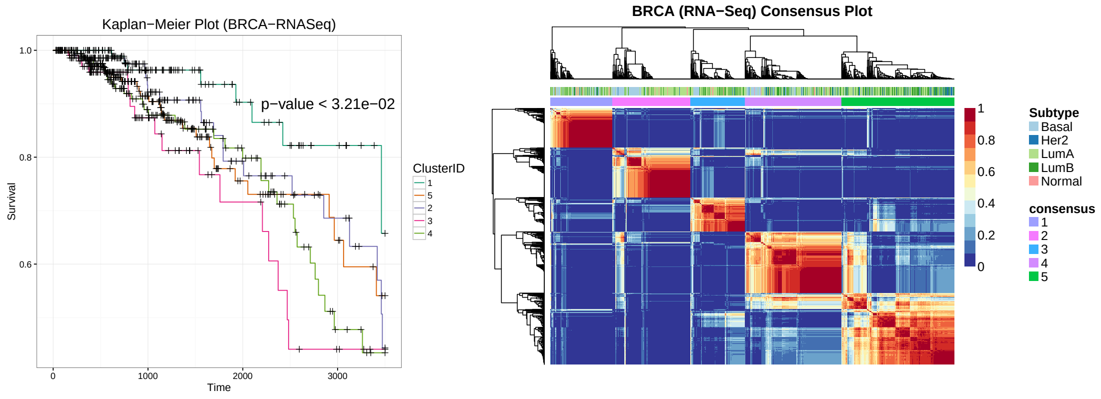
To test the proposed method on breast cancer with data from a different platform, we obtained 791 RNA-Seq samples from TCGA with matching clinical data. QSPLOR identified 200 dysregulated subgraphs. Note that the dataset was not filtered based on prior treatment or patient characteristics hence a heterogeneous dataset was utilized in contrast with breast cancer microarray dataset above. The clustering identified 8 clusters based on cophenetic correlation coefficient and silhouette values. However 8 clusters did not result in significant survival differences hence we have utilized 5 clusters to test whether informative groups were obtained with significant survival differences ($p < 0.05$) (Figure 4a). Reactome pathway enrichment for short survivor group resulted in processes related to cellular division; Mitotic Prometaphase, Separation of Sister Chromatids, Activation of ATR in response to replication stress. Furthermore APC/C-mediated degradation of cell cycle proteins and mitotic proteins pathways were significantly dysregulated. Long survivor group was enriched in immune system related processes; MHC class II antigen presentation, TCR signaling, Cytokine signaling.

We have applied the subgraphs found in microarray dataset to RNA-Seq dataset to check cross-platform application of the proposed method. We were able to identify 5 clusters with significant survival differences. The identified clusters 3 and 4 matched previously identified Basal and Her2 subtypes respectively with further stratification (Figure S16). Pathway enrichment for short and long survivor groups resulted in Keratin metabolims, Signaling by Rho GTPases, Signaling by WNT, Gastrin-CREB signaling pathway via PKC and MAPK, Axon guidance for short survivor group and Signaling by GPCR, EGFR, VEGF, FGFR4, Interleukin-2 signaling for long survivor group (See Supplementary Tables S11-S15).



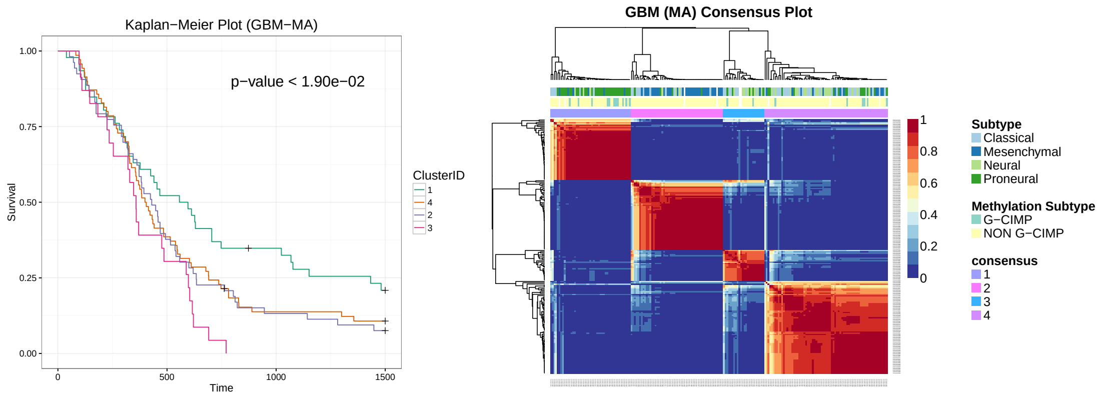
(a) (b)

Fig. 3. Results for breast cancer data analysis used in Curtis *et al.*.[41] **(a)** The Kaplan-Meier plot for 5 groups are shown (Log-rank test $p-value < 9.57E-10$).The x-axis represents days of survival. **(b)** Consensus clustering obtained using NMF is shown. Top bars show novel subtypes clusters, intrinsic subtypes and classification. IntClustMemb shows clusters identified in the Curtis *et al.* study



(a) (b)

Fig. 4. **(a)** Kaplan-Meier and consensus clustering results for breast cancer data obtained from TCGA (Log-rank test $p - value < 3.21E - 02$). Survival is represented as days. **(b)** Top bar in figure shows intrinsic subtypes previously defined, lower bar shows our novel pathway based groups.



(a) (b)

Fig. 5. (a) Survival and consensus clustering results for glioblastoma multiforme microarray data used in.[14] Survival is represented as days and there is a significant difference (Log-rank test $p - value < 1.9E - 02$). (b) Top bar in consensus clustering shows previous classification of GBM patients.

3.3. *Glioblastoma Multiforme (Microarray)*

Using 11861 genes from GBM microarray dataset[14] our method revealed 4 clusters with statistically significant stratification in survival curves ($p-value < 0.05$). The long survivor group 1 consists mostly of proneural subtypes, which also supports the biological implication of our method. A new stratification is visible in Figure 5b for the short survivor group 3.

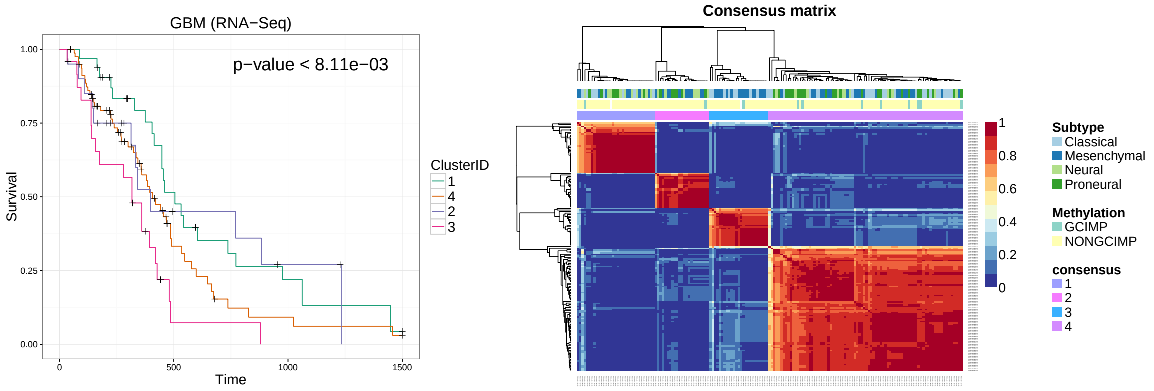
To identify biological implications, we conducted over-representation analysis for Reactome pathways. The long survivor group revealed pathways related to extracellular matrix organization and immune system; axon guidance, collagen degradation, TNFSF mediated activation cascade. The short survivor group was enriched in cell cycle related pathways including: replication, strand elongation and repair. Group 2 shows enrichment for trafficking of GPCR signaling, the Glutamate neurotransmitter release cycle, signaling by Wnt, Gastrin-CREB signaling pathway via PKC and MAPK. Group 4 shows enrichment for respiratory electron transport chain, mitochondrial translation and translation related processes. Overall, the analysis suggests new targets to study for GBM therapy (See Supplementary Tables S16-S19).

3.4. *Glioblastoma Multiforme (RNA-Seq)*

Using GBM data from TCGA[42] which included 15739 genes, our method revealed 4 groups with significant survival *(p-value <0.01)* stratification clustered based on 548 identified subgraphs. As in the microarray data analysis, mesenchymal groups were mostly clustered together in group 3 including the classical subtype. Group 4 is comprised of multiple subtypes suggesting a new classification scheme (Figure 6b). Pathway enrichment results may reveal new biomarkers. Short survivor group 3 was enriched in processes related to cell division; Mitotic prometaphase, Separation of Sister Chromatids, G2/M Transition, DNA Replication. In contrast, long survivor group 1 based on 1 year survival is enriched in Assembly of the primary cilium, Cytokine Signaling in Immune System, Gastrin-CREB Signaling pathway via PKC and MAPK, VEGFA-BEGFR2 Pathway and RET signaling. Interestingly Assembly of the primary cilium is found to be associated with GBM tumors[49,50] (See Supplementary Tables S20-S23).

4. Validation

We compared our method against 2 recently published work integrating PPI and pathway information; *Pathifier* and *NCIS*. (Details of the methods are given in supplementary document) Pathifier identified 6 groups with significant differences in survival (Figure S14a). The number of samples in each group does not suggest biologically relevant clustering (n = 6, and the larger clusters are not significant in terms of survival). The separation distances between groups are not robust with cophenetic correlation coefficient 0.61(Figure S14b). NCIS[25] identified 4 previously established subtypes in the GBM microarray dataset in conjunction with a curated PPI network. The network was constructed by the authors from Reactome, NCI-Nature Curated PID, and KEGG. It consists of 11,648 genes, 211,794 interactions matching 7,183 genes in the GBM dataset. The identified subtypes are similar to established subtypes and have significant differences in survival. However, it is not clear how the patients are clus-



(a)　　(b)

Fig. 6. (a) Kaplan-Meier and (b) consensus clustering results for glioblastoma multiforme samples obtained from TCGA. The RNA-Seq data set showed significant survival difference (Log-rank $p - value < 8.11E - 03$)

tered since previously identified subtypes do not provide overall significant survival difference (Figure S4). Using the data from NCIS study we have identified 5 clusters (based on the clustering metrics) which show separation of survival curves (Figure S15a). We were able to cluster previously proposed mesenchymal and proneural subtypes with further stratification of mesenchymal group (Figure S15b). Based on the survival analysis, proneural clustered groups show the longest survival curves in agreement with previous findings. These results suggest that the proposed method performed better than the NCIS and Pathifier algorithms in terms of significance of survival stratification and relevance of the identified genes and pathways which can be used as precursor targets for future therapeutic studies.

5. Discussion

The proposed method aims to integrate PPI data with gene expression data using a novel approach. In this study we were able to identify networks that play predictive role in clinical outcome and also networks that crosstalk between the established pathways. A crucial development for improving current prognostic methodologies. The presented method is also more general as it does not require apriori identification of important genes.

Several studies have investigated molecular correlation of prognosis and clinical subclasses in GBM. Earlier studies have identified tumor grade as one of the strong predictors of disease outcome,[51] such as *TP53* mutation and *EGFR* amplifications were claimed to stratify patients into subgroups,[52,53] while a later study contests the validity of this classification.[54] Further studies have identified various gene sets that would separate the patient samples by their molecular characterization,[10,15–18] and some have reported prognostic value of these gene sets. However, most of these have identified different sets of genes, a consensus on the functional delivery has not been reached. These proposed subtype classification methods also identified different sets of patient subtypes, classifications greatly rely on selected patient groups and sample size.

Overall the results suggest possible targets and pathways for cancer progression, mechanisms and survival. Additionally enrichment using long and short survivor groups from RNA-Seq data resulted in similar gene targets. Note that results are 'reversed' for RNA-Seq dataset compared to microarray analyzed samples, however since the stratification is based on dysregulation, the method includes both overexpression or underexpression. Hence genes are categorized as possible markers rather than specific targets for long or short survival.

Our validation of the results we presented here, which reproduced similar survival curves over independent studies, presents great potential for prognostic value for this method. Moreover, finding significant mechanisms that can describe the underlying effects of survival and treatment responses can be easily done within these parameters and provide candidate pathways for therapeutic intervention. While follow up studies are needed to further asses the prognostic value, and possible effect of treatments, analysis that we have conducted provide an initial look of the biological mechanisms underlying in these patient groups with different survival which are also supported by various studies.

Gathering multiple omics datasets to better characterize individuals and associating these with extensive phenotype information has been the hallmark achievement of recent years.[3,4,14,41,42] These datasets have paved the road to improved personalized medicine, promising better disease characterization and diagnosis, identification of patient-specific treatment options and improved monitoring of patients in need. While personalized medicine offers great benefit to individuals, the computational approaches to integrate these multiple omic datasets and statistical methods to leverage the underlying disease and patient traits is still under development. This study tackled this problem of integration network data with transcriptomics data to identify classification scheme for both breast and late stage brain tumors (GBM). Our method can be used to group patients in an unsupervised manner, and have prognostic value. The significant separation of patient samples will allow further studies and utility, since these classifications are based on functionally related frequently altered pathway segments. In the future, we plan to investigate the utility of this method for other cancer types, integrating additional genomic features and investigate its value in improving treatment options.

Acknowledgments

Thank you Leigh Henderson for thoughtful discussions and reading drafts of this paper. This research was partially supported by a Grant from NIH/NCRR CTSA KL2TR000440 to GB.

References

1. Q. Li *et al.*, *Cell* **152**, 633 (Jan 2013).
2. C. J. Vaske *et al.*, *Bioinformatics* **26**, i237 (2010).
3. TCGA, *Nature* **474**, 609 (2011).
4. TCGA, *Nature* **490**, 61 (2012).
5. K. Holm *et al.*, *Breast Cancer Res* **12**, p. R36 (2010).
6. T. Sørlie *et al.*, *PNAS* **98**, 10869 (2001).
7. S. Tardito *et al.*, *Nat Cell Biol* **17**, 1556 (Dec 2015).
8. H. Ohgaki and P. Kleihues, *Acta neuropathologica* **109**, 93 (2005).
9. Y. Liang *et al.*, *PNAS* **102**, 5814 (2005).
10. C. L. Nutt *et al.*, *Cancer research* **63**, 1602 (2003).
11. M. Esteller *et al.*, *New England Journal of Medicine* **343**, 1350 (2000).

12. M. E. Hegi *et al.*, *New England Journal of Medicine* **352**, 997 (2005).
13. TCGA, *Nature* **455**, 1061 (Oct 2008).
14. R. G. Verhaak *et al.*, *Cancer cell* **17**, 98 (2010).
15. H. Colman *et al.*, *Neuro-oncology* , p. nop007 (2009).
16. W. A. Freije *et al.*, *Cancer research* **64**, 6503 (2004).
17. J. M. Nigro *et al.*, *Cancer research* **65**, 1678 (2005).
18. H. S. Phillips *et al.*, *Cancer cell* **9**, 157 (2006).
19. R. Tibshirani *et al.*, *PNAS* **99**, 6567 (2002).
20. C. M. Perou *et al.*, *Nature* **406**, 747 (2000).
21. J. S. Parker *et al.*, *J Clin Oncol* **27**, 1160 (Mar 2009).
22. C. Sotiriou *et al.*, *PNAS* **100**, 10393 (2003).
23. C. Fan *et al.*, *New England Journal of Medicine* **355**, 560 (2006).
24. M. L. Gatza *et al.*, *PNAS* **107**, 6994 (2010).
25. Y. Liu *et al.*, *BMC bioinformatics* **15**, p. 1 (2014).
26. Y. Drier, M. Sheffer and E. Domany, *PNAS* **110**, 6388 (2013).
27. D. Croft *et al.*, *Nucleic acids research* **42**, D472 (2014).
28. M. Milacic *et al.*, *Cancers* **4**, 1180 (2012).
29. A. Inokuchi *et al.*, An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data, in *Principles of Data Mining and Knowledge Discovery*, jul 2000 pp. 13–23.
30. S. A. Cook, The complexity of theorem-proving procedures, in *ACM Symposium on Theory of Computing*, (ACM Press, New York, New York, USA, 1971).
31. C. Jiang, F. Coenen and M. Zito, *The Knowledge Engineering Review* **28**, 75 (mar 2013).
32. H. Cheng, X. Yan and J. Han, Mining Graph Patterns, in *Frequent Pattern Mining*, (Springer International Publishing, 2014) pp. 307–338.
33. V. Chaoji, M. Al Hasan, S. Salem, J. Besson and M. J. Zaki, *Stat. Anal. Data Min.* **1**, 67 (2008).
34. M. Al Hasan and M. J. Zaki, Output Space Sampling for Graph Patterns, in *Proceedings of VLDB*, (VLDB Endowment, aug 2009).
35. X. Yan, H. Cheng, J. Han and P. S. Yu, Mining Significant Graph Patterns by Leap Search, in *Proceedings of ACM SIGMOD ICMD*, 2008.
36. D. D. Lee and H. S. Seung, *Nature* **401**, 788 (1999).
37. J.-P. Brunet, P. Tamayo, T. R. Golub and J. P. Mesirov, *PNAS* **101**, 4164 (2004).
38. R. Gaujoux and C. Seoighe, *BMC bioinformatics* **11**, p. 1 (2010).
39. P. J. Rousseeuw, *Journal of computational and applied mathematics* **20**, 53 (1987).
40. J. Z. Sanborn *et al.*, *Nucleic acids research* , p. gkq1113 (2010).
41. C. Curtis *et al.*, *Nature* **486**, 346 (Jun 2012).
42. C. W. Brennan *et al.*, *Cell* **155**, 462 (Oct 2013).
43. L. B. Gardner, *Mol Cancer Res* **8**, 295 (Mar 2010).
44. J. Simões, F. M. Amado, R. Vitorino and L. A. Helguero, *Oncoscience* **2**, 487 (2015).
45. R. L. Schmidt and M. Simonović, *Croat Med J* **53**, 535 (Dec 2012).
46. G.-B. Jang *et al.*, *Sci Rep* **5**, p. 12465 (2015).
47. S. Z. Young and A. Bordey, *Physiology (Bethesda)* **24**, 171 (Jun 2009).
48. A. Singh, J. J. Nunes and B. Ateeq, *Eur J Pharmacol* **763**, 178 (Sep 2015).
49. J. J. Moser, M. J. Fritzler and J. B. Rattner, *BMC cancer* **9**, p. 448 (2009).
50. J. J. Moser, M. J. Fritzler and J. B. Rattner, *BMC clinical pathology* **14**, p. 1 (2014).
51. M. D. Prados and V. Levin, Biology and treatment of malignant glioma., in *Semin Oncol*, 2000.
52. A. von Deimling, D. N. Louis and O. D. Wiestler, *Glia* **15**, 328 (1995).
53. K. Watanabe *et al.*, *Brain pathology* **6**, 217 (1996).
54. Y. Okada *et al.*, *Cancer research* **63**, 413 (2003).

HUMAN KINASES DISPLAY MUTATIONAL HOTSPOTS AT COGNATE POSITIONS WITHIN CANCER

JONATHAN GALLION[†]

Structural Computational Biology and Molecular Biophysics, Baylor College of Medicine, One Baylor Plaza
Houston, TX, 77030, USA
Email: Jonathan.Gallion@bcm.edu

ANGELA D. WILKINS

Immunology, Lichtarge Laboratory BCM, One Baylor Plaza
Houston, TX, 77030, USA
Email: aw11@bcm.edu

OLIVIER LICHTARGE

Structural Computational Biology and Molecular Biophysics, Baylor College of Medicine, One Baylor Plaza
Houston, TX, 77030, USA
Email: lichtarge@bcm.edu

The discovery of driver genes is a major pursuit of cancer genomics, usually based on observing the same mutation in different patients. But the heterogeneity of cancer pathways plus the high background mutational frequency of tumor cells often cloud the distinction between less frequent drivers and innocent passenger mutations. Here, to overcome these disadvantages, we grouped together mutations from close kinase paralogs under the hypothesis that cognate mutations may functionally favor cancer cells in similar ways. Indeed, we find that kinase paralogs often bear mutations to the same substituted amino acid at the same aligned positions and with a large predicted Evolutionary Action. Functionally, these high Evolutionary Action, non-random mutations affect known kinase motifs, but strikingly, they do so differently among different kinase types and cancers, consistent with differences in selective pressures. Taken together, these results suggest that cancer pathways may flexibly distribute a dependence on a given functional mutation among multiple close kinase paralogs. The recognition of this "mutational delocalization" of cancer drivers among groups of paralogs is a new phenomena that may help better identify relevant mechanisms and therefore eventually guide personalized therapy.

[†] The authors gratefully acknowledge support from the National Institutes of Health (GM066099 and GM079656), from the National Science Foundation (DBI-1356569), and from DARPA (N66001-15-C-4042)

1. Introduction

A major focus of recent cancer sequencing projects, such as the TCGA, is to identify causal driver mutations responsible for tumorigenesis (*1*) . To this end, many computational tools have been produced to predict the impact of mutations on protein function in order to screen out null function or low impact mutations (*2*). The efforts of these approaches have identified many proteins and mutations driving cancer progression. Unfortunately, the inherent mutational heterogeneity displayed within cancer often limits the statistical power of these methods so as to capture only the most frequent driver mutations in a large cohort of patients (*3*). By contrast, low frequency drivers or smaller patient cohorts suffer from a lack of statistical significance and are therefore easily missed.

While infrequent mutations in a single gene may, at first glance, appear to indicate insignificance in cancer progression, this may be an oversimplification. Driver mutations in cancer may not only target a single gene but rather groups of genes or functional pathways, distributing the mutational burden across many functionally related genes (*4, 5*); while a single gene may lack significance, combining mutations across a regulatory pathway can increase the power of the analysis and identify gene groups driving cancer progression (*3, 6*). Prior studies have taken these groups from databases such as KEGG (*7*), Reactome (*8*), and analyses of gene association networks like STRING (*9*). However, these approaches are not limited to functional or hierarchical pathways but rather could be applied to any group of proteins that share functionality such as, Gene Ontology terms or even groups of protein homologs sharing significant functional overlap.

Further confounding the prediction of cancer drivers, single gene analyses group mutations regardless of their structural location and, therefore, do not account for the functional heterogeneity of these mutations. To account for these difference, an analysis in Colon and Breast Cancers grouped mutations from various genes occurring in homologous protein domains, finding specific domains enriched for high frequency mutations across many individual proteins (*10*). Furthermore, an analysis of disease-related mutations across all human kinases showed that these mutations preferentially localized in specific structural domains, affected certain residues types, and had conserved amino acid substitutions (*11*). These studies show disease-related mutations can preferentially occur at specific structural domains in homologous proteins, such as kinases, and that kinase mutations share conserved patterns of substitution. Here, we expand upon this work and ask whether there are mutational biases in individual positions in the context of cancer.

For the purpose of this study, we focus on human kinases in order to better understand this essential protein family and how it contributes to cancer. There are over 500 human kinases sharing substantial homology in both the kinase structure and the catalytic mechanism (*12*). The kinase family has been further subdivided into 7 classes based on substrate specificity and evolutionary lineage. Kinases are ubiquitous proteins involved in a diverse array of cellular functions; as a result, numerous perturbations in kinase coding regions, translation, and expression lead to disease and cancer progression (*13*). Moreover, after G protein-coupled receptors, kinases are the second most drugged protein family (*11*). While some kinases such as BRAF, EGFR, and PI3-kinase demonstrate a remarkably high mutation rate within cancer (*14, 15*), many kinases are mutated at a much lower frequency making it difficult to access their influence on cancer progression

Here, we hypothesize that some closely related kinases may act as a single functional group from the perspective of a cancer type. That is, mutations at the same (cognate) position across a group of kinases may have a similar functional effect and fulfill the same selective pressure, leading to positional enrichment of impactful mutations within the cancer. To test this possibility, we used kinase alignments and exomic mutations from the TCGA to group all mutations occurring at the same sequence position and then quantified the predicted functional impact using Evolutionary Action (EA). We identified highly conserved, functionally related positions with a significantly increased mutation rate in a pan-cancer and pan-kinase analysis. Additionally, mutational differences are clear between the various kinase subclasses and additional differences across cancer types. This work shows a novel method that moves beyond a single gene approach and which suggests that functionally related homologous proteins may bear driver mutations that substitute for each other to support cancer progression.

2. Methods

2.1. *Evolutionary Trace and Action Analysis*

To identify evolutionarily important residues, we performed Evolutionary Trace (ET) analysis on each of the kinase sub-families as previously described (*16*). ET utilizes changes in genotype and corresponding phenotypic divergences in the phylogenetic tree to score the evolutionary importance of each residue in a protein sequence. In previous work, ET has identified functional sites and their determinants so as to guide mutational engineering in case studies (*17, 18*).

Evolutionary Action (*19*) builds upon ET to predict the impact a mutation has on protein function by multiplying the importance of the position (ET) by the magnitude of the substitution (evolutionary substitution odds). Prediction scores are then normalized for each individual kinase so the range falls between a predicted effect that is null, 0, to one that is most impactful, 100. EA has been repeatedly validated. It was shown to correctly predict mutation impact in multiple systems (e.g. P53, RecA, bacteriophage T4 lysozyme, etc.), it also outcompeted state of the art methods in the past 3 CAGI challenges (Critical Assessment of Genome Interpretation) (19), and in a clinical context, it can stratify patients with head and neck cancer based on their p53 mutational status (28). Using this technique we score each mutations predicted impact.

2.2. *Kinase Alignment, Mutation Acquisition and Mapping*

In order to compare mutations across all human kinases, we aligned separately each of the 7 major subclasses from The Human Kinome project (*20*). These alignments were used as a translation tool, in order to map mutations across human kinases onto canonical protein sequences. Representative crystallized structures were selected for each sub-family to visualize analysis. Representative proteins can be found in the supplement and were manually chosen based on: 1) the availability of a high resolution crystal structure 2) their similarity to other proteins within that class and finally 3) with a focus on longer proteins so as to limit the number of blank alignment positions when mapping other proteins onto the structure.

Mutation data was acquired from the TCGA for 21 major cancer types using the computationally annotated calls. Chromosome positions were converted to protein position using ANNOVAR (*21*) and then were each mapped onto the representative sequence within the alignments. In this way we were able to measure how mutations within kinases distribute throughout the conserved kinase domain.

Unless otherwise stated, all mutation numbering is in relation to the representative structure from TKL kinases (ACTR2B-2QLU). For visualization purposes on the structure, sphere size of each position was scaled based on frequency of high impact mutations (EA>40) according to the equation:

$$\text{Sphere Size}=2*(\text{Frequency}/\text{Maximal Frequency}) \tag{1}$$

Initially this analysis was performed on each of the seven kinase subclasses (358 individual kinases total) using separate alignments and representative structures for each subclass. CK1 kinases were dropped from the analysis due to insufficient mutations. The remaining six individual representative structures were then aligned and merged into a complete pan-cancer analysis.

2.3. *Random Controls*

See Supplement for additional Methods at http://mammoth.bcm.tmc.edu/GallionEtAlPSB/

3. Results

3.1. *Evolutionary Trace Identifies Functionally Important and Divergent Kinase Positions*

In order to gauge the impact of kinase mutations we first sought to identify key functional residues and sites in kinases. This was done using Evolutionary Trace (ET). Figure 1 shows the ET ranks from most to least important (red to blue) mapped onto the structure of ACTR2B, 2QLU (PDB-ID). As expected, functionally essential motifs, such as the magnesium binding DFG motif and the catalytic HRD motif emerge as ET hotspots. ET also suggests functionally relevant residues

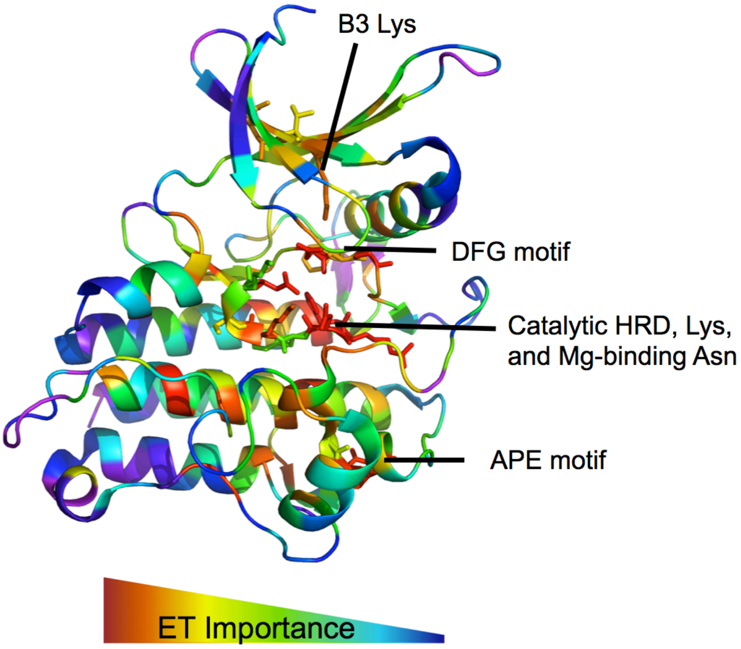


Fig 1: Evolutionary Trace Analysis of ACTR2B (2QLU) identifies evolutionarily important residues corresponding to known motifs.

throughout the substrate pocket and allosteric sites consistent with known protein functionality. Positions predicted to be the least important tend to cluster near the edges of helices, the loop regions, and near solvent exposed positions. Repeating ET analysis on each individual class, we are able to identify positions important to each group. These results confirm that in kinases, ET is able to identify both universally important positions as well as the positions that are evolutionarily divergent among subfamilies correlating to divergent functionalities.

3.2. *Kinase Mutations Demonstrate Non-Random Structural Pattern In TCGA*

To explore structural biases of kinase mutations in cancer, we next conducted a pan-cancer analysis of TCGA data. This analysis grouped mutations occurring at the same sequence position across kinase evolutionary history. This broad pan-cancer analysis identifies 77 residues with a statistically significant mutation rate (p-value<0.01) compared to control (See Supplementary). Then, in order to focus on the subset of impactful mutations and screen out low impact polymorphisms, we repeated the above analysis only using mutations with EA scores greater than 40, and mapped them onto the ET analysis of ACTR2B (Figure 2A). All positions are numbered based on the 2QLU structure unless otherwise specified. For example, the well-known driver mutations from BRAF-V600 (equivalent position V344 in figure) and CHEK2-K373 (R345 in figure) are the most frequently mutated, high impact mutations. Other frequently mutated positions with high impact substitutions occur at known functional residues, such as the glycine-rich region G199, the DFG motif D339, the HRD domain R320 and D321, and a conserved ion-pairing residue R468. Since these mutations involve positions with large ET scores, they are likely to impair protein function. By contrast, and as seen in Figure 2B, there are 54 residues mutated at a lower rate than expected (p-value<0.01). These seldom mutated positions, shown by the small

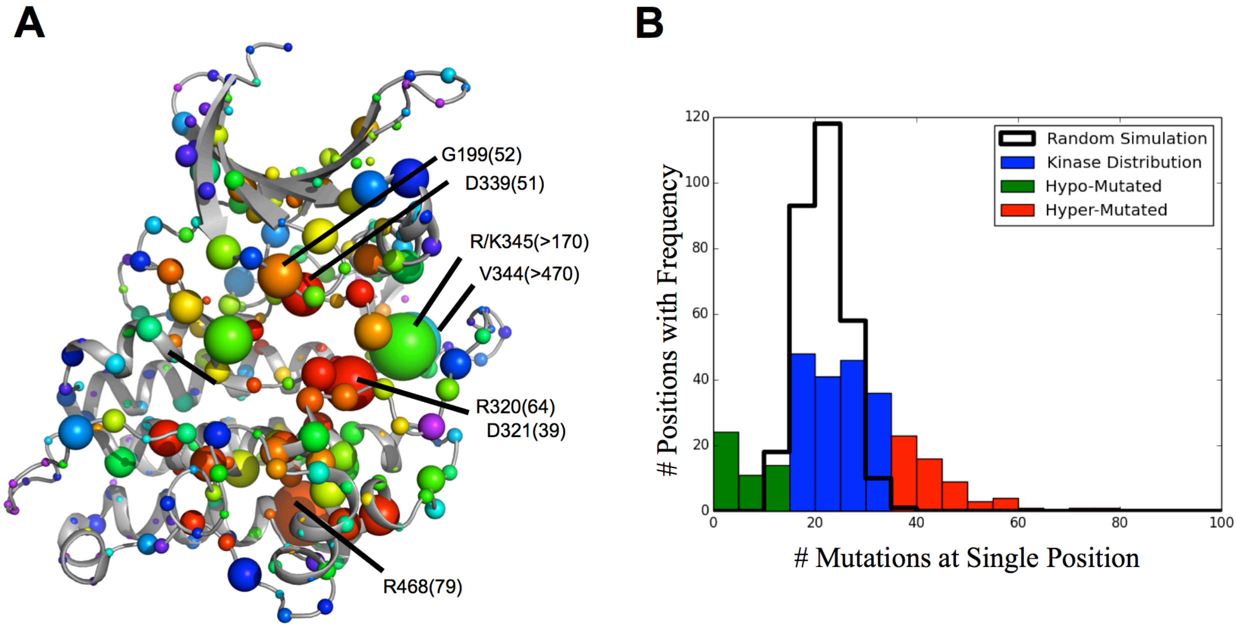


Fig 2: Kinase Mutation Pattern Pancancer (A) Pan-cancer mutations, with an EA>40, mapped onto ACTR2B structure where sphere size=frequency, color=ET importance. (B) Actual mutation frequency significantly varies from Poisson Distribution.

sphere size in Figure 2A fall preferentially in the solvent exposed loop regions of the kinase that are evolutionarily less important according to ET and thus unlikely to have much functional consequence. These data show that kinase mutations in cancer are not evenly distributed throughout the structure. Rather many mutations preferentially fall non-randomly so as to recurrently involve functionally important cognate positions within conserved motifs, where they are likely to be disruptive; conversely, in the loop regions, which are less important, mutations are more rare and involve positions of lesser importance.

3.3. *Frequently Mutated Positions are Enriched for Mutations Predicted to Have a Significant Impact on Protein Function*

To further explore the functional consequences of these mutations, we used EA to predict the functional impact of each mutation on protein function. EA combines the evolutionary importance of the position (ET) with the likelihood of that substitution, based on all evolutionary history, in order to predict the impact of a mutation on protein function. We compared the EA score distribution of frequent positions and infrequent positions (*p*-value<0.01) to the distribution of all kinase domain mutations from the TCGA using a two-sided t-test (Figure 3A). In agreement with the structural and ET biases, the frequently mutated positions are predicted to have a higher impact on protein function (*p*-value=10^{-28}) while the infrequently mutated positions are biased towards lower impact mutations (*p*-value= 10^{-5}). These data show the frequently mutated positions from the TCGA are further enriched for high impact mutations, while those positions infrequently mutated are predicted to have little functional effect.

3.4. *Frequently Mutated Positions Occur in Many Different Kinases at a Low Individual Frequency*

While these cancer somatic mutations demonstrate site specificity, we next investigated which individual kinases carried these mutations and whether specific proteins drove this pattern. The mutation frequency of each individual kinase is displayed in Figure 3B and is compared against a random simulation in which the same number of mutations were randomly distributed to an equal number of proteins. The random distribution had a mean value of 21.4 mutations per kinase while the experimental distribution, after dropping out the outliers BRAF and CHEK2 (550 and 160 mutations, respectively), had a mean of 19.5. We note that the mutation rate in individual kinases is more variable than expected. Overall the distribution is leftward shifted compared to control with a select number of proteins hypermutated: 29% of kinases were mutated at a decreased frequency (*p*-value<0.05) while only 14% of kinases were significantly hypermutated (*p*-value<0.05). Of the hypermutated kinases, nine were mutated at an exceptionally high rate (>50 mutations/protein); many of these however, represent known, high frequency driver mutations occurring at the same location in the same kinase (e.g. BRAF, CHEK2, and EGFR). These data show that within cancer cells, certain kinases experience a remarkably increased mutation rate while the majority of the remaining kinases are hypomutated, typically with fewer that 20 SNVs across a pan-cancer analysis.

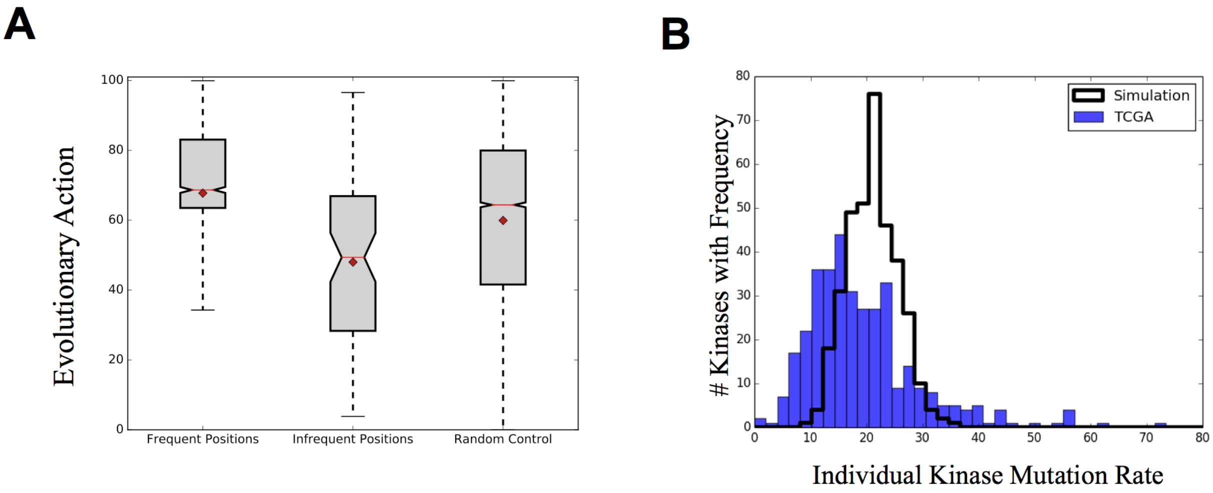


Fig 3: (A) High frequency mutations are significantly biased towards high impact mutations (Pvalue=5*10^{-28}) while low frequency mutations are biased towards low predicted impact (Pvalue 2.5*10^{-05}). Mean=diamond Median=red line Whisk=2STD (B) Observed kinase mutation rate compared to computer simulation of random mutations. BRAF and CHEK2 (550 and 160 mutations, respectively) not shown on plot.

However, while this analysis recapitulates known drivers such as L858R within EGFR, it further identifies mutations at a single residue that individually occur at a low frequency but, taken as a whole, occur at a high frequency. For instance, Table 1 displays a random selection of mutations occurring at the Asp residue of the HRD domain (*p*-value=2x10^{-4}). While each individual mutation has a conserved amino acid transition, individual proteins are mutated infrequently with a median value of 1 and a maximal value of 5 mutations (occurring within MAP2K7). Of the original 54 positions with a *p*-value<0.01 only 6 are at least partially driven by a single protein (1 protein with >20% of the mutations), while all remaining positions were significant only through this combination. These data show that while individual mutations may occur at low frequency, they frequently occur at homologous structural positions with the same native residue and amino acid substitution. Furthermore this pattern is distributed across many individual kinases without a single driver protein.

Table 1. Random sample of mutations occurring at catalytic Asp residue from the HRD domain.

Protein	Substitution	EA Score	Kinase Class	Cancer Type
PRKCI	D378N	64.36	AGC	PAAD
PRKG2	D576Y	98.63	AGC	READ
CHEK1	D130Y	98.73	CAMK	LUSC
STK17B	D158N	74.06	CAMK	SKCM
CLK4	D286N	63.19	CMGC	COAD
MAPK4	D149G	90.95	CMGC	STAD
MAP2K3	D190N	71.77	STE	SKCM
MAP2K7	D243N	61.93	STE	COAD
MAP2K7	D243N	61.93	STE	PAAD
PAK7	D568N	75.15	STE	SKCM
EPHA3	D746N	43.15	TK	SKCM
FES	D683E	67.14	TK	BRCA
ROR2	D615N	74.82	TK	SKCM
MAP3K7	D156Y	96.77	TKL	LIHC

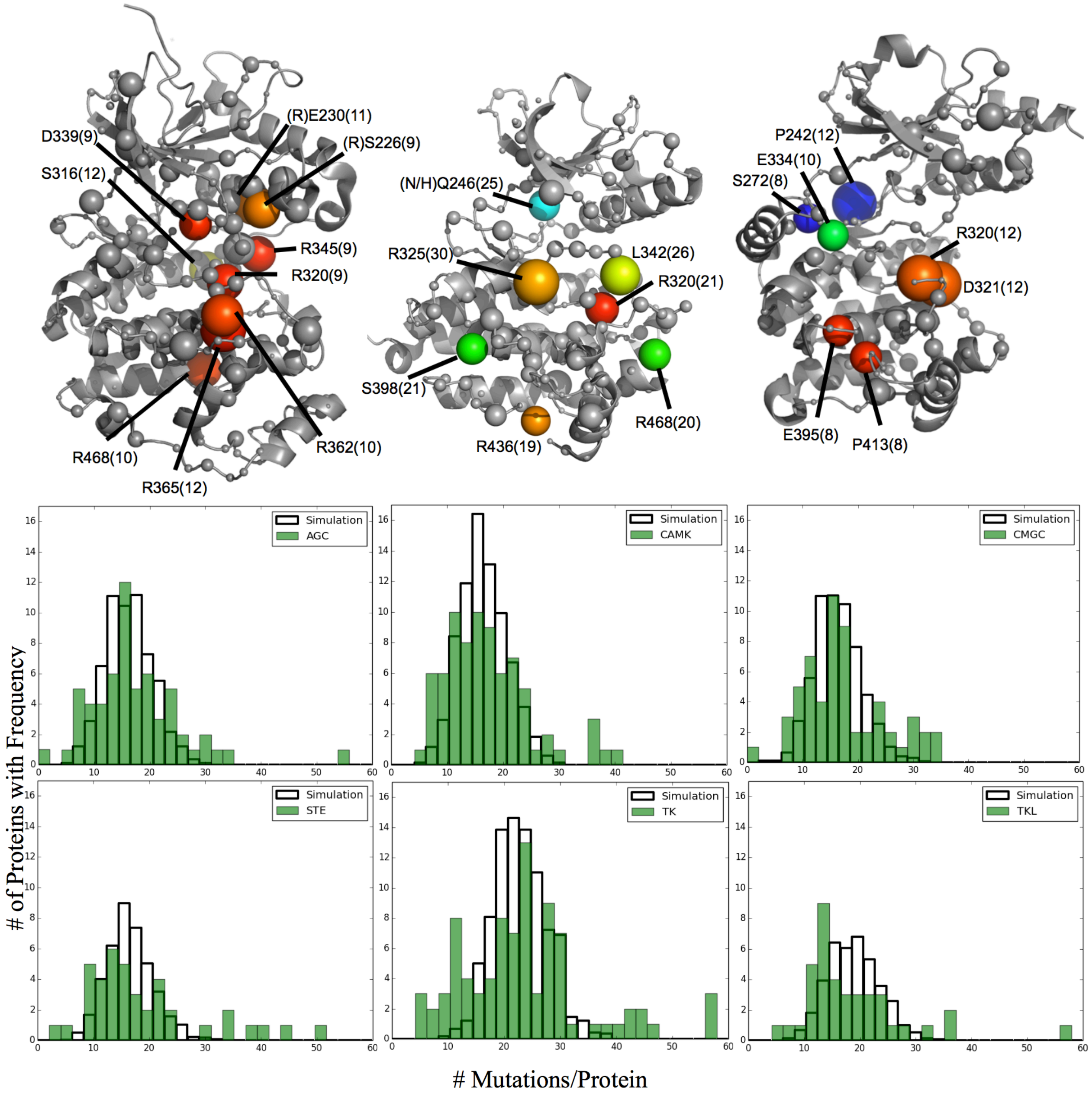


Fig 4: Individual kinase subclasses are frequently mutated at distinct positions (Left to Right: CMGC, TK, STE kinases). Sphere Size=Frequency, Color= ET importance from high to low (red to blue) for each representative kinase: ERK1, EphA5, PAK1 (respectively). All labels are based on ACTR2B numbering. (B) The protein mutation rate for each kinase class was compared against a simulated random distribution specific to the total number of mutations and proteins in each class. BRAF (TKL) and CHEK2 (CAMK) are not shown on their respective figures

3.5. *Individual Kinase Classes Show Unique Mutational Patterns*

Individual subclasses of kinases display marked functional and structural differences corresponding to their target specialization (12). To test if our conclusions held true despite these differences, we repeated the above analysis for each kinase class. As an example, three of these classes are displayed in Figure 4A. While the general location of these residues tend to stay near the catalytic site, the frequently mutated positions from each class vary. Nine residues in CMGC kinases form a statistically significant cluster (*z*-score=4.58) roughly localized around and occurring within the HRD domain. Seven residues in TK kinases are more broadly distributed throughout the structure with the three most frequent near the HRD domain. Finally, STE kinases seem to show two distinct areas of mutation, the HRD region and the ATP-binding hinge region. In all three cases, similar to the pan-cancer analysis, the most frequent positions tend to occur at evolutionarily important residues in functional motifs with high impact mutations. In addition to these differences, significant positions from the pan-kinase analysis are still significant in multiple classes (e.g. R320 (HRD motif) and R468 (Ion pair)). These data indicate that within cancer, certain positions are preferentially enriched in select kinase subclasses while other positions demonstrate broad enrichment across many or all kinase types.

We further note differences in the mutation frequency of proteins from each of the kinase classes (Figure 4B). In each class, some proteins are mutated at a significantly higher rate than expected. Proteins from the AGC kinase class are normally distributed with an exaggerated variance compared to random simulation, indicating that mutations within this class are fairly distributed to many proteins. Likewise, the mutation rate in CMGC and TK kinases is even more varied but still follow a roughly normal distribution centered around the expected mean. The distribution from CAMK, STE, and TKL kinases match the pan-kinase analysis with a leftward shifted distribution displaying many hypomutated proteins and several hypermutated proteins. As 43% of all mutations within TKL kinases occur in BRAF, we have removed these mutations from this analysis. However, creating a random distribution for this class without first removing this outlier shifts the random distribution right (mean=30). This data shows that, in addition to structural differences, the individual kinase classes are mutated at different rates, with some classes having broadly distributed mutations to many individual proteins while other classes are primarily mutated in a select few proteins.

3.6. *Kinases Further Demonstrate Cancer Type Specific Mutational Patterns*

Variances between kinase subtypes led us to next speculate that certain protein positions could have varying functional importance to specific cancer types as well. The above analysis was repeated, now grouping all kinases together and instead performing a cancer-specific analysis for 7 cancer types within the TCGA [Breast invasive carcinoma (BRCA), Bladder Urothelial Carcinoma (BLCA), Colon adenocarcinoma (COAD), Head and Neck squamous cell carcinoma (HNSC), Lung adenocarcinoma (LUAD), Skin Cutaneous Melanoma (SKCM), and Stomach adenocarcinoma (STAD)]. The most frequently mutated position for all but BRCA and STAD was 345 and 346, driven by the high frequency driver mutations BRAF-V600 and CHEK2-K373 (respectively); these mutations were then removed from this analysis in order to search for novel other positions. Figure 5 shows a selection of positions that were significantly mutated within specific cancer types. Interestingly, the analyses from LUAD and STAD resulted in clusters of mutations within the kinase domain. Some positions were significant in two cancer types, such as

L325 in LUAD and BRCA. In agreement with the pan-cancer analysis, R468 was frequently mutated in many cancer types including STAD and COAD. These data indicate that individual cancer types are enriched for varying structural positions across many individual kinases.

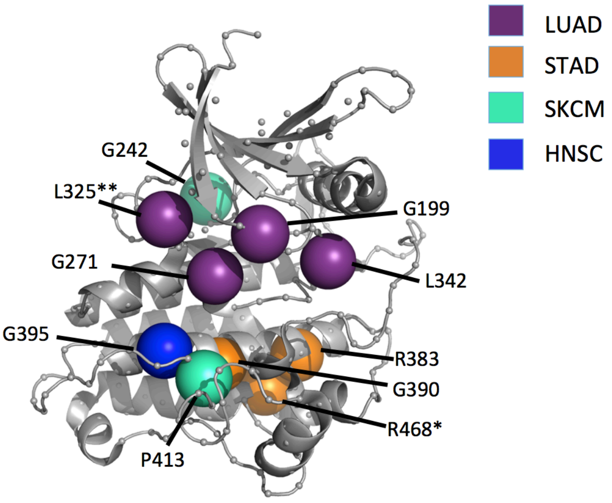


Fig 5: Cancer types demonstrate some specificity towards certain mutation positions. *Occurs in STAD and COAD **Occurs in both LUAD and BRCA

4. Discussion

In order to better predict driver mutations within cancer, computational methods have been extended from gene-by-gene analyses to consider instead groupings of mutations in functional pathways or subnetworks (*3, 6, 22*). In this manner, driver proteins mutated at a low frequency due to the heterogeneity within cancer that are missed by a single gene analysis can still be identified despite their low individual frequency. Being able to predict these diverse infrequent drivers of cancer helps move medicine closer to personalized diagnoses and care. Here, as an alternate way to group genes, we explored protein homology rather than curated hierarchical pathways and gene interactions. Strikingly, we find that among kinases, mutations are structurally biased to functional motifs and evolutionarily important residues.

Mutations providing a benefit to cancer cells become clonally enriched, as that cell proliferates more efficiently than others in the tumor population (*5*). From the pan-kinase analysis, we identified positions frequently mutated across many individual kinases. While the known high frequency driver genes were captured in this analysis, an additional 39 positions were mutated at a low frequency in any given kinase but were significantly mutated across the kinase family. These high frequency positions were preferentially biased for high impact mutations, strongly suggesting a significant effect on protein function. In contrast, the infrequently mutated positions all occurred at evolutionarily unimportant loop regions with a bias towards low impact mutations. These data indicate that enrichment is correlated to functional impact. Presumably, the high-impact mutations across many kinases provide a functional benefit within the cancer cell and are therefore enriched, whereas low-impact mutations, providing little benefit to the cancer cell, are lost from the population resulting in a low mutation rate at those positions.

Previous work in kinases has demonstrated that identical mutations in two different kinases can result in the same phenotype (*23, 24*). For instance, mutations conferring resistance to kinase inhibitors in EGFR occur at the same position as drug resistance mutations in BCR-ABL, PDGFRA and KIT (*25, 26*). A systematic study of mutation locations built upon these observations and demonstrated the existence of 'domain hotspots': frequently mutated regions in many proteins leading to the same functional consequence (*22*). In the context of this analysis of exomic mutations from TCGA, these frequently mutated positions, across many different kinases, with the exact same substitution strongly suggests a conserved functional mechanism driving enrichment: the same mutation in two different kinases likely producing a similar benefit in cancer.

The kinase catalytic mechanism itself is highly conserved across all kinases and is orchestrated by groups of functional motifs; these same positions in all kinases are responsible for the same functions (*12*). These motifs are themselves frequently mutated at a high frequency within cancer; in fact, the majority of the frequently mutated positions occur at or nearby conserved motifs. These positions are often studied in the context of kinases enabling us to speculate on their functional consequences within cancer. For instance, the catalytic Asp and Arg residues of the HRD domain are both mutated in many diverse kinases and are furthermore mutated to the same types of residues (D→N and R→Q/W/H respectively) in each case. Previous characterization of the D→N mutations within the Drosophila Src64 kinase indicate this mutation is equivalent to a gene knockout (*27*). Cancer cells carrying this SNV would therefore experience a loss of kinase activity within this protein, possibly suggesting a tumor suppressing mechanism. What remains to be determined is how far these characterization studies can be extrapolated to other kinases. Further experimental studies are needed in which the same mutation is characterized in multiple proteins to assess how universal these conclusions are. However, given 1) the initial conserved function of these positions, 2) the significant enrichment of the same substitution across many proteins, and 3) the sizeable predicted consequence of these mutations: it becomes tantalizing to suggest that the same mutation in different kinases may produce the same functional benefit in cancer, regardless of the kinase where it occurs.

This hypothesis is further supported by the kinase subclass and cancer type specific analyses. The individual kinase sub-families are evolved to phosphorylate different types of proteins (12). As a result, they have diverged. While the overall structure is conserved, some positions are specific for given target proteins and therefore differ among kinase classes. Likewise, while some positions are broadly mutated, the class specific analyses demonstrate appreciable differences; some positions are enriched in one class but do not occur in another. These variations between kinase classes likely stem from their functional divergence. Mutations occurring at important positions in one class may be beneficial, while the same position in a different class may not, resulting in differential enrichment. Furthermore, cancer types themselves display heterogeneity among their causal driver mutations (*5*), a heterogeneity reflected within kinase mutations as well. Different cancer types are enriched for different kinase positions, again suggesting that some positions may be preferentially beneficial for one cancer type more so than another, and therefore clonally enriched. When the selection pressure varies, either by differing cancer types or by the different kinase classes, the positions of the enriched mutations also vary. This further suggests that a conserved functional mechanism drives this mutational enrichment across many individual kinases.

Cellular homeostasis and function is often maintained by a complex network of proteins with significant functional overlap and crosstalk between functional homologues. For this reason, a single gene approach to predicting driver mutations in cancer may be overly simplistic, therefore requiring a methodology to combine mutations based on functional similarity. Here, we propose that in addition to curated pathways, mutations can also be grouped across homologous protein families. Within kinases, we have demonstrated that individual proteins are enriched for mutations occurring at cognate positions utilizing the same substitutions. These results suggest that the selection pressure within certain cancers may be specific to the mutation's location and not differentiate between which kinase carries the mutation. Taken together, these data show individual kinases may behave in a functionally redundant manner in cancer and that a combined analysis of their mutations could identify individually infrequent driver mutations, previously missed, that occur frequently across the entire class. The conserved nature of these mutations allows speculation as to their predicted functional effect by extrapolating previous characterization studies, in a single protein, to the other kinases. Finally, while these results are specific to kinases, similar analyses could be broadly applicable across many protein families, thereby shifting focus from a 'protein specific' to a 'paralog-wide, cognate position specific' analysis of cancer driver mutations.

References

1. J. S. Kaminker *et al.*, *Cancer Res* **67**, 465-473 (2007).
2. P. Katsonis *et al.*, *Protein Sci* **23**, 1650-1666 (2014).
3. F. Vandin, P. Clay, E. Upfal, B. J. Raphael, *Pac Symp Biocomput*, 55-66 (2012).
4. B. Vogelstein, K. W. Kinzler. *Nat Med* **10**, 789-799 (2004).
5. D. Hanahan, R. A. Weinberg. *Cell* **144**, 646-674 (2011).
6. P. Jia, Z. Zhao. *PLoS Comput Biol* **10**, e1003460 (2014).
7. M. Kanehisa, S. Goto. *Nucleic Acids Res* **28**, 27-30 (2000).
8. D. Croft *et al. Nucleic Acids Res* **42**, D472-477 (2014).
9. D. Szklarczyk *et al.*, *Nucleic Acids Res* **39**, D561-568 (2011).
10. N. L. Nehrt, T. A. Peterson, D. Park, M. G. Kann. *BMC Genomics* **13 Suppl 4**, S9 (2012).
11. A. Torkamani, N. J. Schork. *Genomics* **90**, 49-58 (2007).
12. J. A. Endicott, M. E. Noble, L. N. Johnson. *Annu Rev Biochem* **81**, 587-613 (2012).
13. C. Greenman *et al. Nature* **446**, 153-158 (2007).
14. H. Davies *et al. Nature* **417**, 949-954 (2002).
15. E. Lengyel, K. Sawada, R. Salgia. *Curr Mol Med* **7**, 77-84 (2007).
16. A. Wilkins, S. Erdin, R. Lua, O. Lichtarge, *Methods Mol Biol* **819**, 29-42 (2012).
17. H. J. Kang, A. D. Wilkins, O. Lichtarge, T. G. Wensel. *J Biol Chem* **290**, 2870-2878 (2015).
18. S. M. Peterson *et al. Proc Natl Acad Sci U S A* **112**, 7097-7102 (2015).
19. P. Katsonis, O. Lichtarge. *Genome Res* **24(12)**, 2050-2058 (2014).
20. G. Manning, D. B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam. *Science* **298**, 1912-1934 (2002).
21. K. Wang, M. Li, H. Hakonarson. *Nucleic Acids Res* **38**, e164 (2010).
22. P. Yue *et al. Hum Mutat* **31**, 264-271 (2010).
23. J. L. Marks *et al. PLoS One* **2**, e426 (2007).
24. H. Davies *et al. Cancer Res* **65**, 7591-7595 (2005).
25. W. Pao *et al. PLoS Med* **2**, e73 (2005).
26. S. Kobayashi *et al. N Engl J Med* **352**, 786-792 (2005).
27. T. C. Strong, G. Kaur, J. H. Thomas. *PLoS One* **6**, e28100 (2011).
28. Neskey *et al.* Cancer Research 75,(7); 1527-36 (2015).

MUSE: A MULTI-LOCUS SAMPLING-BASED EPISTASIS ALGORITHM FOR QUANTITATIVE GENETIC TRAIT PREDICTION

DAN HE and LAXMI PARIDA

IBM T.J Watson Research
Yorktown Heights, NY
E-mail: {dhe, parida}@us.ibm.com

Quantitative genetic trait prediction based on high-density genotyping arrays plays an important role for plant and animal breeding, as well as genetic epidemiology such as complex diseases. The prediction can be very helpful to develop breeding strategies and is crucial to translate the findings in genetics to precision medicine. Epistasis, the phenomena where the SNPs interact with each other, has been studied extensively in Genome Wide Association Studies (GWAS) but received relatively less attention for quantitative genetic trait prediction. As the number of possible interactions is generally extremely large, even pairwise interactions is very challenging. To our knowledge, there is no solid solution yet to utilize epistasis to improve genetic trait prediction. In this work, we studied the multi-locus epistasis problem where the interactions with more than two SNPs are considered. We developed an efficient algorithm MUSE to improve the genetic trait prediction with the help of multi-locus epistasis. MUSE is sampling-based and we proposed a few different sampling strategies. Our experiments on real data showed that MUSE is not only efficient but also effective to improve the genetic trait prediction. MUSE also achieved very significant improvements on a real plant data set as well as a real human data set.

Keywords: Genetic Trait Prediction, Mutual Information, Epistasis, Weighted Maximum Independent Set

1. Introduction

Given its relevance in the fields of plant and animal breeding as well as genetic epidemiology,[1–3] whole genome prediction of complex phenotypic traits using high-density genotyping arrays recently received great attentions. Complex traits prediction and association are crucial to translate the findings in genetics to precision medicine. Given the genotype values encoded as $\{0, 1, 2\}$ of a set of biallelic molecular markers (we use “feature”, “marker”, “genotype” interchangeably), such as Single Nucleotide Polymorphisms (SNPs), on a collection of plant, animal or human samples, quantitative genetic traits, such as weight, height, fruit size etc. of these samples can be predicted effectively. More accurate genetic trait prediction can help breeding companies to develop more effective breeding strategies.

One of the most popular algorithms for the genetic trait prediction problem is *rrBLUP* (Ridge-Regression BLUP),[1,4] which assumes all the markers contribute to the trait value more or less. The algorithm fits an additive linear regression model where all the markers are invovled. It fits the coefficient computed for each marker, which quantifies the importance of the marker. The rrBLUP method has the benefits of the underlying hypothesis of normal distribution of the trait value and the marker effects (well suited for highly polygenic traits). Its performance is as good as or better than other popular predictive models such as Elastic-Net, Lasso, Ridge Regression,[5,6] Bayes A, Bayes B,[1] Bayes Cπ,[7] and Bayesian Lasso,[8,9] as well as other machine learning methods.

Epistasis is the phenomenon where different markers, or genes, can interact with each other. The problem of epistasis detection has been widely studied in GWAS (Genome Wide Association Studies). Lots of work, mainly greedy strategies,[10–16] have been proposed to detect epistasis effects. These greedy strategies all assume that significant epistasis effects come from only strong marginal effects, or the markers that are highly relevant to the trait. While most existing methods target epistasis detection on GWAS, some recent developments have been achieved on quantitative genetic trait prediction. He et al.[17] proposed a sampling-based method MINED to detect significant pairwise epistasis effects and to improve the genetic trait prediction. He and Parida[18] further proposed a two-stage sampling algorithm SAME to handle multi-locus epistasis effects where the number of markers involved can be greater than two. They showed that the prediction can be significantly improved with the help of epistasis. In the meanwhile SAME has a few advantages over the existing methods: It is highly scalable; It captures epistasis effects from both strong and weak marginal effects. However, SAME still has a few drawbacks: Its sampling strategy is based on random sampling where for all interactions the same number of samplings is conducted; It does not check the redundancy of the sampled interactions thus many sampled interactions might be redundant given the huge sample space; Its interaction values are based on multiplications of the genotype values, which does not distinguish all the possible genotype combinations.

In this work, we studied the multi-locus epistasis problem where the interactions with more than two SNPs are considered. We developed an efficient algorithm MUSE (Multi-locus Sampling-based Epistasis algorithm) to improve the genetic trait prediction with the help of multi-locus epistasis. MUSE conducts bidirectional sampling: It samples k-locus interactions from (k-1)-locus interactions and it decomposes the k-locus interactions into multiple (k-1)-locus interactions for further sampling. The motivation comes from the observation made in[17] that when a (k-1)-locus interaction is involved in a significant k-locus interaction, no matter whether it is a strong marginal effect or not, it is likely to be involved in multiple significant k-locus interaction. The main contribution of this work is a set of sampling strategies, including constraint-based sampling, encoding-based sampling and iterative sampling. More details will be given in the method section. Our experiments showed that MUSE is not only efficient but also effective to improve the genetic trait prediction. We also observed significant improvements on a real plant data set as well as a real human data set over the state-of-the-art methods.

2. Preliminaries

Genetic trait prediction problem is usually represented as the following linear regression model:

$$Y = \beta_0 + \sum_{i=1}^{d} \beta_i X_i + e$$

where Y is the phenotype and X_i is the i-th genotype value, d is the total number of genotypes and β_i is the regression coefficient for the i-th marker, e is the error term which usually follows a normal distribution. We call the above model *single marker model.*

Epistasis is the phenomenon where different markers can interact with each other. With the pairwise epistasis effects, the traditional linear regression model becomes the following non-linear additive model:

$$Y = \beta_0 + \sum_{i=1}^{d} \beta_i X_i + \sum_{i,j}^{d} \alpha_{i,j} X_i X_j + e \tag{1}$$

where $X_i X_j$ is the product of the genotype values of the i-th and j-th marker and it denotes the interaction of the two genotypes.

Multi-locus epistasis model is more complicated as more than two markers are involved in the interactions. When n-markers are involved in the interaction, we call it *n-locus* interaction or *n-way* interaction, which are interchangeable and we call n as the *order* of the interaction. The model is shown as below:

$$Y = \beta_0 + \sum_{i=1}^{d} \beta_i X_i + \sum_{i,j}^{d} \alpha_{i,j} X_i X_j + \cdots + \sum_{i_1, i_2, \ldots, i_n}^{d} \alpha_{i_1, i_2, \ldots, i_n} X_{i_1} X_{i_2} \ldots X_{i_n} + e \tag{2}$$

For example, the regression model involving both 2-locus and 3-locus interactions is:

$$Y = \beta_0 + \sum_{i=1}^{d} \beta_i X_i + \sum_{i,j}^{d} \alpha_{i,j} X_i X_j + \sum_{i,j,k}^{d} \alpha_{i,j,k} X_i X_j X_k + e$$

3. Multi-locus Sampling-based Epistasis Algorithm

In this work, we follow the pipeline of SAME[18] to conduct the bi-directional search. We start sampling in a forward manner from the significant (k-1)-locus interactions to obtain the significant k-locus interactions. Then we search in backwards where we take the significant k-locus interactions to guide what extra (k-1)-locus interactions we should consider to sample. This is based on the observations made in the work of He et al.[17] that if a (k-1)-locus interaction is involved in a significant k-locus interaction, no matter whether this (k-1)-locus interaction is significant or not, it is likely to be involved in multiple significant k-locus interactions.

We first use a queue Q to store the features (can be 1-locus to (k-1)-locus interactions) from which the sampling is conducted. We define *sampling* a t-locus effect as that for the t-locus effect, we randomly sample a set of single markers to be combined with the t-locus effect to obtain (t+1)-locus effects. We define a feature is *significant* if its r^2 (The square of the Pearson's correlation coefficient between the feature vector and the trait vector) to the trait is higher than a threshold s (We will show how to determine the threshold later). We use r^2 here as it is the most popular metric for genetic trait prediction (or genomic selection). We start from significant single markers and store all of them in Q. Then we sample each single marker X to obtain a set of significant 2-locus interactions where the marker X is involved in. If the 2-locus interaction is significant, we store it in Q. Then for the significant 2-locus interaction, we decompose it into two 1-locus effects, or two single markers. One of the markers will be X, the other one is either a strong or weak marginal effect. If the other marker is not in Q yet, we store it in Q so that it will be sampled later on.

We then repeat the sampling process for 2-locus interactions upto (k-1)-locus interactions. When we sample a (k-1)-locus interaction, if we obtain a significant k-locus interaction, we then decompose the k-locus interaction into k (k-1)-locus interactions and store them in Q. For example, given a significant 2-locus interaction AB, we randomly sample one single marker and by chance we obtain a significant 3-locus interaction ABF. Then we decompose it into three 2-locus interactions AB, BF, AF and store them in Q if they have not been stored yet. They will be sampled in a later stage.

3.1. *Significance Threshold*

The significance threshold s is determined dynamically. This is because we only keep the top K most significant features and thus the threshold is set naturally as the r^2 of the top K-th feature. We maintain a sorted list of the features according to their r^2 score (notice we consider both epistasis effects and single marker effects). When we check an interaction, we insert the interaction into the top-K feature set if its r^2 score is better than s and we remove the last feature from the list. If the interaction does not have a higher r^2 score than s, we do not change the list. We then set the threshold s as the r^2 score of the current K-th feature. We keep on updating the threshold as we insert more interactions, while keeping the order of the list according to the r^2 scores. As the threshold becomes higher, it becomes harder for an interaction to be selected.

3.2. *P-value*

As the feature space is extremely large, in order to avoid over-fitting problem, we also computed the p-value of the features. We ignore features with high r^2 score if the p-value of the features are not small enough. Similar to GWAS, where a typical p-value threshold is 5×10^{-8} after Bonferroni corrections for multiple testing, we used very small p-values. We observed that we can not use a fixed p-value. Instead, for larger feature space, we need to use smaller p-values. For example, for a feature space of size $O(10^7)$, we use p-value 5×10^{-6}. For a feature space of size $O(10^{10})$, we use p-value as 5×10^{-8} to 5×10^{-11}. The p-value to be used is determined by a grid search using cross validation.

3.3. *Estimate Interaction Probability*

Another thing to notice is that when we conduct the sampling, we do not sample all the single markers as it would be very time consuming for a large number of markers. We conduct an initial sampling with size f. It is shown in[17] that the scores follow a truncated normal distribution. Then using the f sampled r^2 scores, we can fit the truncated normal distribution to estimate the mean and the standard deviation. Using this distribution, and given the total number of single markers as d, we compute the probability of seeing at least one significant r^2 score out of the $O(d)$ possible interactions, where a score is significant if it is higher than the current significance threshold s. If the probability is higher than a threshold P, we will test the interactions between the marker and all the remaining markers. In order to capture as many epistasis interactions as possible, we generally use a small value for P, say 0.005.

As we can see, the performance of MUSE is heavily dependent on the sampling strategy. In SAME,[18] a simple random sampling is conducted which has been shown to have certain disadvantages. Next we introduce three sampling strategies that could significantly improve the random sampling:

3.4. *Sampling Strategies*

3.4.1. *Constraint-based Sampling*

Significant interaction selection can be considered as a feature selection process if we consider each significant interaction as a feature. A popular feature selection criteria is called MRMR (Maximum Relevance and Minimum Redundancy),[19] where the objective is to select a set of features which are maximumly relevant to the trait but minimally redundant with each other. It is shown[19] that minimizing the redundancy of the selected features leads to better prediction. In our approach, the selection of the top-k most significant interactions is equivalent to maximizing the relevance of the selected interactions to the trait. However, the redundancy of the selected interactions is not taken into consideration yet.

It is observed in[17] that a t-locus interaction might be involved in multiple significant (t+1)-locus interactions. However, these multiple significant (t+1)-locus interactions might be highly redundant with each other, as all of them share the same t-locus interaction. As the size k is fixed for the top-k most significant interactions, including many redundant interactions might not improve the prediction according to the MRMR criterion. An extreme case is that all the top-k most significant interactions are redundant, which is equivalent to using only one interaction for prediction. This will obviously lead to poor performance.

Thus here we add a constraint on the sampling process: we require every t-locus interaction involved in at most N (t+1)-locus interactions. We call N the *overlap threshold*. Therefore, any of the top-k interactions should at most overlap with N other top-k interactions, where overlap means two (t+1)-locus interactions share the same t-locus interaction. We call this sampling *Constraint-based Sampling*.

To solve the constraint-based sampling problem, we construct an *Interaction Graph*, where the nodes are (t+1)-locus interactions, the edges indicate that the two (t+1)-locus interactions share the same t-locus sub-interaction. Each node is associated with a weight, indicating the r^2 of the node to the trait. Notice we build a graph for each t. Once we moved from t to t+1, we build a new graph and delete the old graph. As an example, we can see in Figure 1, the interaction ABC share the sub-interaction AB with the interaction ABD. Thus the number of edges associated with a node indicates the degree of overlaps of the node and we call it *connectivity*. In this example, the node ABC has connectivity as 3, the node ABD has connectivity as 4. If we set the overlap threshold N as 1, we can only select the nodes that is connected to one other node.

The constraint-based sampling problem is then converted to the problem where we would like to select a set K of k nodes such that the total weights of the nodes is maximized and in the meanwhile the constraint is satisfied, namely in the node set K, there is no node with more than N edges connecting to the other nodes in the set. The problem is similar to a Weighted Maximum Independent Set (WMIS) problem. The WMIS problem seeks to select a set of

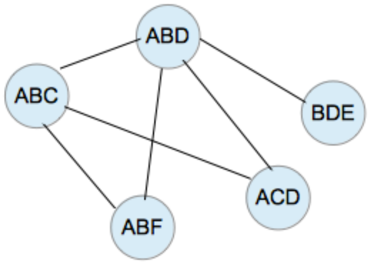


Fig. 1. An example of interaction graph.

nodes from a graph to form an independent set, where all the nodes are not adjacent, such that the sum of the weights on the nodes is maximized. As all the nodes are not adjacent in the independent set, all selected interactions are guaranteed non-overlapping. This is equivalent to allowing the degree of connectivity as 0. In our case, we set the degree of the connectivity of the selected nodes to be no greater than N.

The WMIS problem is NP-complete and what's more, it requires generating the complete interaction graph. However, in our problem, we sample the t-locus interactions one by one. Thus we conducted a greedy algorithm, where we maintain a count for every t-locus interaction. During the samplings, when we sample a t-locus interaction I and find one significant $(t+1)$-locus interaction, we increase the count of I by one. If the count is less than N, we keep on sampling. Otherwise we have two options:

(1) We stop the sampling immediately
(2) We do not stop the sampling, instead we continue the sampling process. However, we maintain only N significant $(t+1)$-locus interactions sampled from I and we call the set S. Once we identify a significant $(t+1)$-locus interaction I', we compare its r^2 score with the r^2 scores of the interactions in S. If its r^2 score is greater than the minimum r^2 score in S, we remove the interaction in S with the minimum r^2 score and replace it with I'.

Obviously, by taking option one, the sampling process can be terminated quickly but it may miss the significant $(t+1)$-locus interactions that might arrive later. By taking option two, we can guarantee that all significant $(t+1)$-locus interactions could be captured. However, we only store N significant $(t+1)$-locus interactions and thus the constraint can be satisfied. By setting N small, we could include more $(t+1)$-locus interactions that have different sub-interactions so that the redundancy of the top-k interactions can be reduced. In MUSE, we choose option two.

3.4.2. *Encoding-based Sampling*

By using the multiplication model and assuming the genotypes are encoded as $\{0, 1, 2\}$, a pairwise epistasis effect contains only 4 different possible values $\{0, 1, 2, 4\}$ (by pairwise multiplication of the values from $\{0, 1, 2\}$) while in reality there are nine different possible combinations of the alleles. It is not clear why a pair of markers with genotypes (0, 1) should have the same interaction value 0 as the pairs with genotypes (0, 2). Thus instead of using the values

$\{0, 1, 2, 4\}$, we could consider using nine different values {0, 1, 2, 3, 4, 5, 6, 7, 8} to differentiate the nine different combinations. However, there is no order for the combinations. For example, we can not determine the order of "AA/Bb" and "Aa/BB". Similarly, we can not determine the order of "Aa/bb" and "aa/Bb". Thus we do not have a systematic way to assign the nine different values {0, 1, 2, 3, 4, 5, 6, 7, 8} to the nine different combinations.

Therefore, we developed the following encoding formula:

$$encoding = \sum_{i=1}^{n} X_i \times 10^{(n-i)}$$

where n is the number of markers involved, X_i is the encoding of the genotype of the i-th marker, which is one of $\{0, 1, 2\}$. Thus instead of multiplication, we use the above encodings for the n-way epistasis interactions. For example, for pairwise interactions, assuming the encoding $\{0, 1, 2\}$ are for "AA, Aa, aa" respectively and the same for "BB, Bb, bb" respectively, we have the following encodings for the nine combinations:

$$AA/BB = 0 \times 10 + 0 = 0, \quad AA/Bb = 0 \times 10 + 1 = 1, \quad AA/bb = 0 \times 10 + 2 = 2$$
$$Aa/BB = 1 \times 10 + 0 = 10, \quad Aa/Bb = 1 \times 10 + 1 = 11, \quad Aa/bb = 1 \times 10 + 2 = 12$$
$$aa/BB = 2 \times 10 + 0 = 20, \quad aa/Bb = 2 \times 10 + 1 = 21, \quad aa/bb = 2 \times 10 + 2 = 22$$

Thus using this encoding, we guarantee that different combinations of epistasis effects have different encodings and we do not need to worry about the assignment of different values to these combinations. Another benefit is that the encoding can be applied to any t-locus interactions in a systematic way. We call this sampling *Encoding-based Sampling.*

3.4.3. *Iterative Sampling*

As we are using sampling to estimate the mean and standard deviation of the normal distribution, it is critical to determine the sample size first. Given an expected error rate, we could estimate the sample size via Equation 3.

$$ME = z\frac{s}{\sqrt{n}} \tag{3}$$

Where ME is the desired margin of error, z is the z-score that depends on the desired confidence level, s is the standard deviation and n is the sample size we want to find. Given the desired margin of error and the confidence level, if we know the standard deviation or we could make a guess on it, we could compute the required sample size n.

However, our problem is much more complicated in that every t-locus interaction has different mean and standard deviation. Therefore it is not appropriate to use an universal sample size and there is no systematic way to estimate the standard deviation for each t-locus interaction.

To address the problem, we propose an iterative sampling method. In iteration one, for every t-locus interaction, we start from a small initial sample size, say, 500, and estimate mean μ_1 and standard deviation δ_1. Then we increase the sample size by 500 for every iteration. In iteration i, we estimate mean μ_i and standard deviation δ_i. If $\frac{abs(\mu_i - \mu_{i-1})}{\mu_i} \leq \epsilon$ and $\frac{abs(\delta_i - \delta_{i-1})}{\delta_i} \leq \epsilon$, where ϵ is a small number such as 0.01, or the number of iterations is greater than a prespecified number, such as 10, we say that the sampling converges.

Notice that MUSE selects the top-k most significant interactions. After the selection, we combine these interactions with the original set of single markers as a new data set. Regression methods such as rrBLUP are then applied on the new data set to make predictions. Notice k is a user defined parameter. The smaller k is, the more efficient MUSE is. Ideally k could be selected using cross-validation. However, given the extremely large feature space, it is not feasible to try all possible k's. Therefore in our work, we just simply set k as 500, a small number. Our experiments showed that by setting k as 500, we could already achieve significant improvements and yet the program is highly efficient.

4. Experimental Results

We first evaluated MUSE on a plant data set: Maize data set,[2] the Dent and Flint panels, developed for the European CornFed program. We do not consider using simulated data here as the rational for how high order multi-locus interactions contribute to the trait is indeed not clear. As the number of multi-locus interactions is extremely high when the order is high, it is not clear what is a reasonable number of the interactions that contribute to the trait.

The Maize data set indeed consists of 6 sub data sets. The Dent panel were genotyped using a 50k SNP array, which after removing SNPs with high rate of missing markers and high average heterozygosity, yielded 29,094 and 30,027 SNPs respectively. Both of them contain 261 samples and three traits. In all experiments, we perform 10-fold cross-validations and measure the average r^2 between the true and the predicted outputs, where higher r^2 indicates better performance. The parameters are learned from the training data. The baseline method is rrBLUP with single marker model using all markers. For a fair comparison, we use the top-500 most significant interactions (for k-locus interactions where $k \geq 2$) captured by MUSE and we combine them with the original set of single markers as a new data set where rrBLUP is then applied. This will indicate whether the extra information from the interactions benefit the prediction. Notice we mark the performance as "NA" for cases where no significant interaction is captured.

We evaluate the performance of MUSE with the constraint-based sampling (MUSE-C), with the encoding-based sampling (MUSE-E) and with iterative sampling (MUSE-I). We consider only 2-locus scenarios where the p-value p=5×10^{-8}. For the constraint-based sampling, overlap threshold N=5. The baseline method is rrBLUP with single marker model using all markers. As we can see in Table 1, MUSE improves the performance over rrBLUP significantly. As MINED does not use p-values as a criteria to select interactions, its performance is worse than SAME and MUSE. MUSE with the constraint-based sampling (MUSE-C) generally is able to improve the prediction accuracy over SAME, as the constraint-based sampling is able to naturally reduce the redundancy of the sampled interactions, which further leads to improvement on the prediction. MUSE with both the constraint-based sampling and the encoding-based sampling (MUSE-CE) achieve better results except for Flint Trait 3, indicating that both constraint-based sampling and encoding-based sampling are effective in improving the prediction accuracy. For Flint Trait 3, when constraint-based sampling is used, MUSE can not capture any interaction with p-value lower than 5×10^{-8}. However, after we conducted the iterative sampling, MUSE is able to capture interactions with p-value lower than 5×10^{-8} and thus MUSE-CEI achieved the best performance among all the methods. This clearly indicates the power of iterative sampling. In general combining all three sampling strategies gives us the best performance.

Table 1. The r^2 of rrBLUP, MINED, SAME, MUSE on Maize Dent and Flint data sets. We show only 2-locus scenarios where p-value $p{=}5 \times 10^{-8}$, overlap threshold $N{=}5$. For MUSE-C and MUSE-CE, the number of initial sampling is 500. Here for MUSE, "-C" stands for constraint-based sampling, "-E" stands for encoding-based sampling, "-I" stands for iterative sampling.

Trait	rrBLUP	MINED	SAME	MUSE-C	MUSE-CE	MUSE-CEI
Dent Trait 1	0.59	0.59	0.615	0.65	0.65	0.67
Dent Trait 2	0.552	0.552	0.583	0.572	0.59	0.61
Dent Trait 3	0.321	0.356	0.432	0.39	0.486	0.49
Flint Trait 1	0.47	0.476	0.514	0.558	0.576	0.595
Flint Trait 2	0.301	0.316	0.356	0.364	0.419	0.429
Flint Trait 3	0.057	0.096	0.113	NA	NA	0.135

In Table 2, we evaluated 2-locus, 3-locus and 4-locus interactions for MUSE. As we have already shown that MUSE-CEI in general achieves the best performance, we only evaluate the performance of MUSE-CEI. We also varied the overlap thresholds as 5, 20, 50. The running times for MUSE-CEI are 226 sec., 979 sec. and 2056 sec. respectively. As we can see, although the size of the feature space increased exponentially, the running time of MUSE-CEI did not change much, indicating that MUSE-CEI is highly scalable due to its effective sampling process. The baseline method is again rrBLUP with single marker model using all markers.

Overall, we can see that MUSE-CEI achieved very significant improvements over rrBLUP on the single marker model (For Dent data, 21% for trait 1, 22% for trait 2, 59% for trait 3. For Flint Data, 33% for trait 1, 46% for trait 2, 138% for trait 3). We can see that both the p-value and the overlap threshold N are critical to the prediction. The best p-value and N are usually different without clear pattern for different traits and we need to use grid search to find their best values.

By varying the p-values, the prediction performance varies significantly. In general, the p-value should be small enough to achieve the best prediction. However, we do not see a clear pattern on setting the p-values. For different traits, the best p-value could be different. And it is not necessarily the case that using smaller p-value leads to better prediction accuracy. This is because smaller p-values may only produce a small set of statistically significant epistasis effects where larger p-values may produce a larger set of statistically significant epistasis effects. If the size of the set of statistically significant epistasis effects is too small and in the meanwhile they do not have very high r^2 score, they might not be able to improve the prediction performance. In the worst case, we might not be able to identify any significant k-locus interaction given a too small p-value might lead, such as Dent Trait 3 with 3-locus

$p = 5 \times 10^{-12}$ and Flint Trait 3 with 3-locus $p = 5 \times 10^{-12}$ and 4-locus $p = 5 \times 10^{-11}$. As we did not observe a clear pattern between p-values and the prediction performance, grid search with cross-validation should be applied in order to detect the best p-value.

Table 2. The r^2 of rrBLUP and MUSE on Maize Dent and Flint data set. For MUSE, we tested 2-locus, 3-locus and 4-locus interactions with different p-value thresholds. We applied all the sampling strategies. We vary the p-value and the constraint threshold N.

Methods	N=5	N=20	N=50	N=5	N=20	N=50
	Dent Trait 1			**Flint Trait 1**		
rrBLUP	0.59			0.47		
MUSE-CEI 2-locus (p=5×10^{-8})	0.67	0.581	0.58	0.595	0.591	0.568
MUSE-CEI 2-locus (p=5×10^{-10})	0.645	0.655	0.616	**0.626**	0.615	0.586
MUSE-CEI 2-locus (p=5×10^{-11})	0.63	0.693	0.656	0.56	0.583	0.556
MUSE-CEI 3-locus (p=5×10^{-10})	0.538	0.644	0.491	0.578	0.618	0.59
MUSE-CEI 3-locus (p=5×10^{-11})	0.675	**0.714**	0.59	0.617	0.62	0.57
MUSE-CEI 3-locus (p=5×10^{-12})	0.606	0.65	0.673	0.601	0.61	0.581
MUSE-CEI 4-locus (p=5×10^{-11})	0.27	0.384	0.601	0.47	0.488	0.301
	Dent Trait 2			**Flint Trait 2**		
rrBLUP	0.552			0.301		
MUSE-CEI 2-locus (p=5×10^{-8})	0.61	0.552	0.563	0.429	0.412	0.403
MUSE-CEI 2-locus (p=5×10^{-10})	0.663	0.557	0.564	0.413	0.427	0.394
MUSE-CEI 2-locus (p=5×10^{-11})	**0.671**	0.595	0.574	0.417	0.415	0.373
MUSE-CEI 3-locus (p=5×10^{-10})	0.608	0.459	0.459	0.428	**0.439**	0.418
MUSE-CEI 3-locus (p=5×10^{-11})	0.623	0.491	0.491	0.423	0.421	0.402
MUSE-CEI 3-locus (p=5×10^{-12})	0.582	0.625	0.549	0.382	0.395	0.399
MUSE-CEI 4-locus (p=5×10^{-11})	0.3	0.335	0.258	0.37	0.365	0.298
	Dent Trait 3			**Flint Trait 3**		
rrBLUP	0.321			0.057		
MUSE-CEI 2-locus (p=5×10^{-8})	0.49	0.424	0.361	0.135	0.12	0.087
MUSE-CEI 2-locus (p=5×10^{-10})	0.355	0.476	0.466	0.115	0.126	0.103
MUSE-CEI 2-locus (p=5×10^{-11})	0.332	0.397	0.465	0.097	0.067	0.048
MUSE-CEI 3-locus (p=5×10^{-10})	0.482	0.391	0.443	0.089	0.111	0.103
MUSE-CEI 3-locus (p=5×10^{-11})	0.453	0.347	0.398	0.120	**0.136**	0.119
MUSE-CEI 3-locus (p=5×10^{-12})	NA	NA	0.358	NA	NA	0.026
MUSE-CEI 4-locus (p=5×10^{-11})	0.341	**0.511**	0.444	NA	NA	0.046

Another observation is that smaller N in general leads to better performance. This clearly indicates the effects of redundancy: when N is large, we allow more redundant interactions to be selected and thus the performance drops. However, a small N may prevent selecting significant interactions as the pool of interactions to be sampled is dramatically reduced for small N. For example, for Dent Trait 3, p=5×10^{-12}, when N=5 and 20, MUSE can not capture any 3-locus significant interactions. However, when $N = 50$, MUSE could capture some 3-locus significant interactions. Similarly, for Flint Trait 3, 3-locus and 4-locus significant interactions are only captured when $N = 50$.

In summary we observed that although there is no clear pattern for the optimal p-value and overlap threshold N, we see that in general a too large N or a too small p-value lead to poorer performance. Also for higher order interactions, the number of detected significant interactions might be too small to lead improvements.

One more thing to notice is that we do not conduct biological validation on the interactions MUSE selected. This is because we assume all the interactions contribute to the trait more or less. The selected interactions also have lots of peers which have similar r^2 scores. However, we are only able to select a small set of interactions due to efficiency concerns. These interactions are selected by random chance from the pool of interactions with similar r^2 scores. But our experiments illustrated that a small set of interactions is sufficient to improve the genetic trait prediction accuracy dramatically.

Besides plant traits, we also conducted experiments on complex trait for humans. Complex traits prediction and association are crucial to translate the findings in genetics to precision medicine. We studied the data set from the Finland-United States Investigation of NIDDM Genetics (FUSION) study,[20] which is a long-term effort to identify genetic variants that predispose to type 2 diabetes (T2D) or that impact the variability of T2D-related quantitative traits. The dataset has 5000 individuals, 317503 SNPs and 10 traits. For illustration purpose, we show the results on two randomly selected traits (trait 2: HDL-cholesterol, trait 10: Height).

In Table 3, we showed the performance of MUSE on two human complex traits. We can see that in general the predictions are poor, indicating the difficulties of complex trait prediction. However, even on complex traits, we see that by integrating interactions into the predictive model, we can still achieve significant improvements. And by tuning the parameters carefully, MUSE can achieve better performance compared with existing methods. Again, we see that with relatively small N and p-value, MUSE achieved better performance.

5. Conclusion and Future Work

In this work, we studied the multi-locus epistasis problem where the interactions with more than two SNPs are considered. We developed an algorithm MUSE which is very efficient for multi-locus epistasis model. We also showed that the algorithm is very effective in improving the performance of the genetic trait prediction. Three sampling strategies are developed which could improve the overall prediction accuracy. More accurate trait predictions can be very helpful to develop breeding strategies and is crucial to translate the findings in genetics to precision medicine.

References

1. T. Meuwissen, B. Hayes and M. Goddard, *Genetics* **157**, 1819 (2001).
2. R. Rincent, D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V. M. Rodriguez, J. Moreno-Gonzalez, A. Melchinger, E. Bauer *et al.*, *Genetics* **192**, 715 (2012).
3. M. A. Cleveland, J. M. Hickey and S. Forni, *G3: Genes— Genomes— Genetics* **2**, 429 (2012).
4. J. Whittaker, R. Thompson and M. Denham, *Genet Res* **75**, 249 (2000).

Table 3. The r^2 of rrBLUP, MINED, SAME and MUSE on Finland data set. For MUSE, we tested 2-locus, 3-locus and 4-locus interactions with different p-value thresholds. We applied all the sampling strategies. We vary the p-value and the constraint threshold N.

Methods	N=5	N=20	N=50	N=5	N=20	N=50
	Trait HDL-cholesterol			**Trait Height**		
rrBLUP	0.11			0.03		
MINED	0.15			0.07		
SAME	0.18			0.10		
MUSE-CEI 2-locus (p=5×10^{-8})	0.14	0.15	0.16	0.04	0.03	0.04
MUSE-CEI 2-locus (p=5×10^{-10})	0.15	0.17	0.17	0.05	0.07	0.05
MUSE-CEI 2-locus (p=5×10^{-11})	0.16	0.18	0.19	0.05	0.06	0.06
MUSE-CEI 3-locus (p=5×10^{-10})	0.16	0.18	0.18	0.07	0.08	0.06
MUSE-CEI 3-locus (p=5×10^{-11})	0.17	0.2	0.19	0.08	0.11	0.1
MUSE-CEI 3-locus (p=5×10^{-12})	0.2	**0.22**	0.21	0.1	0.09	0.08
MUSE-CEI 4-locus (p=5×10^{-11})	0.12	0.15	0.18	0.1	**0.12**	0.11

5. R. Tibshirani, *Journal of the Royal Statistical Society, Series B* **58**, 267 (1994).
6. S. S. Chen, D. L. Donoho, Michael and A. Saunders, *SIAM Journal on Scientific Computing* **20**, 33 (1998).
7. K. Kizilkaya, R. Fernando and D. Garrick, *Journal of animal science* **88**, 544 (2010).
8. A. Legarra, C. Robert-Granié, P. Croiseau, F. Guillaume, S. Fritz *et al.*, *Genetics research* **93**, p. 77 (2011).
9. T. Park and G. Casella, *Journal of the American Statistical Association* **103**, 681 (June 2008).
10. K. A. Pattin, B. C. White, N. Barney, J. Gui, H. H. Nelson, K. T. Kelsey, A. S. Andrew, M. R. Karagas and J. H. Moore, *Genetic epidemiology* **33**, 87 (2009).
11. J. Marchini, P. Donnelly and L. R. Cardon, *Nature genetics* **37**, 413 (2005).
12. N. R. Cook, R. Y. Zee and P. M. Ridker, *Statistics in medicine* **23**, 1439 (2004).
13. C. Yang, Z. He, X. Wan, Q. Yang, H. Xue and W. Yu, *Bioinformatics* **25**, 504 (2009).
14. Y. Zhang and J. S. Liu, *Nature genetics* **39**, 1167 (2007).
15. G. Fang, M. Haznadar, W. Wang, H. Yu, M. Steinbach, T. R. Church, W. S. Oetting, B. Van Ness and V. Kumar, *PloS one* **7**, p. e33531 (2012).
16. X. Zhang, S. Huang, F. Zou and W. Wang, *Bioinformatics* **26**, i217 (2010).
17. D. He, Z. Wang and L. Parada, Mined: An efficient mutual information based epistasis detection method to improve quantitative genetic trait prediction, in *Bioinformatics Research and Applications*, (Springer, 2015) pp. 108–124.
18. D. He and L. Parida, Same: a sampling-based multi-locus epistasis algorithm for quantitative genetic trait prediction, in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, 2015.
19. H. Peng, F. Long and C. Ding, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27**, 1226 (2005).
20. E. Zeggini, L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini, T. Hu, P. I. de Bakker, G. R. Abecasis, P. Almgren, G. Andersen *et al.*, *Nature genetics* **40**, 638 (2008).

ceRNA SEARCH METHOD IDENTIFIED A MET-ACTIVATED SUBGROUP AMONG EGFR DNA AMPLIFIED LUNG ADENOCARCINOMA PATIENTS

HALLA KABAT*

Outreach Program, miRcore, 2929 Plymouth Rd.
Ann Arbor, MI 48105, USA
Email: halla203@gmail.com

LEO TUNKLE*

Outreach Program, miRcore, 2929 Plymouth Rd.
Ann Arbor, MI 48105, USA
Email: leotunkle@gmail.com

INHAN LEE

miRcore, 2929 Plymouth Rd.
Ann Arbor, MI 48105, USA
Email: inhan@mircore.org

Given the diverse molecular pathways involved in tumorigenesis, identifying subgroups among cancer patients is crucial in precision medicine. While most targeted therapies rely on DNA mutation status in tumors, responses to such therapies vary due to the many molecular processes involved in propagating DNA changes to proteins (which constitute the usual drug targets). Though RNA expressions have been extensively used to categorize tumors, identifying clinically important subgroups remains challenging given the difficulty of discerning subgroups within all possible RNA-RNA networks. It is thus essential to incorporate multiple types of data. Recently, RNA was found to regulate other RNA through a common microRNA (miR). These regulating and regulated RNAs are referred to as competing endogenous RNAs (ceRNAs). However, global correlations between mRNA and miR expressions across all samples have not reliably yielded ceRNAs. In this study, we developed a ceRNA-based method to identify subgroups of cancer patients combining DNA copy number variation, mRNA expression, and microRNA (miR) expression data with biological knowledge. Clinical data is used to validate identified subgroups and ceRNAs. Since ceRNAs are causal, ceRNA-based subgroups may present clinical relevance. Using lung adenocarcinoma data from The Cancer Genome Atlas (TCGA) as an example, we focused on EGFR amplification status, since a targeted therapy for EGFR exists. We hypothesized that global correlations between mRNA and miR expressions across all patients would not reveal important subgroups and that clustering of potential ceRNAs might define molecular pathway-relevant subgroups. Using experimentally validated miR-target pairs, we identified EGFR and MET as potential ceRNAs for miR-133b in lung adenocarcinoma. The EGFR-MET up and miR-133b down subgroup showed a higher death rate than the EGFR-MET down and miR-133b up subgroup. Although transactivation between MET and EGFR has been identified previously, our result is the first to propose ceRNA as one of its underlying mechanisms. Furthermore, since MET amplification was seen in the case of resistance to EGFR-targeted therapy, the EGFR-MET up and miR-133b down subgroup may fall into the drug non-response group and thus preclude EGFR target therapy.

* These authors contributed equally to this work.

1. Introduction

Lung cancer accounts for more deaths than any other cancers, with a 5-year survival rate of 10% [1]. Several gene mutations have been shown to play a role in lung adenocarcinoma (LUAD), including KRAS and EGFR [2]. Multiple drugs have been developed to target EGFR proteins and are actively used for those with EGFR mutation cancers. However, some patients do not respond to the targeted therapy and many initially responded patients develop resistance to such drugs. Since diverse molecular pathways are associated with any mutated genes, additional information other than DNA mutation is needed to properly identify which subgroup will benefit from the targeted therapy.

Though mRNA expression data have been used to categorize tumors, correlations across mRNA expression data alone are often difficult to decipher within high dimensional data. Moreover, the most correlated genes or samples often do not provide clinically useful insight. To increase the signals, other types of data such as DNA methylation, microRNA (miR) expression, proteomics, and metabolic data have been incorporated with mRNA expression. In terms of RNA levels, mRNA and miR expression correlations have been heavily mined. However, the lack of known miR targets and excessive false positive target predictions hinder the computational search for significant miR-target gene networks. Worse, since miR effects on most target genes are small in degree, in vitro experimental confirmation is difficult, although the effects may contribute to long term clinical outcomes. Usual mRNA-miR expression analyses calculate correlations among RNAs for all samples.

Studies have recently demonstrated that RNAs can compete with one another for the same regulating miRNAs [3]. One of the earliest of these studies, focused on expression of PTEN, hypothesized that expression levels of “competing endogenous” RNAs (ceRNAs) affected PTEN expression. When siRNAs were used to deplete these RNAs, PTEN expression levels also decreased. Decreased ceRNA levels resulted in fewer miRNAs (which target both the ceRNA and PTEN) being “used-up” in regulation. This frees more of these miRNAs to target PTEN, thereby decreasing its expression. Overall, a decrease in expression of a ceRNA results in a corresponding decrease in PTEN. The same study also demonstrated that an increase in expression of a ceRNA corresponded with an increase in PTEN. This is likely applicable not just to PTEN but to other genes, such as a gene and a similar pseudogene or two genes regulated by the same miRNA. Note that ceRNA by definition entails causality whereas usual mRNA-miR expression results are correlative. RNA expression changes cause other RNA expression changes through miR manipulation.

This RNA-RNA regulation inspired two lines of investigation: biochemical inquiry to identify individual ceRNA pairs [4-6] and bioinformatics research to identify global RNA-RNA networks using RNA expression data along with miR-target predictions [7,8]. Ideal conditions for miRs and ceRNAs have also been explored [9,10]. However, global ceRNA networks are difficult to discern due to imprecise miR target prediction and because, again, the miR effect on one target gene is usually small. Such small degree changes are difficult to identify from multi-layer RNA-RNA regulations of diverse samples though ceRNAs have been associated with diseases and have the potential to uncover disease progression [11].

The Cancer Genome Atlas (TCGA) [12] provides a large amount of various types of data from multiple cancers, enabling new ways of data analysis. For example, LUAD data include the mRNA and miRNA (miR) expressions of 551 patients that could provide insight into multiple biological processes within tumors. This large mass of patient data allows for identification of subgroups based upon very specific traits.

In this study, the concept of ceRNAs was utilized to identify a subgroup related to DNA mutations. We focused on patients with amplified EGFR to identify those who could benefit from EGFR targeted therapy, analyzing multiple datasets including copy number variation (CNV), RNAseq, and miRNAseq from TCGA in order to find the EGFR amplification signature. RNA and miR interactions were then identified using a database of experimentally validated miR-target genes from miRTarBase [13]. Our findings suggest that miR-133b, which targets EGFR, is downregulated due to high mRNA expression for EGFR caused by its DNA amplification, which in turn leads to the upregulation of MET, another gene targeted by miR-133b. In short, EGFR amplification is linked to MET mRNA upregulation through miR-133b, which targets both EGFR and MET in a manner reminiscent of the ceRNA interactions mentioned above. To our knowledge, our research is the first to identify disease subgroups based upon ceRNA interactions, an approach with potential application to other gene mutations or in other types of cancers.

2. Methods

Most research into downstream effects of DNA mutations has focused on protein functions. Here we propose using the ceRNA concept to analyze downstream events of DNA mutation to complement conventional protein-centric biology and to identify RNA-RNA networks (Fig. 1).

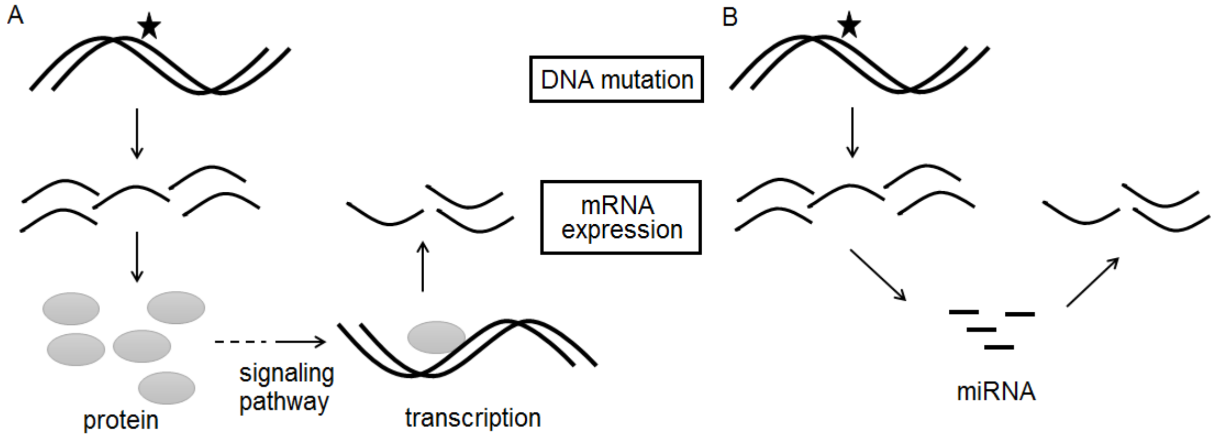


Fig. 1. Underlying biological concepts in mRNA expressions related to DNA mutations. (A) Protein-centric concept. A DNA mutation leads to protein expression changes, resulting in other mRNA changes through signaling pathways. These downstream mRNAs are RNAs of interest. (B) ceRNA concept. If DNA mutation leads to ceRNA upregulation, the “used-up” miRs would fail to regulate the ceRNA pair and thus increase mRNA expression. Similarly, if ceRNA is downregulated, the pairing ceRNA would be downregulated. miR expression data and miR target information are needed to elucidate this process.

2.1. *Overview of data analysis pipeline*

Fig. 2 shows the overall data analysis pipeline to identify subgroups related to a certain DNA amplification [deletion]. Including DNA information may reveal DNA mutation-related ceRNAs, reducing the search space for ceRNA networks. The overall process requires downloading copy number variation (CNV), mRNAseq, miRseq, and clinical data from TCGA and miR-target pairs with strong experimental evidence from miRTarBase.

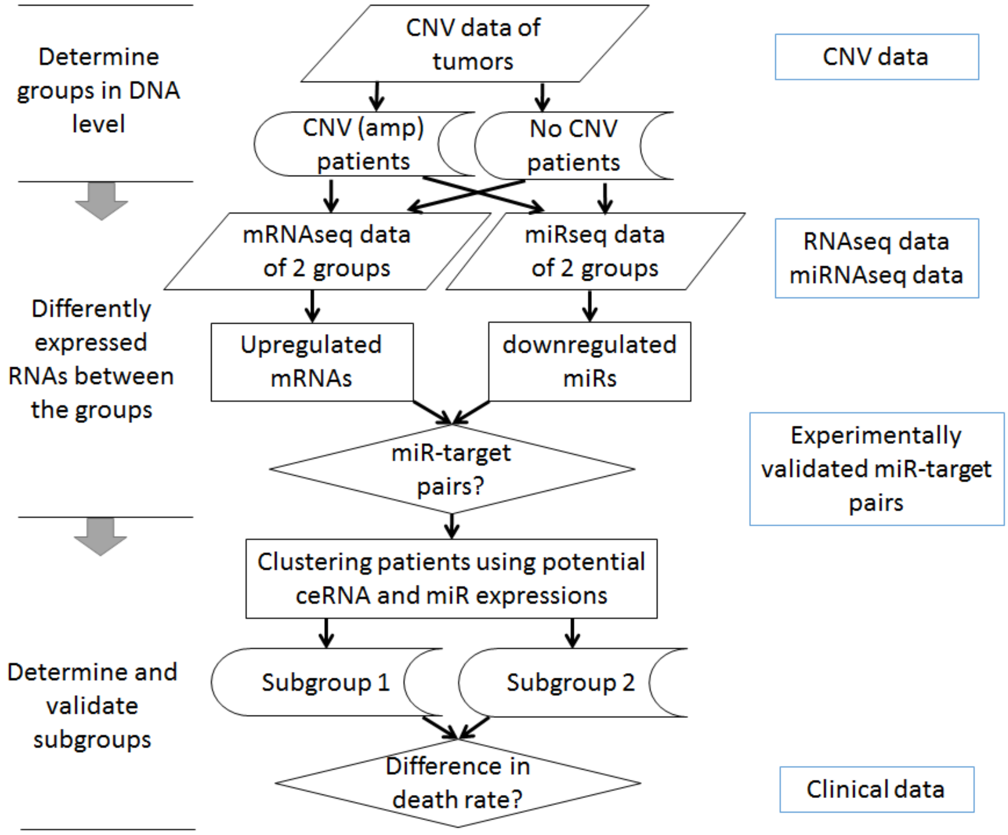

Determine groups in DNA level
CNV data of tumors
CNV data
CNV (amp) patients
No CNV patients
mRNAseq data of 2 groups
miRseq data of 2 groups
RNAseq data miRNAseq data
Differently expressed RNAs between the groups
Upregulated mRNAs
downregulated miRs
miR-target pairs?
Experimentally validated miR-target pairs
Clustering patients using potential ceRNA and miR expressions
Determine and validate subgroups
Subgroup 1
Subgroup 2
Difference in death rate?
Clinical data

Fig. 2. Overall data analysis process to identify ceRNA-based subgroups. Here an example of amplified CNV genes is shown, with only upregulated mRNAs for clarity.

2.2. *TCGA CNV data analysis*

We downloaded CNV data of all LUAD tumor samples from TCGA and translated chromosome locations to gene-level information using TCGA-Assembler [14]. Overall tumor characteristics were assessed by average CNV values of each gene in chromosome seven (EGFR location) for all tumor samples. Individual tumors' EGFR CNV values were then sorted to determine if the sample number of the EGFR group was adequate. We used DNA copy numbers greater than 3 to define the EGFR amplified group (EGFR amp) and defined the control group as having a copy number between 1.97 and 2.03, yielding a sample size similar to that of EGFR amp. The corresponding log2(CNV/2) for the amp and the control groups is 0.58 and -0.02 to 0.02, respectively.

2.3. *TCGA mRNA and miRNA expression data analysis*

To analyze mRNA expression data, rsem.genes.normalized_results files for RNASeqV2 data of all samples were downloaded using TCGA-Assembler. Data for the EGFR amp and control groups were then extracted. Some patients did not have available rsem-normalized RNAseq data or miRseq data, and were removed from any further analysis. After confirmation of normalization across samples, a student t-test was conducted to compare the amp and the control group data.

To analyze miR expressions, isoform.quantification files for miRNAseq data were downloaded from the TCGA Data Matrix and converted to mature miR values. These individual files were then combined to make a matrix file for all patients. The R code for this function can be found in GitHub (https://github.com/rptashkin/TCGA_miRNASeq_matrix). Upper quartile normalization was applied for student t-test analysis between the amp and control groups, upon which the miRNAs with p-values < 0.05 were separated into up- and downregulated groups.

2.4. *Validated miR target finding*

To see if the miRNAs and genes had potential interactions, data from miRTarBase, a database of miRNA-target interactions, were used. The upregulated genes and downregulated miRNAs were compared to the miR-target pairs with strong experimental evidence to search for any pairs.

2.5. *Subgroup determination and validation*

A heatmap of potential ceRNAs and miRNAs of interest was used to determine the subgroups formed. The patients were clustered using Pearson correlation, and subgroups were determined based on the clustering trees where the mRNA and miRNA expressions of all patients within the trees exhibit negative correlations between miR-targets and positive correlations between ceRNAs. A survival graph was prepared using R and the death rate differences between the groups were tested using student t-test.

3. Results

3.1. *EGFR-amplified patients with lung adenocarcinoma*

The average CNV of genes on chromosome seven from 551 LUAD tumor samples was calculated to assess overall CNV signatures across the entire chromosome (Fig. 2A). One of the two peaks in chr7 corresponds to the EGFR location, confirming the existence of EGFR amplification in these tumor samples strong enough for analysis. To understand the EGFR CNV status of individual patients' tumors, we sorted 551 tumor samples in terms of EGFR CNV values (Fig. 2B). The number of tumors with amplified EGFR copy numbers is much more than that with reduced copy numbers; some tumors showed distinctively amplified EGFR. Using the CNV cutoff value of three, there were a total 50 patients in the EGFR amp group and 56 patients in the control group.

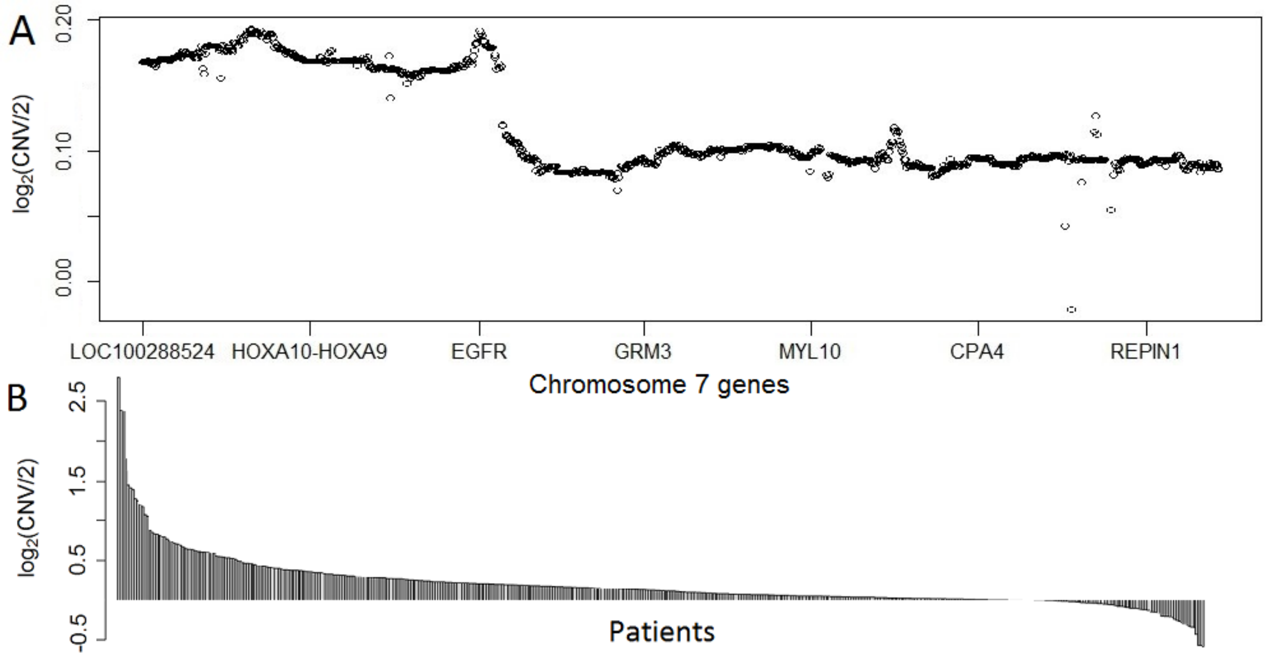


Fig. 3. CNV data for LUAD patients. (A) Average $\log_2$(CNV) values of genes on chromosome 7 for all patients, ordered by chromosome position. (B) EGFR values for the 551 tumor samples.

3.2. *RNA and miRNA expression analysis*

After we downloaded the rsem-normalized data from TCGA, we confirmed the normalization status using box plots. Using the patient lists in the EGFR amp and control groups identified from CNV data, mRNA expression data were extracted and organized for each group. We used isoform.quantification data to obtain mature miR reads for miR expression data analysis. The isoform data were translated to mature miR names and all reads corresponding to the same mature miRs were combined. All EGFR amp and control group patient miR data were merged into a matrix file. Upper quartile normalization was used for miR data and box plots of data before and after normalization were compared to ensure the normalization status. We used only those samples having both mRNA and miR data for further analysis, leaving 42 amp and 35 control patients.

Student t-test was used to identify differently-expressed genes between the two groups of patient samples since the sample number is large. A heatmap of mRNAs with student t-test p-value < 0.0001 (for visual purpose) is shown in Fig. 4A and that of miRs with p-value < 0.05 in Fig. 4B, together with the EGFR amp and control ID labels on top of each heatmap. The unsupervised hierarchical clustering of mRNA expressions identified two large groups: one mostly control and the other mostly amp group. Additionally, the amp group displays a greater number of upregulated genes than does the control group. The mRNA expression of EGFR (p-value of 1.62 x 10^{-6}), is excluded in this heatmap. The miR clustering also identified two large groups: one with all amp and the other generally with control samples.

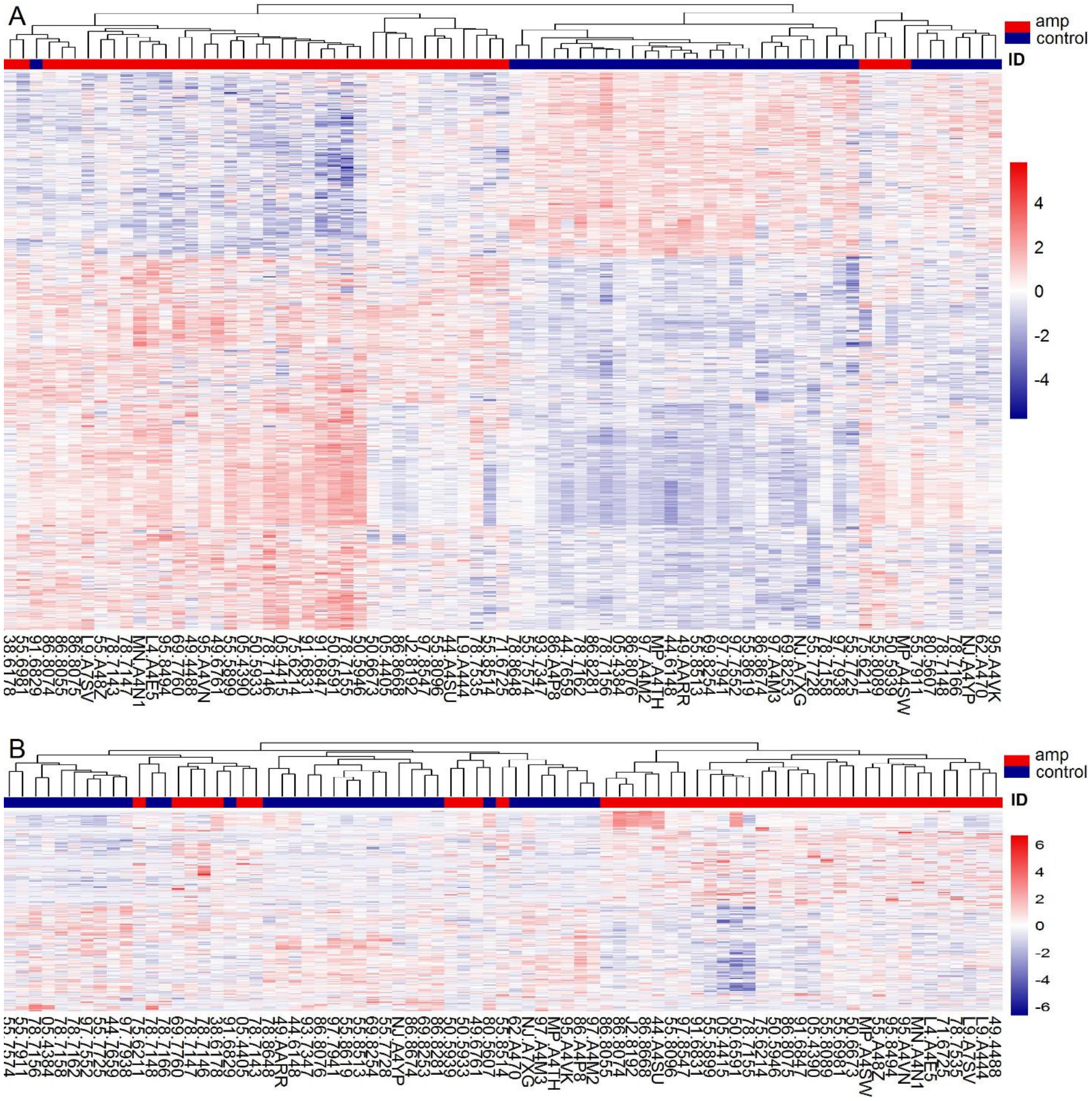


Fig. 4. Hierarchical clustering of mRNA (A) and miRNA expression (B). ID above the heatmap represents the amp group in red and the control in blue. The patient IDs for each group can be found below the heatmap.

3.3. *Identifying miR-target RNA pairs*

We used all mRNAs and miRs with p-values less than 0.05 to find experimentally validated miR-target pairs, since such pairs are still highly limited. To ensure miR-target pair validity, we only used pairs found through strong experimental evidence from miRTarBase. Strong evidence includes validating with a reporter assay, a western blot analysis, or qPCR experiments. Also, since we are looking into direct downstream events of EGFR amplification, only upregulated mRNAs and downregulated miRs in EGFR amp groups were considered.

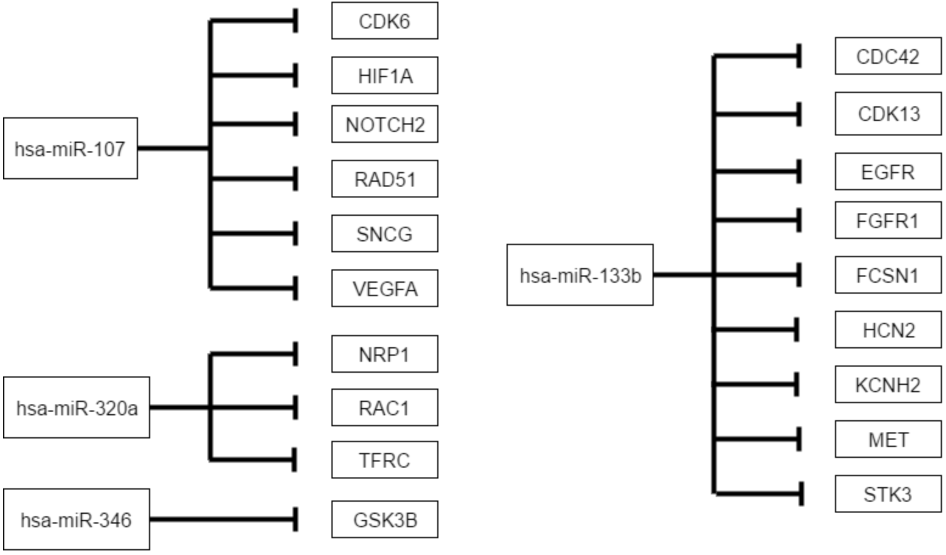


Fig. 5. Validated miRNA-RNA target pairs. The validated target pairs from upregulated mRNAs and downregulated miRNAs with $p < 0.05$.

A total 19 miR-target pairs were identified in the up-mRNAs and down-miR groups, including 4 miRNAs and 19 genes (Fig. 5). One of these pairs included EGFR, a known target of miR-133b. Interestingly, previous studies found miR-133 mediating ceRNAs of mRNA pairs, making miR-133b a good candidate mediator for ceRNAs. Eight other miR-133b targets were found in the upregulated mRNAs, with $p < 0.05$, some possibly functioning as ceRNAs for EGFR through miR-133b in certain patient tumors.

Among them, we decided to focus on MET, given its well-established EGFR and MET crosstalk [15,16], particularly related to drug resistance [17]. The fold changes of EGFR, MET, and miR-133b between EGFR amp and control groups are 6.68, 1.79, and 0.318, respectively; and corresponding p-values for MET and miR-133b are 0.0065 and 0.00085. To exclude other ways of increasing MET mRNA expressions in our dataset, we confirmed that 1) MET copy numbers did not vary in the EGFR amp groups; 2) the expression values of ETS1/2, PAX3, and TCF4, known transcription factors of MET [18], are not upregulated; and 3) ERBB3, known to activate MET [19], is not activated in the EGFR amp groups.

3.4. *Subgroup identification*

To identify patients with potential EGFR-miR-133b-MET interactions, unsupervised hierarchical clustering with only miR-133b, EGFR, and MET were calculated using Pearson correlation distance (Fig. 6A). With a tree cutting of four groups, a subgroup featuring high EGFR-MET and low miR-133g (24 patients) and another subgroup with low EGFR-MET and high miR-133b (24 patients) were identified (boxed in Fig. 6A). Overall Pearson correlation coefficients between

EGFR and MET, EGFR and miR-133b, and MET and miR-133b for all 77 patients are 0.082, -0.030, and 0.082, respectively, unlikely to be identified by global RNA-RNA network analysis of all patients. The correlation coefficients across these 48 patients became 0.22, -0.24, and -0.23, respectively.

To validate these two subgroups, we downloaded patient clinical data from TCGA. As seen in the survival curve (Fig. 6B), these two groups presented different survival rates (student t-test p-value 0.016). Given the known EGFR-MET transactivation, we wondered if subgrouping may also emerge using EGFR and MET expressions alone. We could not see a clear pattern in the clustered heatmap using Pearson correlation distance method, but two clusters showed up using the Euclidean method. The p-value of survival rate differences between these groups was 0.15. Therefore, subgroups identified from EGFR-miR-133b-MET expression data presented stronger clinical implications.

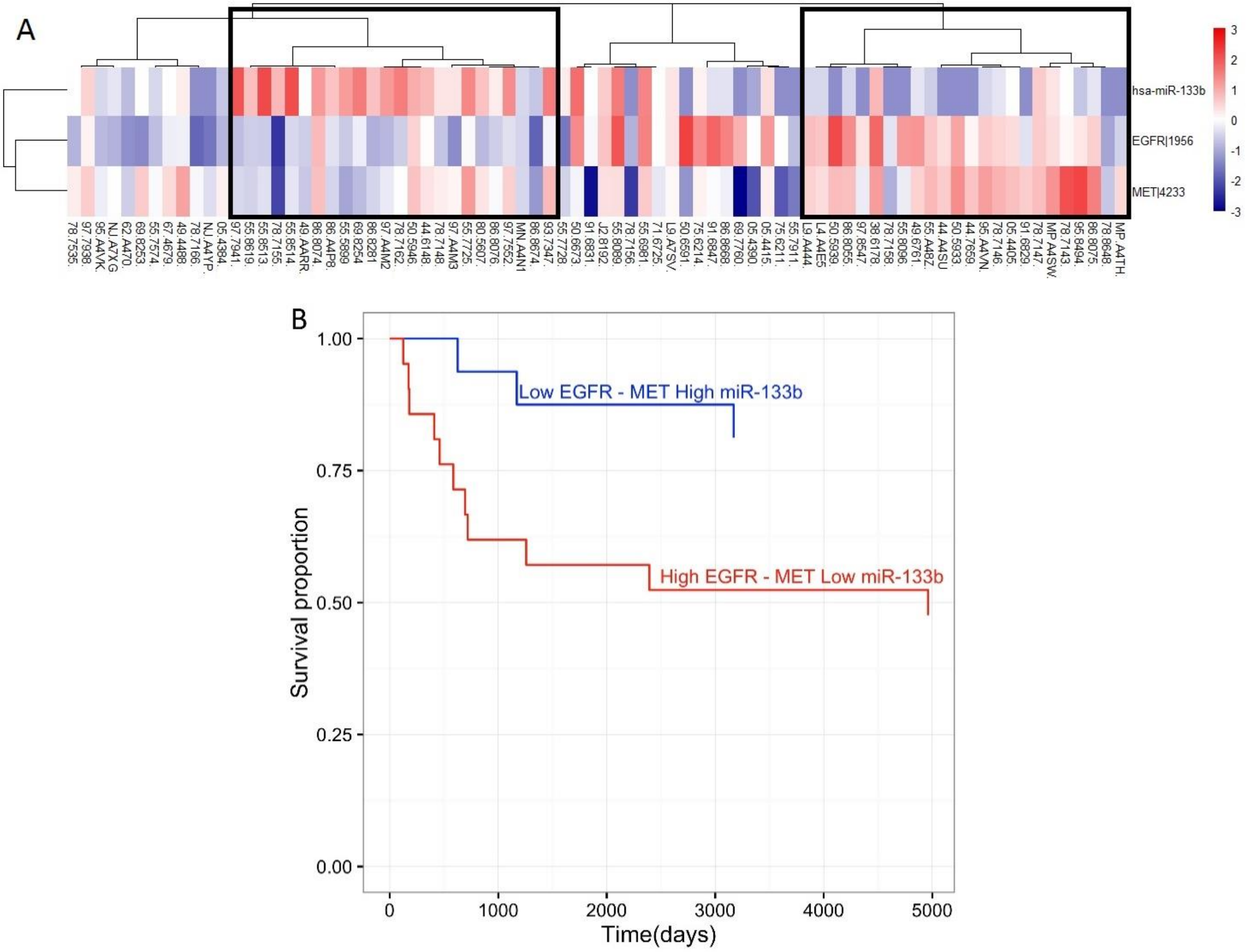


Fig. 6. Subgroup selection and survival curve. (A) Clustered heatmap of EGFR, MET, and miR-133b. The two boxes show the subgroups made through clustering. These subgroups have high miR-133b, low EGFR, and low MET or low miR-133b, high EGFR, and high MET. (B) Survival curves for the subgroups.

4. Discussion

EGFR is one of the more common mutations in lung adenocarcinoma and there exist targeted therapy options for those with this mutation. These currently include drugs such as gefitinib and erlotinib [20]. Though these therapies work well for many patients initially, most patients encounter drug resistance. Of the tumors that develop resistance to these drugs, around 20% have MET amplification [21].

MET, like EGFR, is a growth factor receptor that leads to several signaling cascades including those within the RAS-ERK pathway, which is often targeted by cancer drugs. When functioning normally, MET is essential to such processes as angiogenesis, wound healing, and liver regeneration [22].

Since there is a correlation between MET amplification and drug resistance to an EGFR-targeted therapy, studies have focused on transactivation of EGFR and MET [16-18] though their mechanism has not been cleared elucidated. On the other hand, searching for ceRNA pairs as signature components of DNA level changes, we identified MET as a potential ceRNA for EGFR, suggesting ceRNA as one such mechanism. For a certain subgroup of patients, EGFR and MET were upregulated while their shared regulating miRNA was downregulated. This would fit well with the ceRNA concept, leading to the hypothesis that EGFR CNV amplification “uses up” the regulatory miR-133b, which is then less likely to regulate MET so that EGFR indirectly upregulates MET. Since MET upregulation may be due to MET amplification, we also checked MET CNV values for both the amp and control groups. We found no MET amplification in these groups, confirming that the MET RNA upregulation was not due to DNA amplification.

While we have not biochemically confirmed MET and EGFR to be ceRNAs, EGFR-miR-133b-MET expression clustering could provide subgroups with significantly different survival rates. Since such survival rate difference was not found in groups considering only EGFR-MET expressions, identifying patients with ceRNA function was essential. On the other hand, an EGFR-MET ceRNA pair could have not been found without considering subgroups. Using our method of utilizing multiple-level data consisting of DNA copy number, mRNA expression, and miR expression together with biological information, we may find more clinically relevant potential ceRNA pairs as well as subgroups worthy of pursuit.

Our method can be automated by changing tree distance cutoff values (Pearson correlation distance) in identifying other ceRNAs and related subgroups, which can be validated with survival rates. However, overfitting using survival rate should not be done. Since we started from EGFR CNV-amplified patients, we hypothesized EGFR as the causal mRNA, fit well with ceRNA concept. This kind of biological knowledge is essential to our method.

Acknowledgments

We thank Ryan Ptashkin for the R script to generate a mature miRNA expression matrix file from individual isoform miRNAseq data; Axel Martin for the R script to make mRNA and miRNA matrix files in the same patient order; and Cristina Castillo and Andre Zapico for help in R. This project was extended from a 2015 computational biology summer camp for high school students supported by the University of Michigan WISE (Women in Science and Engineering).

References

1. Survival statistics for lung cancer | Cancer Research (2016) UK.Cancerresearchuk.org.
2. OMIM Entry Search - lung adenocarcinoma. (2016). Omim.org.
3. Cancer Genome Atlas Network. *Nature* **490**, 61–70 (2012).
4. L. Poliseno, *et al*. *Nature* **465**, 1033–1038 (2010).
5. M. S. Kumar, *et al*. *Nature* **505**, 212-217 (2013).
6. Y. Tay, *et al*. *Cell* **147**, 344–357 (2011).
7. F. A. Karreth, *et al*. *Cell* **147**, 382–395 (2011).
8. P. Sumazin, *et al*. *Cell* **147**, 370–381 (2011).
9. Y. C. Chiu, T. H. Hsiao, Y. Chen, E. Y. Chuang. *BMC Genomics* **16** Suppl 4, S1 (2015).
10. L. M. Wee, C. F. Flores-Jasso, W. E. Salomon, P. D. Zamore. *Cell* **151**, 1055–1067 (2012).
11. Y. Yuan, *et al*. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3158-3163 (2015).
12. Y. Tay, J. Rinn, P. P. Pandolfi. *Nature* **505**, 344-352 (2014).
13. C. Chou, *et al*. *Nucleic Acids Res*, **44** (D1), D239-D247 (2015).
14. Y. Zhu, P. Qiu, Y. Ji. *Nature Methods* **11**, 599–600 (2014).
15. A. Guo, *et al*. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 692-697 (2008).
16. N. Puri, R. Salgia. *J Carcinog*. **7**, 9 (2008).
17. M. Acunzo, *et al*. *Proc Natl Acad Sci U.S.A.* **110**, 8573-8578 (2013).
18. S. L. Organ, M. S. Tsao. *Ther Adv Med Oncol*. **3**, S7-S19 (2011).
19. J. A. Engelman, *et al*. *Science* **316,** 1039-1043 (2007).
20. https://clinicaltrials.gov/ct2/show/NCT01024413 (2016).
21. K. Nguyen, S. Kobayashi, D. Costa. *Clinical Lung Cancer*, **10**, 281-289 (2009).
22. http://www.genecards.org/cgi-bin/carddisp.pl?gene=MET (2016).

IMPROVED PERFORMANCE OF GENE SET ANALYSIS ON GENOME-WIDE TRANSCRIPTOMICS DATA WHEN USING GENE ACTIVITY STATE ESTIMATES

THOMAS KAMP

Department of Mathematics, Statistics, and Computer Science, Dordt College
Sioux Center, IA 51250, USA
Email: Thomas.Kamp@dordt.edu

MICAH ADAMS

Department of Mathematics, Statistics, and Computer Science, Dordt College
Sioux Center, IA 51250, USA
Email: Micah.Adams@dordt.edu

CRAIG DISSELKOEN

Department of Mathematics, Statistics, and Computer Science, Dordt College
Sioux Center, IA 51250, USA
Email: Craig.Disselkoen@dordt.edu

NATHAN TINTLE

Department of Mathematics, Statistics, and Computer Science, Dordt College
Sioux Center, IA 51250, USA
Email: Nathan.Tintle@dordt.edu

Gene set analysis methods continue to be a popular and powerful method of evaluating genome-wide transcriptomics data. These approach require *a priori* grouping of genes into biologically meaningful sets, and then conducting downstream analyses at the set (instead of gene) level of analysis. Gene set analysis methods have been shown to yield more powerful statistical conclusions than single-gene analyses due to both reduced multiple testing penalties and potentially larger observed effects due to the aggregation of effects across multiple genes in the set. Traditionally, gene set analysis methods have been applied directly to normalized, log-transformed, transcriptomics data. Recently, efforts have been made to transform transcriptomics data to scales yielding more biologically interpretable results. For example, recently proposed models transform log-transformed transcriptomics data to a confidence metric (ranging between 0 and 100%) that a gene is active (roughly speaking, that the gene product is part of an active cellular mechanism). In this manuscript, we demonstrate, on both real and simulated transcriptomics data, that tests for differential expression between sets of genes using are typically more powerful when using gene activity state estimates as opposed to log-transformed gene expression data. Our analysis suggests further exploration of techniques to transform transcriptomics data to meaningful quantities for improved downstream inference.

1. Introduction

Gene set analysis methods are a popular approach to assessing statistical significance on *a priori*, biologically defined sets of genes, as opposed to on a gene by gene basis [1]. These approaches have now been widely applied to SNP and RNA microarrays, and, more recently, RNA and DNA sequencing. The hope and promise of these methods is a combination of both statistical and biological improvements. Statistically, by analyzing sets of genes, instead of each gene individually, multiple testing penalties can be reduced. Furthermore, by potentially aggregating multiple independent effects (in different genes in the set), the true signal may more easily rise above the 'noise' of other genes in the set. Both reduced multiple testing penalties and aggregated effects have the potential to improve the statistical power of gene set tests. Biologically, by defining gene sets using *a priori* defined sets of genes, there is the increased potential for testing specific and more complex biological hypotheses (e.g., defining a set of genes as all genes in a pathway).

Previously, we discussed application of gene set analysis methods to testing for differential levels of gene expression in a genome-wide transcriptomics setting for bacteria [2]. In particular, we evaluated the performance of novel methods of testing for differential gene expression finding that the novel methods often outperformed, other popular methods, like Fisher's Exact Test (FET) [3]. These novel methods of testing for differential gene expression between two experiments (or bacterial strains) utilize the entire vector of normalized gene expression values for all genes in the set, instead of first defining an arbitrary cutoff (as is the case in FET). By leveraging the entire vector of expression values, instead of suffering from the information loss due to defining an arbitrary cutoff, the methods are generally more powerful than FET.

While gene set analysis typically focus on analyzing 'raw' gene expression data, many current approaches to understanding genome-wide transcriptomics data attempt to further leverage the data by classifying genes into one of two states: *active* (roughly speaking, the gene product is part of an active cellular mechanism) or *inactive* (the cellular mechanism is not active) [4]–[6]. We label this classification a determination of the *gene activity state.* Recently, we published a novel approach, *MultiMM* [7], to address documented deficiencies in many of the current state of the art methods. *MultiMM* is a parametric Bayesian mixture modelling approach which addresses limitations in existing methods as demonstrated through a rigorously grounded statistical framework, better performance than existing methods on simulated and real transcriptomics data, and through improved consistency with well-accepted biological realities and fluxomics data. Full details of, and links to, software for the *MultiMM* method are available elsewhere [7]. Ultimately, the *MultiMM* method yields a confidence estimate, $a_{ij} \in [0,1]$, that gene *i* is active in condition *j*. One stated goal of the

MultiMM method is to improve inference in downstream interpretations of gene expression data.

In this manuscript we consider the performance of a variety of gene set analysis methods on both raw gene expression data, as well as on a_{ij} values (confidence estimates that gene i, is active in experiment *j*) in order to determine if a_{ij} values are advantageous for use when conducting gene set analysis.

2. Methods

2.1. Methods of gene set testing

We consider three broad classes of gene set analysis methods [2], [3], [8].

First, we consider the burden test type of gene set testing method, with test statistic defined as:

$$B_m = \left|\sum_{i=1}^{k} e_{ij_1}^{m} - \sum_{i=1}^{k} e_{ij_2}^{m}\right|^{\frac{1}{m}} (1)$$

Where e_{ij} is the expression value of the i^{th} gene measured in the j^{th} condition, m is a positive constant (including infinity), and *k* is the number of genes in the set. As is discussed elsewhere [8], the Burden (B_m) test class of methods of conducting gene set analysis assumes that the effects of the genes within the test will tend to be in the same direction. For example, all genes in the set of interest are either not changing in underlying expression values, or are increasing, but none are decreasing. In the framework of 'activity states' this means that all genes are either moving from inactive to active (across the two experiments being compared) or are in the same state in both experiments. When this assumption is not met, Burden tests tend to be low powered since effects 'cancel out.' As m increases, increasing weight is put on the most expressed genes, such that if m=∞, $\sum_{i=1}^{k} e_{ij_1}^{m} = argmax(e_{ij_1})$.

The Variance Components class of test methods was envisioned primarily in response to the fact that Burden tests could not appropriately handle changes in multiple directions within the same set of genes (e.g., some genes move from inactive to active and others from active to inactive when comparing two experiments) [9]. The general form of a Variance Components gene set test statistic, VC_m, is given as:

$$VC_m = \left(\sum_{i=1}^{k} \left|e_{ij_1} - e_{ij_2}\right|^{m}\right)^{\frac{1}{m}}$$

Similar to the behavior for Burden tests, Variance components tests put increasing weight on pairwise differences in expression values as m increases, such that when m=∞, the VC statistic takes the value of the largest observed pairwise difference in expression values.

The third class of tests we considered was Fisher's Exact Test (FET). In this approach, an arbitrary cutoff, c, is first chosen, such that if $\left|e_{ij_1} - e_{ij_2}\right| > c$, then the gene is coded '1' (changing state; differentially expressed) and otherwise is coded '0' (not changing state; not differentially expressed). The proportion of genes in the set of interest which are deemed to be differentially expressed (>c) is compared to the proportion of genes not in the set of interest which are deemed to be differentially expressed using Fisher's Exact test, which uses a hypergeometric distribution to assess statistical significance.

2.2. Implementation of methods of gene set testing

In this manuscript we consider nine different tests, applied to both raw expression data (e_{ij}) and gene activity state estimates (a_{ij}; see next section for details). The nine tests are $B_1, B_2, B_\infty, VC_1, VC_2, VC_\infty, FET(1SD), FET(2SD)\ and\ FET\ (3SD)$. The test statistic equations for B and VC are given in the previous section, along with a description of the FET approach. For the FET approach, we use 1SD, 2SD and 3SD to denote how determine a cutoff value, *c.* In short, we find the average within gene SD across genes and experiments for which data is available, and then use that value (1SD), 2 times that value (2SD) or 3 times that value (3SD) to determine the cutoffs. For $e_{ij}\ 1SD = 0.75$ and, for a_{ij}, 1SD=0.3. FET determines statistical significance using the hypergeometric distributions. All other tests are evaluated for statistical significance by comparing the observed statistic to a null distribution of 10,000 randomly generated statistics obtained by randomly choosing 10,000 sets of the same size as the gene set being evaluated and finding the fraction of randomly chosen sets with larger statistics than observed (the p-value).

2.3. Moving from raw expression values to estimates of gene activity states

The *MultiMM* algorithm takes as input a genome-wide matrix of transcriptomics data ***E*** across numerous experimental conditions, such that the entries in ***E*** are denoted e_{ij} and represent the estimated gene expression of gene *i* in condition *j*. Additionally, if available, *MultiMM* allows for *a priori* identification of sets of genes which are known to be co-regulated such that in the same experimental condition, the co-regulated genes are all active or all inactive. The *MultiMM* algorithm starts by using the Bayesian Information Criterion (*BIC*) to assess the fit of a 1-component (univariate or multivariate) Gaussian mixture distribution (gene is always active or inactive in the set of conditions represented) vs. a 2-component mixture distribution (gene is sometimes active and sometime inactive in the set of conditions represented) using the *R* package *Mclust* [10]. Following Raftery et al. [11] we require the BIC to be at least 12 points better for the 1-component model to be chosen vs. the 2-component model. Second, for all genes estimated to come from a 2-component mixture distribution, a Gaussian mixture model is fit and a Gibbs sampler is used in order to yield estimates of the means and standard deviations of the components of the mixture model, along with an estimate of the proportion of experiments for which the gene is active. In the case of co-regulated sets of genes this mixture model is multivariate, whereas for genes that are not known to be co-regulated with other genes, the mixture model is univariate. Finally, the estimated mixture distribution parameters can be used to yield a confidence estimate, $a_{ij} \in [0,1]$, that gene i is active in condition j. For genes inferred as being always active or always inactive in the dataset in step one of the algorithm, multiple imputation is used to impute a_{ij} values. Full details of, and links to, software for the *MultiMM* method are available elsewhere [7].

2.4. Simulation of gene expression data

We simulated expression data with 'known' gene activity states (active/inactive). The simulation of expression data was informed by the *E. coli* expression data described later. We first ran the Screening Method described above (BIC with MClust) and dropped all operons (co-regulated gene sets), including single gene operons, for which the two-component model did not yield the highest BIC (n=697 dropped). We then randomly selected 26.3% (=697/2648) of the remaining 1951 operons to be single component in the simulated data, with each of the single component operons having an equal likelihood of being always active or always inactive.

To calculate the mixing parameter, π, used in the simulation for the 1438 two-component operons we chose a random value for π between 0.2 and 0.8. Values for $\vec{\mu}_0, \vec{\mu}_1, \Sigma_0$=$\Sigma_1$ are all as estimated by the *MultiMM* method computed on the real expression data. To generate simulated expression values, ϵ_{ij}^{s}, we drew $907(\pi_i)$ random values from a multivariate normal distribution $(\vec{\mu}_{1i}, \Sigma_{1i})$ and $907(1-\pi_i)$ random values from a multivariate normal distribution $(\vec{\mu}_{0i}, \Sigma_{0i})$. Thus, we generated a 907 by 3435 matrix of ϵ_{ij}^{s} values. Prior analysis has shown this simulated data to have good properties and behave in reasonable ways [7].

2.5. Simulation of gene sets for analysis

We used the simulated gene expression data described above to generate random sets of genes for evaluation of different methods of gene set analysis. We selected random sets of 8, 20 or 40 genes from among genes which were not changing or changing states between the two experiments of interest. In particular, we looked at the following proportions of genes in the set which were not changing state (0, 25, 50, 75 and 100%), and either 0%, 50% or 100% of the genes in the set active in the first experiment. Thus, we explored 45 simulation settings (3 (set size) by 5 (not changing) by 3 (starting state). Of these 45 simulation settings, 9 represent settings for which we can evaluate the empirical type I rate and 36 will be used to evaluate statistical power. Each of the nine test statistics is computed for the set, and then each of the nine statistics is compared to a distribution of the same statistic across 10,000 randomly selected sets of the same size (an approach termed 'gene sampling' which uses a 'competitive null hypothesis'[12]). We considered 1000 randomly selected sets at each of the 45 simulation settings. Full simulation results are available in Supplemental File #1. We also analyzed 574 *a priori* defined operon (co-regulated) sets based on operon definitions for *E. coli* as provided by Microbes Online [13]. Full results are available in Supplemental File #2. Supplemental Files are available at: http://homepages.dordt.edu/ntintle/gsa_supp.zip

2.6. Real data

We also used genome-wide gene expression data from 907 different microarray data sets collected on 4329 *Escherichia coli* genes via the M3D data repository [14]–[16] both to inform simulated data analysis and when considering the actual performance of the methods. Raw data from Affymetrix [17] CEL files were normalized using RMA [18]. Details of data processing are described elsewhere [19], [20].

2.7. Statistical analysis

Empirical power and type I error rate estimates are computed as the proportion of times that the p-value was less than the significance level for a particular test and simulation setting. We considered significance levels of 5%, 0.5% and 0.05%.

Results

Across 36 simulation settings where at least one gene in the set changed activity states, power was consistently better when using gene activity state estimates, than raw expression data (see Table 1 for overall summary). Across the 9 simulation settings where none of the genes in the set changed state (type I error setting), the Type I error rate was generally controlled for all methods (detailed results not shown). Table 1 shows that gains in power can be high across all methods, whereas when power is worse when using activity states, the reduction in power is usually quite minimal (19 to 82 average percentage point increase vs. 0.3 to 2.3 average percentage point decrease).

Table 1. Power improvements comparing raw expression data to gene activity state estimates using a variety of gene set analysis approaches

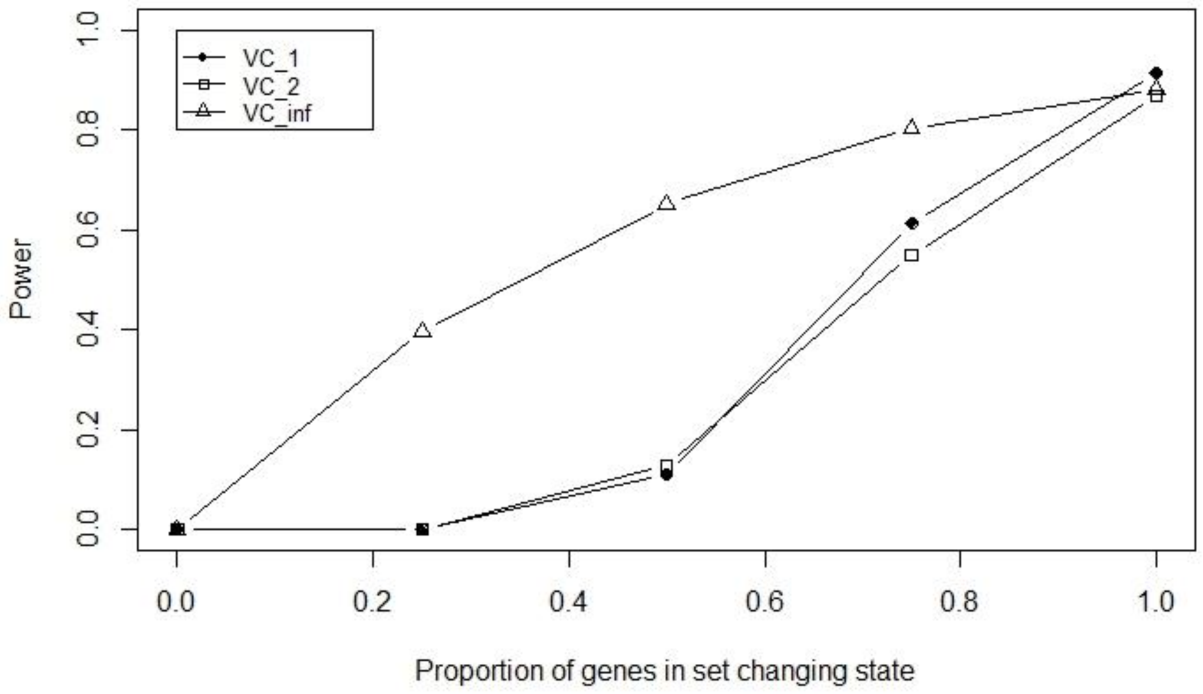
Gene set analysis approach		Proportion of 36 simulation settings where power is better using a_{ij}	Average (SD) power gain when power is better using a_{ij}[1]	Proportion of 36 simulation settings where power is the same using a_{ij}	Proportion of 36 simulation settings where power is worse using a_{ij}	Average (SD) power loss when power is worse using a_{ij}[2]
Fisher's exact test	Cutoff=3SD	73.1%	24.9% (21.8%)	17.6%	9.3%	0.3% (0.2%)
	Cutoff=2SD	63.9%	28.4% (20.5%)	16.7%	19.4%	0.7% (0.9%)
	Cutoff=1SD	66.7%	25.2% (21.0%	16.7%	16.7%	0.9% (0.8%)
Burden	m=1	48.1%	19.1% (16.4%)	29.6%	22.2%	1.4% (1.1%)
	m=2	46.3%	22.1% (17.7%)	25.0%	28.7%	1.2% (1.3%)
	m=∞	55.6%	39.0% (29.2%)	10.2%	34.3%	2.3% (2.9%)
Variance components	m=1	61.1%	28.9% (20.6%)	19.4%	19.4%	0.8% (0.6%)
	m=2	64.8%	40.4% (28.0%)	10.2%	25.0%	1.0% (1.1%)
	m=∞	100%	82.2% (17.4%)	0	0	-

1. In situations when the power is better using a_{ij} vs. $e_{ij,}$ what is the difference in powe$_r$ estimates between the two different methods. For example, for VC_∞ the difference power between using a_{ij} and e_{ij} averaged 82.2% percentage points, reflecting the fact that VC_∞ is substantially better when using a_{ij}
2. In situations when the power is worse using a_{ij} vs. $e_{ij,}$ what is the difference in powe$_r$ estimates between the two different methods. For example, for B_∞ the difference power between using a_{ij} and e_{ij} averaged 2.3% percentage points, reflecting the fact that B_∞ is not much worse using a_{ij} and e_{ij} in the 34.3% of cases when it is worse

For each of the thirty-six simulation settings used to estimate power, the power was always highest across all 18 methods (nine different test statistics using either e_{ij} or a_{ij}) for a method using gene activity state estimates. This was true for each of the 3 different significant levels. VC_∞ was frequently the most powerful approach (16 out of 36 times for significance level 5%; 26 out of 36 times for significance level 0.5% and 33 times for significance level 0.05%). While other B and VC methods were periodically most powerful, notably, the FET methods were never the most powerful, even when using gene activity state estimates (a_{ij}).

Figure 1 illustrates typical performance of the VC methods as the proportion of genes in the set changes, by highlighting the performance of the methods on sets of size 8. VC_∞ is most robust to lower proportions of genes in the set changing state, while all methods perform well when the proportion of genes in the set changing state is relatively large.

Figure 1. Power of different VC tests as the proportion of genes in the set changing state varies



Analysis of the 574 real, operon based sets of genes showed similar performance to the randomly generated gene sets, with even better performance of the activity state informed methods in many cases (detailed results not shown).

Real data example

The L-arabinose (*ara*) operon is a well-studied set of three co-located genes (*araB, araA, araD*) which encode enzymes needed for the catabolism of arabinose in *E. coli* [52]. Across the 907 experiments in our dataset, L-arabinose is present in the media in 227 cases. We randomly selected 1000 pairs of experiments where one experiment had L-arabinose present in the media and one experiment did not. We then computed different gene set analysis test statistics for the L-arabinose operon using both raw expression data and activity state estimates, as compared to 100,000 randomly selected sets of 3 genes. Table 2 illustrates that methods using activity state estimates were always more powerful than methods which were based on raw expression values.

Table 2. Empirical power estimates for detecting significant changes in activity for the L-arabinose operon in *E. coli* when comparing an experiment with L-arabinose present in the media vs. one without

Sig. Level	Method	B_1	B_2	B_∞	VC_1	VC_2	VC_∞
0.05%	Raw expression (e_{ij})	96.6%	98.1%	1.6%	95.7%	52.3%	1.7%
	Activity state estimates (a_{ij})	100%	100%	99.6%	100%	100%	99.6%
0.005%	Raw expression (e_{ij})	85.3%	86.1%	0%	58.0%	3.9%	0%
	Activity state estimates (a_{ij})	99.6%	99.6%	99.6%	99.6%	99.6%	99.6%

4. Discussion

Gene set analysis remains a statistically promising and biological relevant approach to the analysis of genome-wide transcriptomics data. Here we demonstrate that, in line with previous work [2], methods which don't arbitrarily introduce a cutoff and lose information, are generally more powerful than methods that do (e.g., Fisher's exact test). We also demonstrate that using a more statistically grounded metric to quantify gene expression (activity state estimates, a_{ij}) generally leads to more powerful tests than using raw gene expression data (e_{ij}) on simulated data, with promising results also observed on real data in well-understood biological systems.

We note that the VC_∞ method performed particularly well, especially at low significance thresholds. This finding reflects the use of gene-sampling (a competitive null hypothesis). Briefly, when using gene sampling to assess statistical significance, test statistics generated for the gene set of interest, are compared to randomly chosen gene sets. The VC_∞ method performs relatively better as compared to other methods as the significance level decreases because it is focused on the most extreme observed difference in activity state estimates and, thus, is more robust than other methods to small numbers of randomly selected sets of genes with extreme values of the test statistic. This performance was particularly notable in the example with the L-arabinose operon, where the VC_∞ method using activity state estimates (a_{ij}) outperformed its performance on raw expression values (e_{ij}) by nearly 100%. While other test statistics did not show as large of a difference, in all cases the power was higher when using activity state estimates. Thus, when attempting to determine if sets of genes are differentially active in two conditions, inferring gene activity state estimates prior to applying gene set analysis methods will maximize the likelihood of identifying differential activity. In short, use of these methods will maximize our ability to identify sets of genes associated with differential activity between two conditions.

We note numerous opportunities for future work, including (1) the ability to expand these methods to incorporate information from multiple, similar experimental conditions, instead of only comparing two conditions, (2) integrating directionality and/or gene set topology, (3) potential improvements by further leveraging the statistical properties of well-calibrated a_{ij} (the posterior likelihood that gene *i* is active in gene *j*), (4) potential further improvements in power by using non-competitive null hypotheses, which may be possible through statistical quantification of the null distributions of particular methods when using well-calibrated a_{ij}'s and (5) use of this general framework to test for whether a set of genes in a single experiment shows evidence of significant 'activity' (vs. only a change in activity levels between two experiments, as we considered here).

The most notable limitation of our analysis here is the limited application to real data, though initial results are promising and performance on real (operon-based sets) was also quite encouraging. Further work is necessary to ensure transferability of these promising initial findings to additional organisms. For example, to determine if these methods will successfully distinguish sets of differentially active genes between diseased and non-diseased tissue. Furthermore, further work is necessary to explore validation in other well-understood biological systems and as compared to the results of other –omics data (e.g., genome-scale metabolic models; fluxomics, etc.).

Acknowledgments

This work is supported by NSF MCB-1330734. We gratefully acknowledge the use of the Silicon Mechanics grant funded beaker computer cluster on the campus of Dordt College for computations.

References

[1] C. de Leeuw, B. M. Neale, T. Heskes, and D. Posthuma, "The statistical properties of gene-set analysis," *Nat. Rev. Genet.*, vol. 17, pp. 353–364, 2016.

[2] N. L. Tintle, A. A. Best, M. DeJongh, D. Van Bruggen, F. Heffron, S. Porwollik, and R. C. Taylor, "Gene set analyses for interpreting microarray experiments on prokaryotic organisms.," *BMC Bioinformatics*, vol. 9, no. 1, p. 469, 2008.

[3] P. Khatri and S. Drăghici, "Ontological analysis of gene expression data: Current tools, limitations, and open problems," *Bioinformatics*, vol. 21, no. 18, pp. 3587–3595, 2005.

[4] S. Abel, T. Bucher, M. Nicollier, I. Hug, V. Kaever, P. Abel zur Wiesch, and U. Jenal, "Bi-modal Distribution of the Second Messenger c-di-GMP Controls Cell Fate and Asymmetry during the Caulobacter Cell Cycle," *PLoS Genet.*, vol. 9, no. 9, p. e1003744, 2013.

[5] C. A. Gallo, R. L. Cecchini, J. A. Carballido, S. Micheletto, and I. Ponzoni, "Discretization of gene expression data revised," *Brief. Bioinform.*, no. May, pp. 1–13, 2015.

[6] J. E. Ferrell, "Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability," *Curr. Opin. Cell Biol.*, vol. 14, no. 2, pp. 140–148, 2002.

[7] C. Disselkoen, B. Greco, K. Cook, K. Koch, R. Lerebours, C. Viss, J. Cape, E. Held, Y. Ashenafi, K. Fischer, A. Acosta, M. Cunningham, A. A. Best, M. DeJongh, and N. L. Tintle, "A Bayesian framework for the classification of microbial gene activity states," *Front. Microbiol.*, vol. 7, no. 1191, 2016.

[8] K. Liu, S. Fast, M. Zawistowski, and N. L. Tintle, "A geometric framework for evaluating rare variant tests of association," *Genet. Epidemiolgoy*, vol. 37, no. 4, pp. 712–722, 2013.

[9] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test.," *Am. J. Hum. Genet.*, vol. 89, no. 1, pp. 82–93, Jul. 2011.

[10] C. Fraley, A. Raftery, L. Scurcca, T. B. Murphy, and M. Fop, "mclust: Normal mixutre modelling for model-based clustering, classification and density estimation," *CRAN*, 2015. [Online]. Available: https://cran.r-project.org/web/packages/mclust/index.html.

[11] A. Raftery, "Bayesian model selection in social research," *Sociol. Methods*, vol. 25, pp. 111–163, 1995.

[12] J. Goeman and P. Buhlmann, "Analyzing gene expression data in terms of gene sets:

methodological issues," *Bioinformatics*, vol. 23, no. 8, pp. 980–987, 2007.

[13] M. N. Price, K. H. Huang, E. J. Alm, and A. P. Arkin, "A novel method for accurate operon predictions in all sequenced prokaryotes.," *Nucleic Acids Res.*, vol. 33, no. 3, pp. 880–92, Jan. 2005.

[14] "Many Microbes Database." [Online]. Available: http://m3d.mssm.edu.

[15] J. J. Faith, M. E. Driscoll, V. a Fusaro, E. J. Cosgrove, B. Hayete, F. S. Juhn, S. J. Schneider, and T. S. Gardner, "Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata.," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D866–70, 2008.

[16] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.," *PLoS Biol.*, vol. 5, no. 1, p. e8, Jan. 2007.

[17] "Affymetrix." [Online]. Available: http://www.affymetrix.com.

[18] T. Irizarry, R. a., Bolstad, Benjamin, Collin, Francois, Cope, Leslie, Hobbs, Bridget, Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Res.*, vol. 31, no. 4, p. 15e–15, Feb. 2003.

[19] S. Powers, M. DeJongh, A. a Best, and N. L. Tintle, "Cautions about the reliability of pairwise gene correlations based on expression data.," *Front. Microbiol.*, vol. 6, no. June, p. 650, Jan. 2015.

[20] N. Tintle, A. Sitarik, B. Boerema, K. Young, A. Best, and M. De Jongh, "Evaluating the consistency of gene sets used in the analysis of bacterial gene expression data.," *BMC Bioinformatics*, vol. 13, no. 1, p. 193, Jan. 2012.

methylDMV: SIMULTANEOUS DETECTION OF DIFFERENTIAL DNA METHYLATION AND VARIABILITY WITH CONFOUNDER ADJUSTMENT

PEI FEN KUAN*, JUNYAN SONG and SHUYAO HE

Department of Applied Mathematics and Statistics,
Stony Brook University,
Stony Brook, NY 11794, USA
**E-mail: peifen.kuan@stonybrook.edu*
http://www.stonybrook.edu/commcms/ams2/

DNA methylation has emerged as promising epigenetic markers for disease diagnosis. Both the differential mean (DM) and differential variability (DV) in methylation have been shown to contribute to transcriptional aberration and disease pathogenesis. The presence of confounding factors in large scale EWAS may affect the methylation values and hamper accurate marker discovery. In this paper, we propose a flexible framework called `methylDMV` which allows for confounding factors adjustment and enables simultaneous characterization and identification of CpGs exhibiting DM only, DV only and both DM and DV. The proposed framework also allows for prioritization and selection of candidate features to be included in the prediction algorithm. We illustrate the utility of `methylDMV` in several TCGA datasets. An R package `methylDMV` implementing our proposed method is available at `http://www.ams.sunysb.edu/~pfkuan/softwares.html#methylDMV`.

Keywords: DNA methylation; Differential variability; Feature selection; Elastic net.

1. Introduction

DNA methylation is an important hallmark of genomic imprinting, transcriptional regulation, X-inactivation and chromosomal stability.[1] The most common DNA methylation process in human involves the addition of a methyl group to the 5-carbon of the cytosine ring. In human, this modification mostly occurs at a CpG site in which a cytosine nucleotide is followed by a guanine nucleotide. Aberrant patterns of DNA methylation have been shown to be a critical mechanism in the development and progression of various diseases, in particular cancer.[2] DNA methylation is one of the most widely studied epigenetics event and has been profiled extensively in large consortiums including the Cancer Genome Altas (TCGA), NIH Roadmap and the Encyclopedia of DNA Elements (ENCODE) projects. These efforts provide research opportunities for secondary analyses of the large datasets to further understand the biology of the disease.

Most of the work in DNA methylation have been focused on identifying DNA methylation markers that exhibit differential average or mean methylation (DM).[3,4] These epigenetic markers have been shown to be promising biomarkers in designing platform for disease diagnosis.[5] Over the last few years, there has been an increasing interest in identifying DNA methylation markers that exhibit differential variability in various diseases, including cancer[6–8] and obesity.[9] These epigenetic variabilities can be attributed to increased plasticity arising from changing environment including varying oxygen tension[10] and is associated with the risk of morphological and neoplastic transformation.[11] These studies opened up new avenues to the study of DNA methylation, which indicated that simultanenous investigation of both differential mean and variability may delineate the complex patterns of epigenetic regulation in pathophysiology and development of diseases.

One of the most widely used DNA methylation platforms is the Illumina Infinium HumanMethylation450 BeadChip which profiles more than 450,000 CpGs genome wide. The latest phase of the Illumina methylation array is the MethylationEPIC BeadChip which covers approximately 850,000 methylation sites including CpG islands, enhancers and regulatory regions identified from the ENCODE project. The methylation value for each CpG is represented as a *beta* (β) value, which is the ratio of methylated probe intensities to the total probe intensities, where $0 \leq \beta \leq 1$; $\beta = 0$ and $\beta = 1$ indicate that the CpG is fully unmethylated and methylated, respectively.

An important aspect of differential methylation analysis is to identify CpGs which exhibit differential mean or variance in large scale hypothesis testing. Statistical tests for detecting CpGs which exhibit differential mean methylation include t-tests, non-parametric Wilcoxon rank sum test or limma[12] based on linear models and empirical Bayes approach. On the other hand, several algorithms have been proposed in recent years to identify CpGs which exhibit differential variability in large scale hypothesis testing. For instance, Teschendorff et al. (2012)[8] proposed a regularized version of the Bartlett's test, Ahn et al. (2013)[13] used a score test from generalized regression model, Phipson et al. (2014)[14] proposed a modification of Levene's test, Wahl et al. (2014)[15] introduced a generalized additive models for location, scale and shape (GAMLSS) framework and Kuan (2014)[16] proposed a general linear model with propensity score method for detecting CpGs with differential variability.

CpGs which exhibit differential mean methylation have been utilized in classification algorithm to define methylation signatures for disease subtypes.[17,18] As the methylation arrays encompass $> 450,0000$ CpGs, a common approach in training the classification algorithm is to pre-select features ranked highly by the univariate differential mean methylation as candidate CpGs in the classification algorithm to improve the stability of the algorithm. Motivated by the biological insights of differential variability in methylation, Teschendorff et al. (2012)[8] proposed a method which selected differential variable CpGs using Bartlett's test for inclusion in the prediction algorithm.

Large scale differential methylation analysis requires proper adjustment for confounders to reduce the biases associated with the identified methylation markers. For instance, age[19,20] and cigarette smoking[21,22] have been shown to be associated with DNA methylation; thus in studies to identify methylation markers for cancer or other disease phenotypes, appropriate adjustment for these factors is necessary. In the analysis of differential mean methylation, this can be achieved via a regression framework where confounders are included as covariates in the model. However, in the analysis of differential variability, potential biases due to confounding variables are usually ignored.[8,14]

This paper aims to develop a unified framework to address the limitation of existing work: (1) incorporates adjustment for confounding variables that potentially affect methylation levels, and allows for simultaneous detection of differential mean (DM) and differential variability (DV) in methylation analysis, (2) systematic selection of CpGs which exhibit differential mean and/or differential variability in the prediction algorithm to improve prediction accuracy and biological interpretation. In Section 2, we describe our proposed approach. This is followed by simulation studies and real data applications in Sections 3 and 4, respectively. The paper concludes with a discussion in Section 5.

2. Methods

2.1. *A framework for simultaneous detection of differential mean (DM) and differential variability (DV)*

Without loss of generality, we describe our proposed framework for detecting differential mean and differential variability between two conditions or groups (e.g., tumor versus normal). A common distribution to model the *beta* values from Illumina methylation arrays is the beta distribution.[23] Since the variance of a beta distribution is a function of the mean, the β values exhibit significant heteroscedasticity.[24] To overcome the heteroscedasticity issue, we consider a variance stabilizing transformation via the logit function to the β values, i.e., $logit(\beta) = \log[\beta/(1-\beta)]$. Let x_{ij} denote the logit transformed methylation value for sample i and CpG j. We first define a deviation measure $r_{ij} = |x_{ij} - \text{wt.med}_i(x_{ij})|$ where $\text{wt.med}_i(x_{ij})$ is the weighted median of CpG j with weights $w_i = 1/2n_{g_i}$, $g_i = 0$ if sample i is a control and $g_i = 1$ if sample i is a case, and n_0 and n_1 are the respective sample sizes.

We recast the model for simultaneous detection of differential mean and differential variable CpGs using a logistic regression model. Let y_i denote the group membership of sample i, where $y_i = 0$ if the sample is a control/normal and $y_i = 1$ if the sample is a case/tumor. y_i is assumed to follow a binomial distribution with $P(y_i = 1) = \pi_i$ and $\log[\pi_i/(1-\pi_i)] = \theta_i$. We consider the four competing models for each CpG:

Model 1: $\theta_i = \beta_0 + \sum_{k=1}^{K} \gamma_k Z_{ik}$ (no DM or DV)
Model 2: $\theta_i = \beta_0 + \beta_m x_{ij} + \sum_{k=1}^{K} \gamma_k Z_{ik}$ (DM only)
Model 3: $\theta_i = \beta_0 + \beta_v r_{ij} + \sum_{k=1}^{K} \gamma_k Z_{ik}$ (DV only)
Model 4: $\theta_i = \beta_0 + \beta_m x_{ij} + \beta_v r_{ij} + \sum_{k=1}^{K} \gamma_k Z_{ik}$ (both DM and DV)

In all models, $\mathbf{Z}_k = (Z_{ik})'$ corresponds to confounding variable k, for instance age, smoking status or alcohol consumption. Model 1 is the baseline model which adjusts for confounding variables and assumes that the phenotype is not associated with differential mean (DM) or differential variability (DV). Model 2 (Model 3) assumes that the phenotype is associated with DM (DV) after adjusting for confounders, whereas Model 4 assumes that the phenotype is associated with both DM and DV for a CpG. To identify CpGs which exhibit DM, one can compare Model 1 to Model 2 using likelihood ratio tests or score tests.[25] On the other hand, Model 3 can be compared to Model 1 to obtain p-values associated with DV for each CpG. The comparison of Model 4 and Model 1 identifies CpGs which exhibit either DM or DV. The vector of p-values from each analysis are adjusted via the false discovery rate (FDR)[26] to account for multiple testings. In addition to large scale hypothesis testing framework to identify DM and DV CpGs, another advantage of our proposed model is that it allows for automatic classification of the CpGs into the four classes (1) no DM or DV, (2) DM only, (3) DV only and (4) both DM and DV. This is carried out via a Bayesian Information Criterion

(BIC) to rank the four models for each CpG, i.e., the CpG is categorized into the class with the smallest BIC score.

2.2. *Candidate feature selection for prediction modeling*

The BIC used for model ranking within each CpG can also be utilized to aid candidate feature selection to improve the stability of the prediction algorithm. The proposed framework provides flexibility to the user for including top ranking features in constructing prediction model. For instance, if the user is interested in a prediction model using CpGs which exhibit the largest discriminative power in terms of both DV and DM after adjustment for confounding variables, then the subset of CpGs which show the lowest BIC scores for Model 4 are selected as candidate features. On the other hand, if the user is interested in a prediction model using only DM CpGs , then the candidate features correspond to the CpGs which identify Model 2 as the best model using BIC scores.

The selected candidate features are used in the prediction algorithm for constructing classification rule discriminating case from control. In this paper, we consider the elastic net algorithm.[27] The objective function of elastic net consists of a loss function + penalty:

$$\min_{\beta} ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda \left\{\alpha||\beta||_1 + (1-\alpha)||\beta||^2\right\}$$

where $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ and $||\beta||^2 = \sum_{j=1}^{p} \beta_j^2$. The parameters λ and α are tuned via cross-validation. Other types of machine learning prediction algorithm can also be used on the selected candidate features, for instance the random forest[28] which is a non-parametric ensemble approach based on a large number of classification trees trained on bootstrap samples.

An R package `methylDMV` implementing our proposed method for testing DM and DV, as well as CpGs ranking by BIC and candidate feature selection is available at `http://www.ams.sunysb.edu/~pfkuan/softwares.html#methylDMV`.

3. Simulation studies

We carried out simulation studies to evaluate the effect of confounders on CpG ranking. Specifically, denote Z_{i1} and Z_{i2} as the two confounders, where $Z_{i1} \sim N(0,1)$ and $Z_{i2} \sim \text{Bernoulli}(0.6)$ for sample i, $i = 1, 2, \ldots, n$. The group indicator y_i was generated from the following model

$$\text{logit}(p_i) = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i}$$
$$y_i \sim \text{Bernoulli}(p_i)$$

For each CpG j ($j = 1, 2, \ldots, p$), the measurements x_{ij}'s were generated from the Gaussian distribution under the assumption that the beta values have been properly transformed (e.g., logit or arcsine transformation), i.e., $x_{ij} \sim N(\mu_{ij}, \sigma_{ij}^2)$ where

(i) $\mu_{ij} = \mu_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}$ and $\sigma_{ij}^2 = \sigma_0^2$ if CpG j is from Model 1 (no DM or DV)
(ii) $\mu_{ij} = \mu_0 + \alpha_g y_i + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}$ and $\sigma_{ij}^2 = \sigma_0^2$ if CpG j is from Model 2 (DM only)
(iii) $\mu_{ij} = \mu_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}$ and $\sigma_{ij}^2 = \sigma_0^2 + \beta_g y_i$ if CpG j is from Model 3 (DV only)
(iv) $\mu_{ij} = \mu_0 + \alpha_g y_i + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}$ and $\sigma_{ij}^2 = \sigma_0^2 + \beta_g y_i$ if CpG j is from Model 4 (both DM and DV)

The proportion of CpGs from Models 1-4 were drawn from a multinomial distribution with $\pi = \left(\pi_1, \frac{1-\pi_1}{3}, \frac{1-\pi_1}{3}, \frac{1-\pi_1}{3}\right)$. We set $\gamma_0 = 1, \gamma_1 = 2, \gamma_2 = -2$ to obtain approximately equal number of cases and controls; and $\alpha_g = 1, \beta_g = 1, \mu_0 = 0, \sigma_0^2 = 1$. We varied $\alpha_1 = \alpha_2 = 0, 0.5, 1, 3, 5$ to reflect the different degrees of confounding in the methylation measurements and $\pi_1 = 0.4, 0.6, 0.8$ for the different mixing proportions of DM and DV CpGs. To evaluate the effect of confounders on the phenotype, i.e., case/control, we also considered the case in which the y_i's were not affected by confounders. Under this scenario, $y_i = 0$ for $i = 1, 2, \ldots, n/2$ and $y_i = 1$ for $i = n/2 + 1, \ldots, n$. For each scenario, the simulation was conducted for $n = 200$ samples and $p = 10000$ CpGs over 100 iterations.

We compared the average accuracy of the BIC ranking procedure in classifying the CpGs into Models 1-4 with (BICadj) and without (BICnoadj) adjustment for confounders. We also included comparison to method which performed tests for DM and DV separately. Two sample t-test and Levene's test were used to identify DM and DV CpGs, respectively. CpG j was classified as DM (DV) if the p-value from t-test (Levene's test) adjusted via the Benjamini-Hochberg procedure[26] $\leq$ FDR. We considered FDR 0.05 and 0.1, and referred to this method as SepTest0.05 and SepTest0.1, respectively.

Figure 1 summarizes the average accuracy for the four methods across the different settings. In scenarios where both the phenotype (case/control status) and methylation measurements were affected by confounders (top row of Figure 1 for $\alpha_1 \neq 0$), the methods which did not adjust for confounders exhibited poor accuracy across different mixing proportions π_1. For the case where $\alpha_1 = 0$, i.e., methylation measurements were not affected by confounders, the BICadj method showed a slight decrease in accuracy compared to other methods. Bottom row of Figure 1 displays the results for the scenarios where only the methylation measurements were confounded while the phenotype was not affected by confounders. For these cases, the performance of the methods were comparable for $\alpha_1 \leq 1$. The advantages of adjusting for confounders were apparent for $\alpha_1 = 3, 5$, i.e., strong confounding effect in the methylation measurements even in the absence of confounding in case/control status.

4. Case studies

4.1. *Data preprocessing and normalization*

We illustrated our proposed method, `methylDMV` on three datasets, namely the breast cancer (BRCA), kidney cancer (KIRC) and liver cancer (LIHC) dataset. The breast cancer dataset consisted of 909 samples downloaded from the TCGA data portal and the NCBI gene expression omnibus under accession number GSE67919, whereas the kidney and liver cancer consisted of 475 and 404 samples from the TCGA data portal, respectively. All the samples were profiled using the Illumina Infinium HumanMethylation450 BeadChip.

Preprocessing of the methylation data at the 485,557 CpGs were performed as follows. Probes with detection p-value > 0.05 were set to missing and probes with more than 20% missing were filtered. A beta mixture quantile (BMIQ) normalization[29] was applied to the beta values for correction of bias due to the type I and type II probes. Non-specific, cross-hybridized probes,[30,31] probes overlapping with a SNP and probes mapping to repeat regions were filtered. For KIRC and LIHC, we further filtered for CpGs mapping to chromosomes X

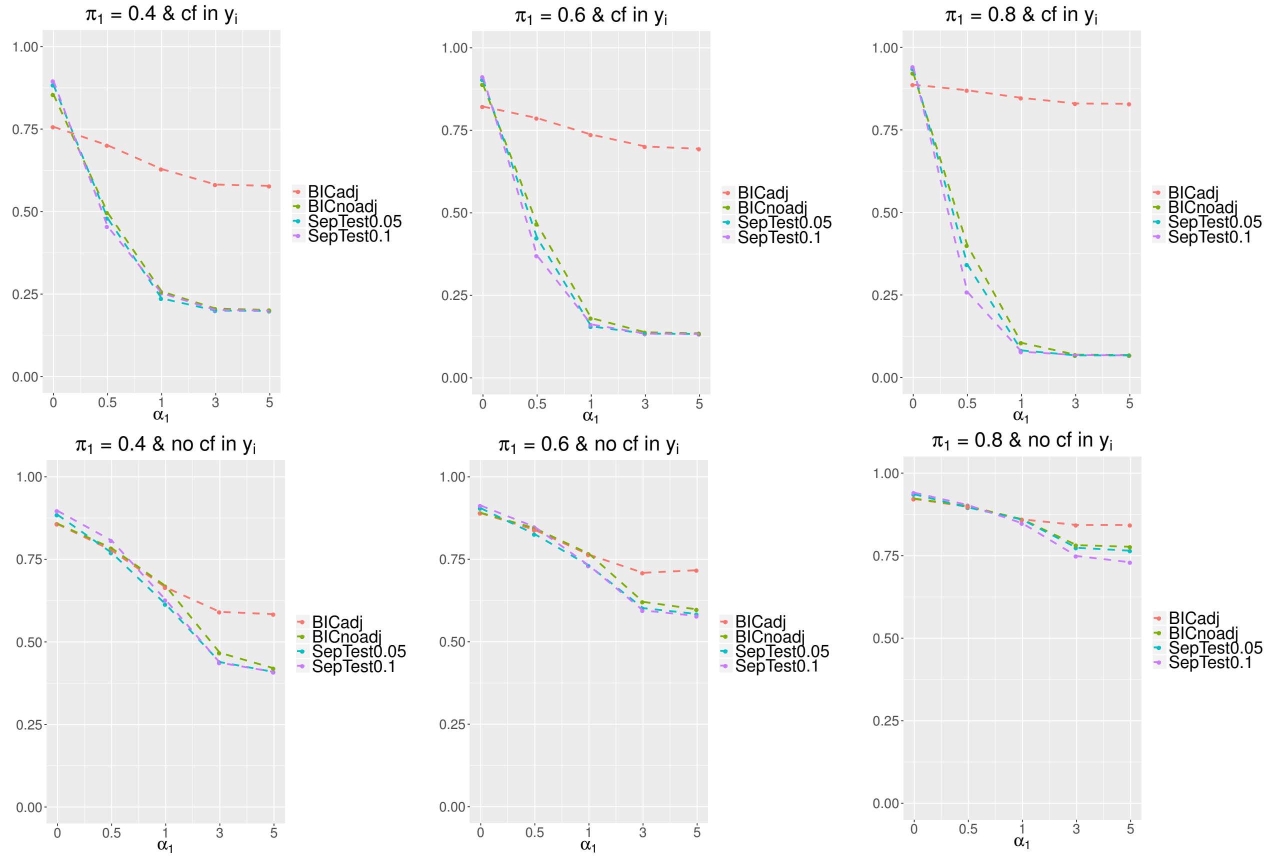


Fig. 1. Average accuracy of CpG classification across α_1's for our proposed BIC ranking with confounding adjustment (BICadj, orange), BIC ranking without confounding adjustment (BICnoadj, green), separate two-sample t-test and Levene's test for DM and DV at FDR 0.05 (SepTest0.05, turquoise) and 0.1 (SepTest0.1, purple). Each panel corresponds to a specific π_1 value and whether the case control status was affected by confounders (top row: $y_i \sim$ Bernoulli(p_i), i.e, affected by confounders; bottom row $y_i = 0, i = 1, \ldots n/2$, $y_i = 1, i = n/2+1, \ldots n$, i.e, not affected by confounders).

and Y. The normalized datasets consisted of 374,680, 365,896 and 365,658 CpGs for BRCA, KIRC and LIHC, respectively. We performed the following pairwise comparisons:

(i) **KIRC (tumor vs normal)**: Models 1-4 were fitted on $n_0 = 156$ normal (control) and $n_1 = 319$ tumor (case), adjusting for age and race.

(ii) **LIHC (tumor vs normal)**: Models 1-4 were fitted on $n_0 = 47$ normal (control) and $n_1 = 357$ tumor (case), adjusting for age and race.

(iii) **BRCA (tumor vs normal)**: Models 1-4 were fitted on $n_0 = 180$ normal (control) and $n_1 = 729$ tumor (case), adjusting for age and race.

(iv) **BRCA (basal vs luminal A)**: Models 1-4 were fitted on $n_0 = 93$ luminal A (control) and $n_1 = 30$ basal (case), adjusting for age and race.

(v) **BRCA (basal vs luminal B)**: Models 1-4 were fitted on $n_0 = 40$ luminal B (control) and $n_1 = 30$ basal (case), adjusting for age and race.

(vi) **BRCA (luminal B vs luminal A)**: Models 1-4 were fitted on $n_0 = 93$ luminal A (control) and $n_1 = 40$ luminal B (case), adjusting for age and race.

4.2. *Feature ranking by BIC scores*

In tumor versus normal comparison within KIRC, LIHC and BRCA datasets, majority of the CpGs were showing either DM or DV or both as shown in Table 1. A large number of CpGs ranked Model 4 (DM and DV) as the best model which indicated that both differential mean and differential variability play important role in distinguishing tumor from normal. In KIRC and BRCA, CpGs showing DM only (Model 2) were enriched in CpG islands, first exons, 200 bp upstream of the transcription start sites (TSS200); whereas CpGs showing DV only (Model 3) were enriched in CpG shores and gene body as shown in Figures 2 and 3. In LIHC, the proportions of DM and DV CpGs mapping to CpG islands were fairly similar, whereas the proportion of DM CpGs mapping to gene body was higher compared to DV CpGs. On the other hand, the subtypes comparison within BRCA identified fewer number of CpGs exhibiting DM or DV. In basal versus luminal A or luminal B comparisons, the proportions of DV CpGs mapping to CpG island and TSS200 were higher than DM CpGs.

Among the lists of DM only CpGs (Model 2) identified by tumor versus normal comparison within KIRC, LIHC and BRCA datasets, 4814 CpGs were in common. On the other hand, there were 1223 and 46885 common CpGs in DV only (Model 3) and both DV and DM (DM&DV) (Model 4) categories, respectively. DAVID (`https://david-d.ncifcrf.gov/home.jsp`) functional annotation enrichment analysis was performed on the genes of mapping to each of the top 1000 common DM only CpGs, DV only CpGs and DM&DV CpGs to identify enriched canonical pathways and biological process ontologies. At FDR ≤ 0.05, enriched canonical pathways for DM only CpGs include Rho GTPase cycle, Rap1 signaling pathway and NRAGE signals death through JNK; whereas DM&DV CpGs identified olfactory transduction and signaling pathway among the top enriched pathways. On the other hand, DM only CpGs, DV only CpGs and DM&DV CpGs identified processes related to GTPase regulation, regulation of transcription from RNA polymerase II promoter and regulation of ion transmembrane transport, respectively.

Table 1. Number of CpGs identified for each model based on BIC scores for the different datasets and comparisons.

Data	Model 1	Model 2	Model 3	Model 4
KIRC: tumor vs normal	18685	94948	44291	207972
LIHC: tumor vs normal	85769	52315	83296	144278
BRCA: tumor vs normal	33735	104575	43880	192490
BRCA: basal vs luminal A	201378	131085	23193	19024
BRCA: basal vs luminal B	198192	124764	31393	20331
BRCA: luminal B vs luminal A	290963	47145	31327	5245

4.3. *Elastic net predictive modeling*

The elastic net algorithm[27] was applied to each dataset for constructing a prediction model differentiating case from control. We randomly split the dataset into 80% training and 20%

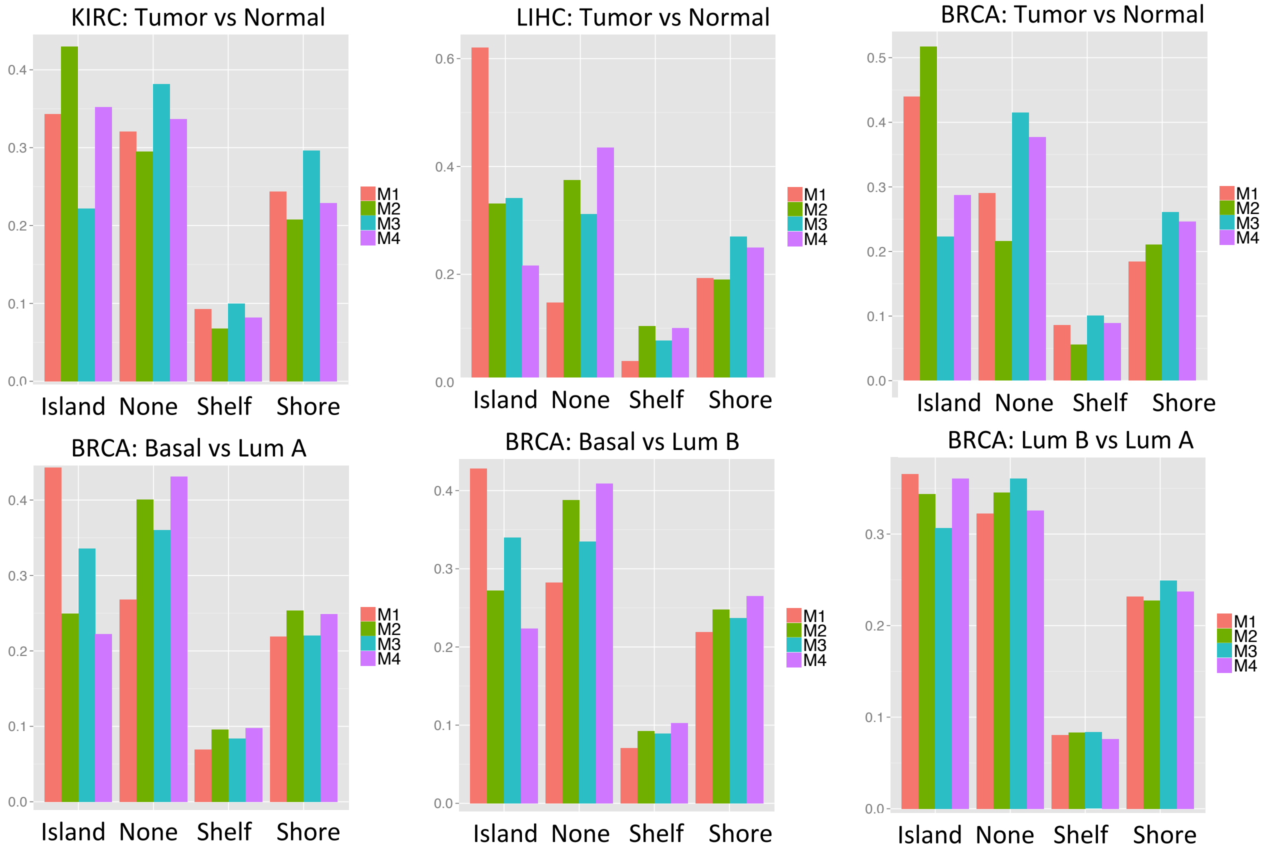


Fig. 2. CpG island, shelf and shore annotation for the proportion of CpGs identified by each model (color code: orange (Model 1), green (Model 2), turquoise (Model 3), purple (Model 4)) for the different datasets and comparisons.

test set. The parameters λ and α were tuned using 10 fold cross-validation on the training set. The random partitioning of data into training and test set was repeated 10 times. We compared the following methods for selecting top 2000 CpGs from the training set to be included as candidate features:

(i) **Set 1**: Logit transformed beta values x_{ij} of the top 2000 CpGs among the CpGs which ranked model 2 as the best model.

(ii) **Set 2**: Absolute deviation measure r_{ij} of the top 2000 CpGs among the CpGs which ranked model 3 as the best model.

(iii) **Set 3**: Both the logit transformed beta values x_{ij} and absolute deviation measure r_{ij} of the top 2000 CpGs among the CpGs which ranked model 4 as the best model.

We evaluated the performance of the prediction algorithm on the test set in terms of area under the receiver operating characteristics curve (AUC), accuracy (Acc)= $\frac{TP+TN}{n_0+n_1}$, sensitivity (Sn)= $\frac{TP}{TP+FN}$, specificity (Sp)= $\frac{TN}{TN+FP}$ and Matthew's correlation coefficient (Mcc)= $\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, averaged over the 10 iterations. The results are presented in Table 2. The prediction model for predicting tumor from normal in KIRC, LIHC

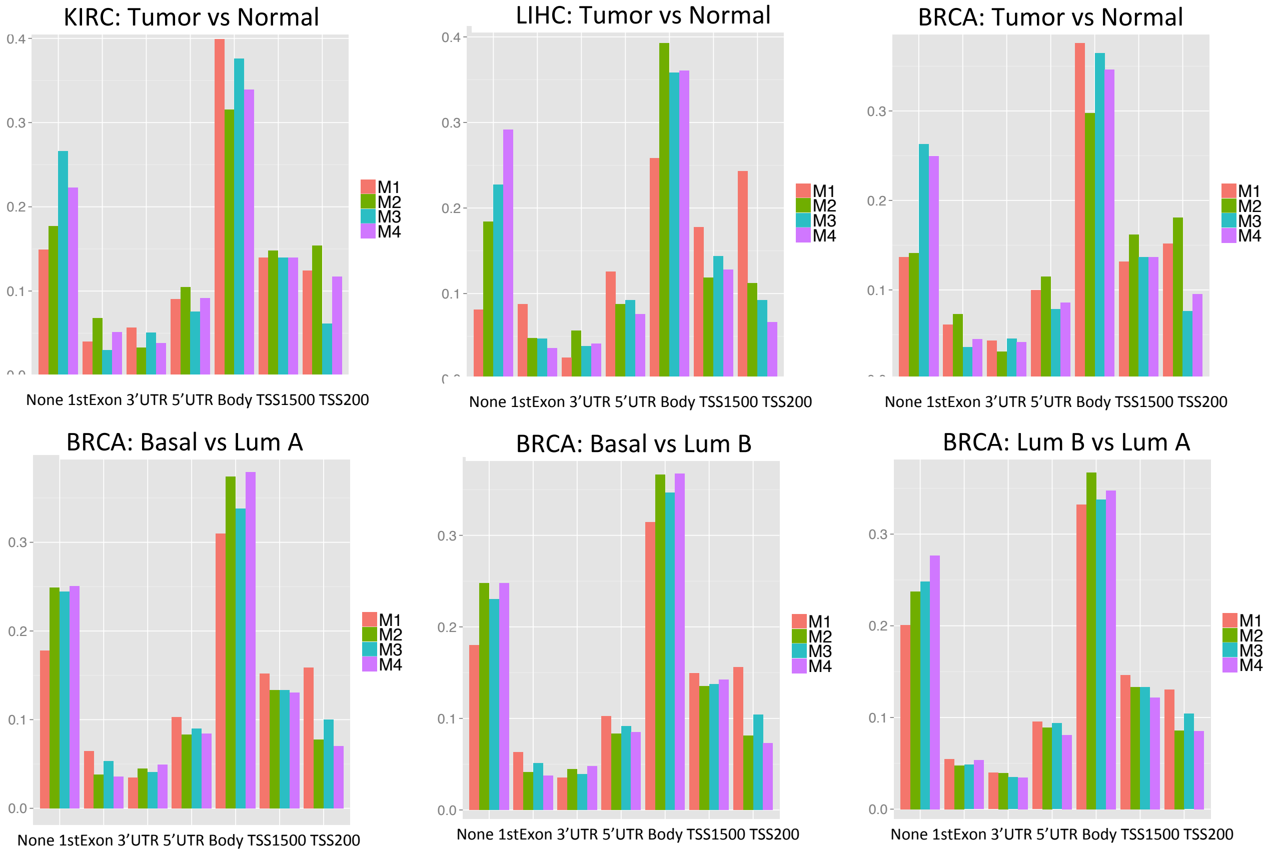


Fig. 3. Gene annotation for the proportion of CpGs identified by each model (color code: orange (Model 1), green (Model 2), turquoise (Model 3), purple (Model 4)) for the different datasets and comparisons.

and BRCA had high accuracy and AUC, and were comparable across the different candidate feature sets. Similar patterns were observed in basal versus luminal A and basal versus luminal B comparisons, indicating that DNA methylation was able to differentiate the more aggressive subtype (basal) from the less aggressive subtypes (luminals A and B) regardless of whether DM or DV CpGs were used. On the other hand, the prediction algorithm for predicting luminal A from luminal B subtypes exhibited lower accuracy compared to the previous comparisons, indicating that it is harder to differentiate these two subtypes based on DNA methylation.

5. Discussion

The promise and power of DNA methylation for therapeutics and diagnostics have been demonstrated in various diseases including cancer. Advancements in biotechnology enable large scale and population based epigenome-wide profiling of DNA methylation for identifying differential mean (DM) and differential variability (DV) CpGs. In these studies, covariates such as demographic and clinical factors may be confounded with both DNA methylation and disease phenotypes. One way to circumvent this problem is via randomization. However, this approach is not always feasible especially in case control studies. Moreover, in DNA

Table 2. Average AUC, Mcc, Accuracy (Acc), Sensitivity (Sn) and Specificity (Sp) for the different datasets and comparisons.

Candidate feature	AUC	Mcc	Acc	Sn	Sp
KIRC: tumor vs normal					
Set 1	1.000	0.998	0.999	0.998	1.000
Set 2	1.000	0.991	0.996	0.994	1.000
Set 3	1.000	0.998	0.999	0.998	1.000
LIHC: tumor vs normal					
Set 1	0.996	0.933	0.986	0.992	0.940
Set 2	0.994	0.913	0.981	0.985	0.950
Set 3	0.997	0.929	0.984	0.987	0.960
BRCA: tumor vs normal					
Set 1	1.000	0.976	0.992	0.997	0.975
Set 2	0.999	0.969	0.990	0.993	0.978
Set 3	1.000	0.976	0.992	0.997	0.972
BRCA: basal vs luminal A					
Set 1	0.996	0.947	0.980	0.950	0.989
Set 2	0.987	0.848	0.944	0.817	0.984
Set 3	0.995	0.947	0.980	0.950	0.989
BRCA: basal vs luminal B					
Set 1	0.998	0.905	0.950	0.950	0.950
Set 2	0.996	0.905	0.950	0.967	0.938
Set 3	0.998	0.889	0.943	0.950	0.938
BRCA: luminal B vs luminal A					
Set 1	0.798	0.339	0.741	0.425	0.874
Set 2	0.720	0.287	0.722	0.413	0.853
Set 3	0.791	0.380	0.767	0.413	0.916

methylation studies using whole blood sample, the different cell types have been shown to be confounded with the measured methylation levels.[32] In such cases, confounding factors need to be properly accounted for to avoid biases in DNA methylation biomarker detection. There are several approaches for DM analysis which allow for confounders adjustment,[33] however to the best of our knowledge existing DV analysis approaches are not tailored for confounders adjustments, except for our earlier work[16] which proposed a DV only analysis in the presence of confounders within large scale hypothesis testings framework. This paper extends our earlier work which allows for simultaneous detection of DM and DV in large scale hypothesis testings framework, and at the same time provides a candidate feature selection mechanism for the prediction algorithm.

We showed that the analysis on KIRC, LIHC and BRCA TCGA datasets identified DM and DV CpGs which mapped to different CpG and gene annotations. For instance, in tumor versus normal comparisons, a larger proportion of DM CpGs mapped to CpG island and TSS200, whereas in basal versus luminal A or B comparisons, a larger proportion of DV CpGs mapped to these regions, suggesting that DM and DV CpGs regulate transcription differently. An R package `methylDMV` implementing this flexible framework is available at `http://www.ams.sunysb.edu/~pfkuan/softwares.html#methylDMV`.

DNA methylation generated from high resolution arrays including Illumina Infinium HumanMethylation450 BeadChip may induce a natural correlation structure among neighboring CpGs. An immediate extension of our current framework is to model the dependence structure and borrow information from nearby CpGs to improve the power of detecting DM and DV CpGs. Two of such approaches are (1) the hidden Markov model and local index of significance method as in Kuan et al. (2012),[34] and (2) the smoothing and bump hunting method as in Jaffe et al (2012),[7] which can possibly be adapted into our current `methylDMV` framework for detecting DM and DV CpGs.

Acknowledgments

The authors thank the reviewers for their constructive comments and suggestions.

References

1. V. Rakyan, T. Down, N. Thorne, P. Flicek, E. Kulesha, S. Graf, E. Tomazou, L. Backdahl, N. Johnson, M. Herberth, K. Howe, D. Jackson, M. Miretti, H. Fiegler, J. Marioni, E. Birney, T. Hubbard, N. Carter, S. Tavare and S. Beck, *Genome Research* **18**, 1518 (2008).
2. M. Esteller, *Annual Review Pharmacological Toxicology* **45**, 629 (2005).
3. R. Irizarry, C. Ladd-Acosta, B. Carvalho, H. Wu, S. Brandenburg, J. Jeddeloh, B. Wen and A. Feinberg, *Genome Research* **18**, 780 (2008).
4. P. Wang, Q. Dong, Z. Chong, P. Kuan, Y. Liu, W. Jeck, W. Jiang, G. S. nd T. Tan, J. Andersen, T. Auman, J. Hoskins, A. Misher, C. Moser, S. Yourstone, J. Kim, K. Cibulskis, S. Getz, H. Hunt, S. Thorgerisson, L. Roberts, D. Ye, K. Guan, Y. Xiong, L. Qin and D. Chiang, *Oncogene* **32**, 3091 (2012).
5. K. Conway, S. Edmiston, Z. Khondker, P. Groben, X. Zhou, H. Chu, P. Kuan, H. Hao, C. Carson, M. Berwick, D. Olilla and N. Thomas, *Pigment Cell and Melanoma Research* **24**, 352 (2011).
6. K. Hansen, W. Timp, H. Bravo, S. Sabunciyan, B. Langmead, O. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. Irizarry and A. Feinberg, *Nature Genetics* **26**, 768 (2011).
7. A. Jaffe, A. Feinberg, R. Irizarry and J. Leek, *Biostatistics* **13**, 166 (2012).
8. A. Teschendorff and M. Widschwendter, *Bioinformatics* **28**, 1487 (2012).
9. X. Xu, S. Su, V. Barnes, C. Miguel, J. Pollock, D. Ownby, H. Shi, H. Zhu, H. Snieder and X. Wang, *Epigenetics* **8**, 522 (2013).
10. A. Feinberg and R. Irizarry, *Proc. Natl Acad. Sci. USA* **107**, 1757 (2010).
11. A. Teschendorff, A. Jones, H. Fiegl, A. Sargent, J. Zhuang, H. Kitchener and M. Widschwendter, *Genome Medicine* **4**, p. DOI: 10.1186/gm323 (2012).
12. G. Smyth, *Statistical Application in Genetics and Molecular Biology* **3**, p. 3 (2004).
13. S. Ahn and T. Wang, *Pacific Symposium of Biocomputing* , 69 (2013).

14. B. Phipson and A. Oshlack, *Genome Biology* **15**, DOI: 10.1186/s13059 (2014).
15. S. Wahl, N. Fenske, S. Zeilinger, K. Suhre, C. Gieger, M. Waldenberger, H. Grallert and M. Schmidt, *BMC Bioinformatics* **15**, DOI: 10.1186/1471 (2014).
16. P. Kuan, *Statistical Applications in Genetics and Molecular Biology* **13**, 645 (2014).
17. O. Stefansson, S. Moran, A. Gomez, S. Sayols, C. Arribas-Jorba, J. Sandoval, H. Hilmarsdottir, E. Olasfdottir, L. Tryggvadottir, J. Jonasson, J. Eyfjord and M. Esteller, *Molecular Oncology* **9**, 555 (2015).
18. J. Zhuang, M. Widschwendter and A. Teschendorff, *BMC Bioinformatics* **13**, DOI: 10.1186/1471 (2012).
19. S. Horvath, *Genome Biology* **14**, p. R115 (2013).
20. M. Jung and G. Pfeifer, *BMC Biology* **13**, doi: 10.1186/s12915 (2015).
21. M. Dogan, B. Shields, C. Cutrona, L. Gao, F. Gibbons, R. Simons, M. Monick, G. Brody, K. Tan, S. Beach and R. Philibert, *BMC Genomics* **15**, DOI: 10.1186/1471 (2014).
22. K. Lee and Z. Pausova, *Frontiers in Genetics* **4**, p. doi: 10.3389/fgene.2013.00132 (2013).
23. A. Houseman, B. Christensen, R. Yeh, C. Marsit, M. Karagas, M. Wrensch, H. Nelson, J. Wiemels, S. Zheng, J. Wiencke and K. Kelsey, *BMC Bioinformatics* **9**, doi:10.1186/1471 (2008).
24. P. Du, X. Zhang, C. Huang, N. Jafari, W. Kibbe, L. Hou and S. Lin, *BMC Bioinformatics* **11** (2010).
25. C. Rao, *Proceedings of the Cambridge Philosophical Society* **44**, 50 (1948).
26. Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **57**, 289 (1995).
27. H. Zou and T. Hastie, *Journal of the Royal Statistical Society, Series B* **67**, 301 (2005).
28. L. Breiman, *Journal of Machine Learning* **45**, 5 (2001).
29. A. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero and S. Beck, *Bioinformatics* **29**, 189 (2013).
30. E. Price, A. Cotton, L. Lam, P. Farre, E. Emberly, C. Brown, W. Robinson and M. Kobor, *Epigenetics and Chromatin* **6** (2013).
31. Y. Chen, M. Lemire, S. Choufani, D. Butcher, D. Grafodatskaya, B. Zanke, S. Gallinger, T. Hudson and R. Weksberg, *Epigenetics* **8**, 203 (2013).
32. A. Houseman, W. Accomando, D. Koestler, B. Christensen, C. Marsit, H. Nelson, J. Wiencke and K. Kelsey, *BMC Bioinformatics* **13**, 189 (2012).
33. M. Ritchie, B. Phipson, D. Wu, Y. Hu, C. Law, W. Shi and G. Smyth, *Nucleic Acids Research* **43**, p. e47 (2015).
34. P. Kuan and D. Chiang, *Biometrics* **68**, 774 (2012).

IDENTIFY CANCER DRIVER GENES THROUGH SHARED MENDELIAN DISEASE PATHOGENIC VARIANTS AND CANCER SOMATIC MUTATIONS

MENG MA[1], CHANGCHANG WANG[2], BENJAMIN S. GLICKSBERG[1], ERIC E. SCHADT[1], SHUYU D. LI[1]*, RONG CHEN[1]*

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl.
New York City, NY 10029, USA

[2]School of Computer Science, Anhui University, Anhui, P.R. China
**Email: rong.chen@mssm.edu; shuyudan.li@mssm.edu.*

Genomic sequencing studies in the past several years have yielded a large number of cancer somatic mutations. There remains a major challenge in delineating a small fraction of somatic mutations that are oncogenic drivers from a background of predominantly passenger mutations. Although computational tools have been developed to predict the functional impact of mutations, their utility is limited. In this study, we applied an alternative approach to identify potentially novel cancer drivers as those somatic mutations that overlap with known pathogenic mutations in Mendelian diseases. We hypothesize that those shared mutations are more likely to be cancer drivers because they have the established molecular mechanisms to impact protein functions. We first show that the overlap between somatic mutations in COSMIC and pathogenic genetic variants in HGMD is associated with high mutation frequency in cancers and is enriched for known cancer genes. We then attempted to identify putative tumor suppressors based on the number of distinct HGMD/COSMIC overlapping mutations in a given gene, and our results suggest that ion channels, collagens and Marfan syndrome associated genes may represent new classes of tumor suppressors. To elucidate potentially novel oncogenes, we identified those HGMD/COSMIC overlapping mutations that are not only highly recurrent but also mutually exclusive from previously characterized oncogenic mutations in each specific cancer type. Taken together, our study represents a novel approach to discover new cancer genes from the vast amount of cancer genome sequencing data.

1. Introduction

Significant efforts in the past several years in cancer genomic sequencing by individual investigators and large consortium such as The Cancer Genome Atlas (TCGA) and The International Cancer Genome Consortium (ICGC) have uncovered a large number of novel oncogenic drivers. These studies not only advanced our understanding on the genetic basis of tumorigenesis and cancer progression, but also significantly enabled the development of personalized cancer therapeutics [1, 2]. Cancer genome or exome sequencing data have been generated from approximately 25,000 tumor samples covering more than 50 tumor types [3, 4], representing a comprehensive cancer genomic atlas. While data generation has been greatly facilitated by rapid technology development, interpretation of cancer sequence information still remains a major challenge. As most solid tumors harbor a median of 40-80 non-synonymous somatic mutations per tumor, only three to six of them are driver mutations [5]. The most commonly used approach to distinguish a small number of driver mutations from those background passenger mutations is to identify significantly mutated genes in a cohort study [6]. The underlying rationale is if a gene is mutated at significantly greater rate than the background mutation rate, it is more likely to be oncogenic, as the mutations conferring tumor growth advantage are evolutionarily selected during cancer development. To complement this approach, various computational tools have been developed to assess the effects of missense mutations on protein functions [7]. While such an approach has further characterized numerous novel cancer drivers and oncogenic pathways from cancer genomic sequencing data, it requires a large number of samples to uncover those drivers mutated at low population frequency in a given tumor type. This is particularly problematic for those cancers with high background mutation rates such as melanomas and lung cancers. For example, it has been estimated that it would require approximately 4,000 melanoma patient samples to detect cancer genes mutated at 2% frequency, and more than 20,000 samples for genes mutated at 1% with 90% power for 90% of genes [8].

Many human genetic diseases are Mendelian disorders caused by one or more aberrations in the genome. These diseases are often heritable as the disease causing, pathogenic variants are passed on from parents' genome. To date, approximately 180,000 genetic variants in more than 7,000 genes have been identified as pathogenic for more than 4,000 Mendelian diseases [9]. Some of the first established cancer genes with frequent somatic mutations were originally identified from their associations with familial cancer syndromes. The first tumor suppressor RB1 was discovered by studying the familial form of retinoblastoma [10]. The most frequently mutated gene in cancers, p53, was also identified as a tumor suppressor inactivated in Li–Fraumeni syndrome, a rare cancer predisposition hereditary disorder. Other well-known cancer genes harboring high frequency somatic mutations and that are associated with Mendelian diseases include VHL in Von Hippel-Lindau syndrome, MLH1, MSH2, MSH6 in Lynch syndrome, TSC1, TSC2 in Tuberous sclerosis, and ATM in ataxia-telangiectasia [11]. Notably, a recent study has revealed potentially novel cancer-associated genes through analysis of comorbidity between cancers and Mendelian diseases [12].

By definition, germline pathogenic variants impact the functions of key proteins involved in the developmental process and consequently cause heritable diseases. If the same germline pathogenic variants occur as somatic mutations in cancers, these mutations would also alter protein functions and may play a role in tumor initiation and progression, even though the same proteins can have very different functions during development than in adult tissues. Indeed in a recent report, several genes sharing identical mutations in Mendelian diseases and cancers were proposed as novel cancer genes [13]. Based on this underlying hypothesis, we carried out a systematic comparative analysis of the reported pathogenic variants in Mendelian diseases and cancer somatic mutations. There are several repositories for pathogenic variants. A comparison of four of the most comprehensive databases showed that HGMD is currently the largest collection of human disease variants, although each database has its own advantages in terms of the information collected as well as database infrastructure [9]. For cancer somatic mutations, COSMIC is recognized as the most comprehensive resource for somatic mutations in human cancers [14], with more than 1.4 million confirmed somatic mutations identified from 1.1 million tumor samples including genome-wide sequencing data from more than 20,000 tumors. In this study, we first identified overlapping mutations between pathogenic variants in HGMD [15] and cancer somatic mutations from the COSMIC database [14]. Further characterization of these mutations show that the mutation-harboring genes are significantly enriched for known cancer genes, supporting the above described hypothesis. We then examined those genes harboring the shared pathogenic variants and somatic mutations in cancers by applying additional filters such as the number of overlapping HGMD/COSMIC mutations in a given gene or the frequency of overlapping mutations in each tumor type. Moreover, those overlapping mutations with high recurrence in cancers were subjected to mutual exclusivity analysis with known oncogenes in each tumor type in order to identify novel oncogenic drivers. Taken together, our study represents a novel approach to discover new cancer genes from the vast amount of cancer genome sequencing data.

2. Methods

COSMIC V73 was downloaded from sftp-cancer.sanger.ac.uk using GUI client WinSCP under protocol sftp and port 22. HGMD Professional can be accessed from https://www.qiagenbioinformatics.com/products/human-gene-mutation-database/ with an authorized license. 1000 Genome Phase3 was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. ExAC database was downloaded from ftp://ftp.broadinstitute.org/pub/ExAC_release/. All RefSeq Exons were downloaded from UCSC table refGene through UCSC Table Browser (clade: Mammal, genome: Human, assembly: Feb.2009 (GRCH37/hg19), group: Genes and Gene Predictions, track: RefSeq Genes, Table: refGene). Cancer Gene Census dataset was downloaded from http://cancer.sanger.ac.uk/census/.

All the analyses were performed using shell scripts, mysql scripts and R scripts. The mutual exclusivity heat map was generated using gitools (http://www.gitools.org/). The survival analysis was done through cBioPortal (http://www.cbioportal.org/). Several major scripts for database query and statistical analyses are available on github (https://github.com/CosmicHGMD/CancerMendelian).

3. Results

3.1. *Identification of overlapping pathogenic variants in HGMD and somatic mutations in COSMIC*

HGMD includes six classes of variants, and we only included disease-causing mutations (DM and DM?) in our analysis. The DM class variants have been demonstrated in literature to confer the associated clinical phenotype of the affected individuals. The DM? class variants have some degree of uncertainty, but nevertheless have strong evidence supporting their pathogenicity. At the time of this writing, there are a total of 153,593 DM/DM? class variants in HGMD database. 11,523 of these variants are present in the COSMIC database, representing 0.54% of the total mutations in COSMIC (Table 1). When we only include the confirmed somatic mutations in COSMIC, there are 8,582 mutations (0.6%) that overlap with HGMD DM/DM? variants. As the majority of the somatic mutation data in COSMIC are from cancer genomic sequencing studies, some of these mutations are likely false positives, particularly those from early whole genome/exome sequencing when computational methods for calling somatic mutation were less reliable or if the identified somatic mutations were not validated by a different sequencing platform. Therefore, we further restrict COSMIC data to include only those somatic mutations occurred in more than one tumor samples. Although the total number of overlapping mutations with HGMD DM/DM? variants is reduced to 3,470, using this limited but more reliable somatic mutation list, the percentage with respect to the total number of these recurrent mutations (215, 436) in COSMIC increases to 1.6% (Table 1), suggesting Mendelian disease pathogenic variants are over-represented in recurrent somatic mutations in cancers.

Then we randomly selected the same number of genetic variants (153,593) from 1000 genome (exonic region) or the ExAC database as control variant datasets, and performed the same analysis. The analysis of randomly selected, mostly non-pathogenic common genetic variants was repeated 1000 times, and the results indicated that percentages of common non-pathogenic variants overlapping with COSMIC mutations are lower than the HGMD pathogenic variants (Table 1). The statistical significance was assessed based on the distribution of results from 1000 simulations. This finding supports our initial hypothesis that overlapping pathogenic variants in HGMD with cancer somatic mutations could enable identification of novel cancer genes.

Table 1. Enrichment of HGMD pathogenic variants in cancer somatic mutations.

Variant dataset (total number of variants)	**Randomly selected variants**	**Overlap with COSMIC mutations (percentage)**		
		All mutations in COSMIC (2,132,117)	**Somatic mutations in COSMIC (1,425,978)**	**Recurrent somatic mutations in COSMIC (215,436)**
HGMD DM/DM? (153,593)	-	11,523 (0.54%)	8,582 (0.60%)	3,470 (1.6%)
1000 Genome exonic region (2,156,973)	153,593	8,092 (0.38%); p<0.001	5,983 (0.42%); p<0.001	1,975 (0.92%); p<0.001
ExAc (10,450,722)	153,593	6,919 (0.32%); p<0.001	4,841 (0.34%); p<0.001	1,325 (0.62%); p<0.001

Next, we tested if HGMD/COSMIC overlapping mutations are more likely to occur at high frequency in cancers than those somatic mutations non-overlapping with HGMD. We first divided the confirmed COSMIC somatic mutations into two groups. The first group includes those mutations overlapped with HGMD DM/DM? variants and the second group includes the rest of somatic mutations that are only present in the COSMIC database. Then, for a given recurrence frequency cutoff *c*, we computed the percentage of somatic mutations with recurrence frequency (*f*) greater than *c* in group 1 (denoted as $\%G1_{f>c}$) and those in group 2 (denoted as $\%G2_{f>c}$). This is followed by computing the ratio of $\%G1_{f>c}$ over $\%G2_{f>c}$ at various mutation frequencies. As illustrated in Figure 1A, as the recurrence frequency (x-axis) increases, this ratio (y-axis) also increases. For example, the ratio is approximately 25 for recurrence frequency 20, indicating that COSMIC mutations overlapping with HGMD pathogenic variants are 25 fold more likely to occur in more than 20 tumor samples than those not overlapping with HGMD variants. We also directly plotted $\%G1_{f>c}$ and $\%G2_{f>c}$ (Figure 1B), and it clearly shows the HGMD/COSMIC overlapping mutations have higher mutation frequencies than those mutations only in the COSMIC database with mean recurrence in 8.0 and 1.3 tumors respectively (p = 1.5E-5, one-sided t-test). Because the likelihood that a somatic mutation is a cancer driver increases with its mutation frequency in cancers, this result is consistent with the hypothesis that cancer mutations overlapping with germline disease pathogenic variants in HGMD are more likely to be oncogenic. We further examined the presence of known cancer genes in the two groups using cancer gene census annotation [16]. While only 4.3% of the COSMIC somatic mutations do not overlap with HGMD are in the cancer gene census list, there are approximately 10% of the somatic mutations overlapping with HGMD occur in cancer census genes.

To determine if the combination of somatic mutation frequency and the presence of overlap with HGMD pathogenic mutations would facilitate cancer gene discovery, we computed the percentage of somatic mutations mapped to cancer census genes in all COSMIC confirmed somatic mutations or only in those overlapped with HGMD DM/DM? variants. This procedure was then repeated for mutations with increasing frequencies (Figure 2). Two observations were notable from the results. First, a somatic mutation is more likely to be in a cancer gene as its frequency increases, evidenced by increasing percentage of cancer census genes. Second, the probability that the mutation-harboring genes are cancer-related increases if there are overlapping with HGMD variants (Figure 2, red bars vs. blue bars).

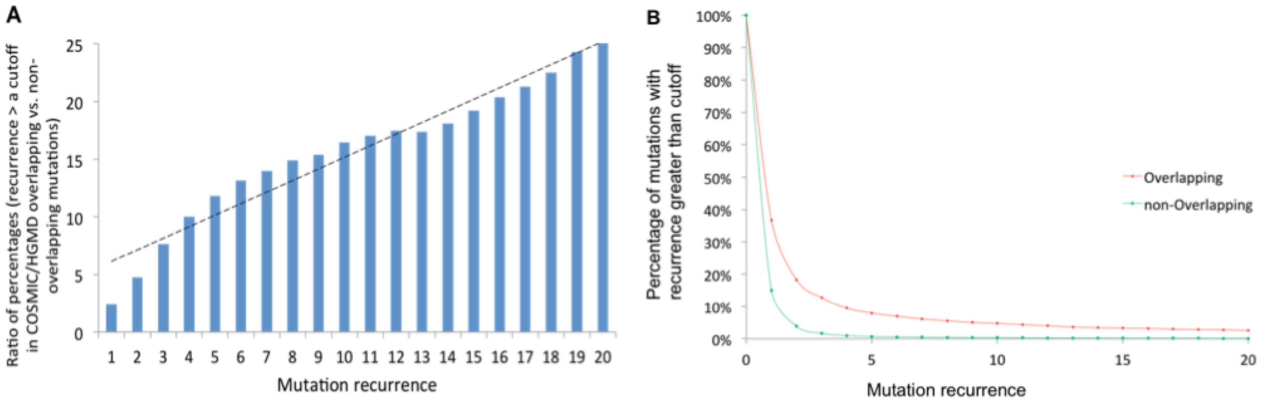


Figure 1. Overlap of HGMD variants with cancer somatic mutations is correlated with high mutation recurrence in cancers.

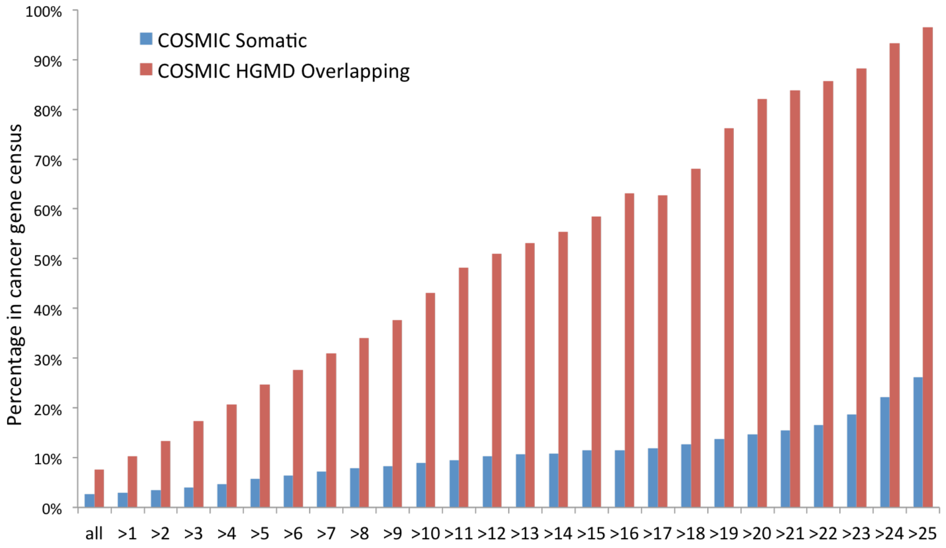


Figure 2. Novel cancer gene discovery through overlapping with HGMD and high mutation recurrence. X-axis represents mutation recurrence in COSMIC.

3.2. *Identification of potential tumor suppressors*

We examined whether the number of distinct overlapping HGMD/COSMIC mutations in a given gene is associated with the probability that the gene is a cancer gene. Figure 3 shows that as the number of distinct overlapping HGMD/COSMIC mutations in a given gene increases, the percentage of genes that belong to cancer gene census increases as well. Of those genes with more than six distinct overlapping mutations, approximately 25% are present in cancer gene census. It is generally recognized that while oncogenes are often mutated recurrently at certain positions (referred as hotspots), tumor suppressors tend to lack such mutational hotspots and are mutated at many positions across the gene sequences. Therefore, we reason that identifying genes with high number of distinct overlapping HGMD/COSMIC mutations would allow us to discover potentially novel tumor suppressors.

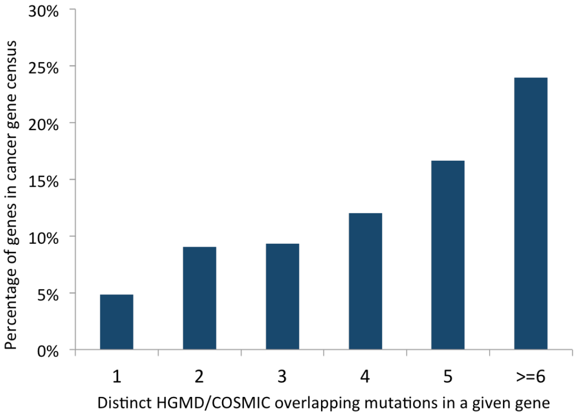


Figure 3. Identification of novel tumor suppressors based on the number of distinct HGMD/COSMIC overlapping mutations.

Accordingly, we ranked all the genes in COSMIC based on the number of distinct somatic mutations that overlap with HGMD DM/DM? variants and provided both cancer gene census annotations as well as oncogene/tumor suppressor classifications according to Vogelstein et al. [5] in Table 2. Almost half (23/48) of the genes with at least 20 overlapping HGMD/COSMIC mutations are in the cancer gene census list and/or annotated as an oncogene or a tumor suppressor, furthering the notion that HGMD pathogenic variant annotation may help distinguish driver oncogenic mutations from the passenger mutations in tumors. As expected, most of those genes with oncogene/tumor suppressor annotations are classified as tumor suppressors (19/21, 90%; Table 2). A literature search has provided support that some of the remaining genes are likely novel tumor suppressors. There are several genes that encode ion channels with many HGMD/COSMIC overlapping mutations, including SCN5A (67 overlapping mutations), SCN1A (52), CFTR (48), RYR1 (36) and RYR2 (30) (Table 2). While ion channels have not been recognized as a major class of cancer related genes, emerging evidence suggest at least some ion channels are involved in promoting malignancy. For example, CFTR, the cystic fibrosis (CF) gene, has been postulated to be a tumor suppressor because loss of CFTR enhanced tumor cell proliferation and epithelial-to-mesenchymal transition, and is associated with poor prognosis in several cancer types [17, 18, 19]. We also observed multiple collagen family genes with significant overlap between HGMD and COSMIC mutations, such as COL3A1 (29 overlapping mutations), COL7A1 (22) (Table 2), COL1A2 (19), COL4A5 (18), COL2A1 (14), COL6A3 (13), COL1A1 (13), and COL4A4 (11) (data not shown). Although collagens are considered as a barrier to suppress angiogenesis since they are key components of extracellular matrix in tumor microenvironment, only recent functional studies have shown a causal relationship between loss of collagens and tumor progression [20]. Our results suggest that collagens may represent another new class of tumor suppressors. Notably, two genes FBN1 and TGFBR2, associated with a genetic disorder of connective tissue known as Marfan syndrome [21, 22], had 43 and 22 HGMD/COSMIC overlapping mutations respectively. Upon further investigation, we found that the two genes are mutated frequently in lung squamous cell carcinomas (SCCs) with a combined mutation frequency 10% in the TCGA cohort [23]. Moreover, FBN1 and TGFBR2 mutations are associated with poor survival. As shown in Figure 4, FBN1 mutation-harboring lung SCCs had poor disease progression free survival (DFS) (Figure 4A), and those patients with TGFBR2 mutations had both poor DFS and overall survival (OS) (Figure 4B, 4E). The combined FBN1 and TFGBR2 mutations are associated with both poor DFS and OS (Figure 4C, 4F).

Table 2. Genes ranked by the number of distinct overlapping HGMD-COSMIC mutations. Only genes with at least 20 overlapping mutations are shown. CGC: cancer gene census. TSG: tumor suppressor gene.

Gene	Mutations	CGC	Oncogene/TSG	Gene	Mutations	CGC	Oncogene/TSG
TP53	198	Yes	TSG	F9	33		
APC	192	Yes	TSG	DMD	33		
VHL	173	Yes	TSG	PKHD1	31		
NF1	148	Yes	TSG	SMAD4	31	Yes	TSG
PTEN	145	Yes	TSG	PTPN11	31	Yes	Oncogene
RB1	91	Yes	TSG	RYR2	30		
SCN5A	67			COL3A1	29		
CDKN2A	66	Yes	TSG	MLH1	29	Yes	TSG
NF2	65	Yes	TSG	BRCA1	29	Yes	TSG
KMT2D	56	Yes		MSH2	28	Yes	TSG
F8	56			VWF	27		
MYH7	54			TSC2	27	Yes	
SCN1A	52			STK11	27	Yes	TSG
USH2A	50			PTCH1	27	Yes	TSG
ATM	50	Yes	TSG	ATP7B	25		
MEN1	49	Yes	TSG	WT1	24	Yes	TSG
CFTR	48			TGFBR2	22		
FBN1	43			PAH	22		
HNF1A	39	Yes	TSG	IRF6	22		
RET	39	Yes	Oncogene	COL7A1	22		
ABCA4	37			CASR	22		
RYR1	36			APOB	22		
BRCA2	36	Yes	TSG	GCK	21		
LDLR	35			MYBPC3	20		

3.3. *Identification of potential oncogenes*

To identify putative oncogenes from the overlapping HGMD/COSMIC mutations, we applied two criteria. First, as most well-known oncogenic, activating mutations are highly recurrent in a specific tumor type, we ranked HGMD/COSMIC overlapping mutations by their mutation frequency. This was done separately for each tumor type in COSMIC. Second, because different oncogenic mutations in a given tumor type are often mutually exclusive, we performed mutual exclusivity analysis to identify those HGMD/COSMIC overlapping mutations that are not only highly recurrent but also mutually exclusive from mutations in known oncogenes based on oncogene classification by Vogelstein et al [5].

To achieve sufficient statistical power in mutual exclusivity analysis, we only analyzed 19 tumor types with at least 200 samples that had whole genome or exome sequencing data in COSMIC and focused on those HGMD/COSMIC overlapping mutations in non-cancer genes (oncogene or tumor suppressor according to Vogelstein et al.) that are mutated in at least 1% of the total samples in a specific tumor type. Interestingly, of the 19 tumor types we analyzed, only endometrium, large intestine, and upper aero-digestive tract (UADT) cancers had such mutations, indicating that while only a very small percentage of COSMIC somatic mutations overlap with HGMD pathogenic variants (Table 1), even fewer are mutated in cancers with high recurrence. Notably, the ACVR1 R206H mutation occurred in 3 endometrium cancer samples, and an additional endometrium tumor harbors the ACVR1 G356D mutation. Mutual exclusivity analysis revealed that 3 of these 4 samples are mutually exclusive from the most frequently mutated oncogene PIK3CA, CTNNB1 and KRAS in this tumor type (p-value = 0.078; Figure 5).

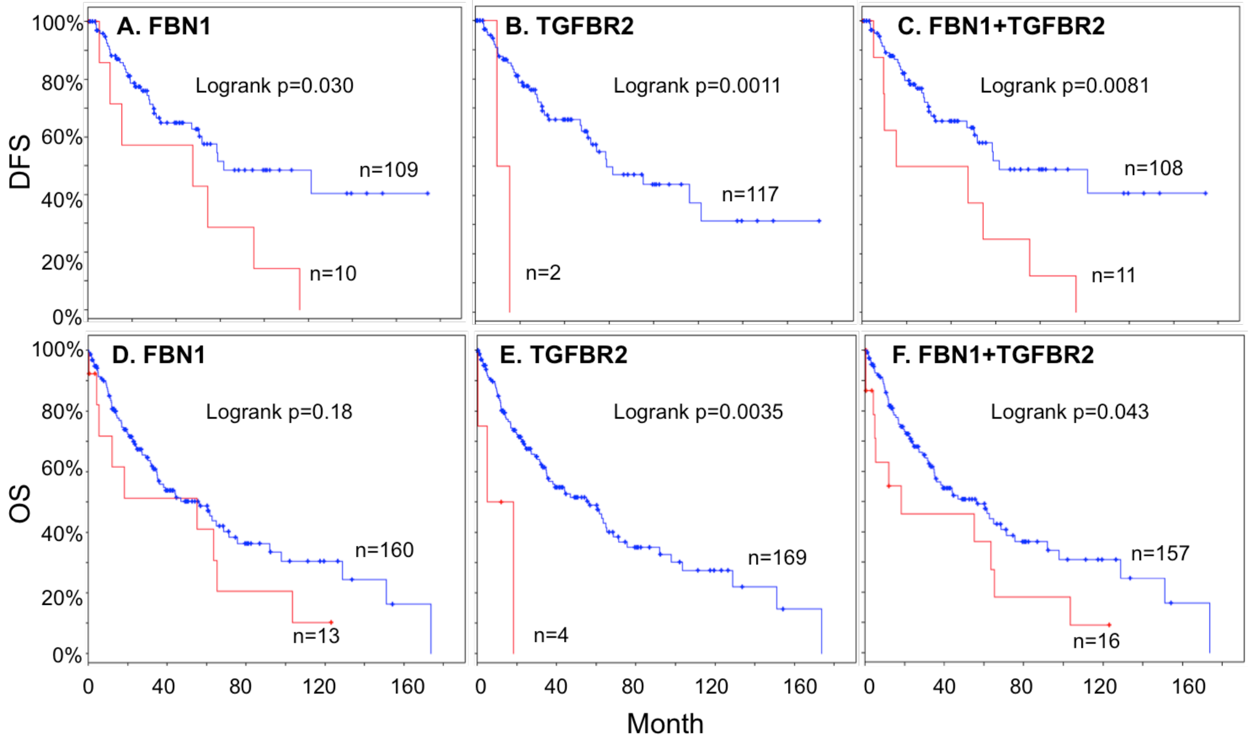


Figure 4. FBN1 and TGFBR2 mutations are associated with poor survival in lung squamous cell carcinomas. Disease free survival (DFS) are shown in panel A-C, and overall survival (OS) are shown in panel D-F. Red curves represent patients harboring somatic mutations for the indicated gene and blue curves represent patients with wild type gene. Sample size in red and blue curves, and logrank p-values in survival analysis are shown in each panel.

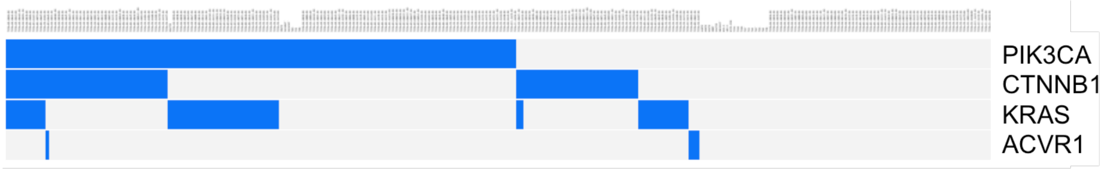


Figure 5. Mutual exclusivity of HGMD/COSMIC overlapping ACVR1 mutations from most frequently mutated oncogenes in endometrium cancers. Each column represents a tumor sample. The presence of a mutation in each gene in a given tumor sample is indicated by the blue color.

Since the above approach combining high mutation frequency and mutual exclusivity from known oncogenic drivers in each specific tumor type led to very few candidates as putative oncogenic mutations, we ranked somatic mutations only based on frequency across all cancer types in COSMIC without taking mutual exclusivity into consideration (Table 3). Many oncogenes have multiple mutational hotspots, and therefore for each gene we only show the mutation (at amino acid level) with the highest recurrence. Of genes with the most recurrent amino acid change occurring in at least 15 tumors, 18 had oncogene/tumor suppressor annotations (Table 3). While 50% (9/18) are classified as oncogenes, the presence of many tumor suppressors is not surprising because mutational hotspots (typically dominant negative mutations) are also observed in some tumor suppressors such TP53 [24]. The remaining genes without oncogene/tumor suppressor annotations provide possible candidate oncogenes due to the presence of mutational hotspots. It is noteworthy that there are 3 protein kinases that had a recurrent somatic mutation detected in more than 10 but less than 15 tumor samples: RAF1, p.S257L, 13 tumors; FGFR4, p.G388R, 12 tumors; TYK2, p.V362F, 12 tumors. Although the 3 kinases are not recognized as oncogenes, there are strong evidences that these recurrent mutations are activating and/or oncogenic [25, 26, 27], suggesting RAF1, FGFR4 and TYK2 are likely novel oncogenes.

4. Discussion

Owing to technological advancement and cost reduction, genomic sequencing is a new paradigm in cancer research and personalized cancer therapeutics. A large number of cancer somatic mutations have been described from whole genome/exome sequencing studies. As only a small percentage of somatic mutations are cancer drivers, it is of paramount importance to distinguish those driver mutations from a background of predominantly passenger mutations. Although many computational methods have been developed to predict the functional consequences of mutations [7], it has been indicated that their utility is limited [28]. In this study, we applied an alternative approach to discover cancer drivers from genomic sequencing data. By overlapping cancer somatic mutations and well-defined pathogenic disease-causing germline variants in Mendelian diseases, we identified putative tumor suppressors and oncogenes, which warrant follow-up functional studies. Our analyses suggested that ion channels, collagens and Marfan syndrome-related genes may represent new classes of tumor suppressors. More significantly, mutations in two Marfan syndrome-related genes FBN1 and TGFBR2 are associated with poor prognosis in lung squamous cell carcinomas, providing novel biomarkers with potential clinical relevance in areas of prevention, diagnosis and treatment [29]. Although the previous report by Zhao and Pritchard [13] also interrogated overlapping pathogenic mutations in inherited diseases and cancer somatic mutations, we applied a novel approach to identify candidate tumor suppressors and oncogenes separately based on different criteria. Our approach is particularly useful in identifying the above highlighted putative tumor suppressors.

Table 3. Genes ranked by mutation frequency of the most recurrent HGMD-COSMIC overlapping mutation (at amino acid level) for each gene. Tumor samples with genome-wide sequencing data were used in the analysis. Only genes with the most recurrent amino acid change in at least 15 tumors are shown. TSG: tumor suppressor gene.

Gene	Mutation	Tumors	Oncogene/TSG	Gene	Mutation	Tumors	Oncogene/TSG
KRAS	p.G12D	524	Oncogene	TMEM106B	p.T185S	19	
IDH1	p.R132H	293	Oncogene	TAS2R43	p.H212R	19	
PIK3CA	p.H1047R	274	Oncogene	ROCK2	p.T431N	18	
TP53	p.R175H	226	TSG	PRNP	p.M129V	18	
APC	p.R1450*	66	TSG	GZMB	p.P94A	18	
PTEN	p.R130Q	42	TSG	PON2	p.S311C	17	
CDKN2A	p.R80*	40	TSG	KRT14	p.A94T	17	
CHEK2	p.Y390C	37		HNF1A	p.I27L	17	TSG
SMAD4	p.R361H	31	TSG	FGFR2	p.S252W	17	Oncogene
ABCD1	p.S606P	29		NRAS	p.G13D	16	Oncogene
KMT2C	p.T316S	26		IL1A	p.A114S	16	
OPRD1	p.C27F	25		HLA-DPB1	p.M105V	16	
PRDM9	p.T681S	24		EME1	p.I350T	16	
IDH2	p.R140Q	24	Oncogene	ALK	p.R1275Q	16	Oncogene
ARID1A	p.R1989*	24	TSG	ABCA1	p.R219K	16	
AR	p.Q58L	24	Oncogene	POU5F1B	p.E238Q	15	
UGT2A1	p.R75K	23		LTF	p.K47R	15	
PRSS1	p.K170E	23		IFIH1	p.A946T	15	
UGT1A7	p.N129K	22		HLA-A	p.L180*	15	
USH2A	p.C3416G	21		GRIN3B	p.T577M	15	
TGFB1	p.P10L	20		FGFR3	p.Y373C	15	Oncogene
RAD21L1	p.C90R	20		BRCA2	p.N372H	15	TSG
HRG	p.P204S	20		ATM	p.R337C	15	TSG

From our analyses, we rediscovered genes with cancer predisposing mutations, including TP53, APC, VHL, RB1 and many others (Table 2), which enhanced our confidence in the approach. However, as these genes have been well studied with respect to both germline mutations in familial cancer syndromes and somatic mutations in cancers, our focus lies on those genes with unknown connections between Mendelian diseases and specific cancers associated with the identical mutations. As genes often function differently in development versus in adult tissues, it is critical to further investigate the molecular pathways modulated by those genes in order to understand the mechanisms by which the same mutations can cause Mendelian diseases during development and drive tumor growth in adult tissues. This is best illustrated by an example in our oncogene discovery that revealed 5 HGMD/COSMIC overlapping mutations in ACVR1 gene cumulatively occurred in 19 central nervous system (CNS) cancers (data not shown). While the 5 ACVR1 mutations in germline cause fibrodysplasia ossificans progressiva (FOP), an autosomal dominant disorder of skeletal malformation and disabling heterotopic ossification [30], the same mutations are somatic oncogenic drivers in a subtype of CNS cancers, specifically diffuse intrinsic pontine glioma (DIPG) [31]. Functional studies have demonstrated the ACVR1 mutations in germline activate the canonical bone morphogenic protein (BMP) pathway to promote osteogenic differentiation and endochondral bone formation resulting in FOP, and the same BMP pathway activated by these mutations in astrocyte cells in the brain accelerates cell proliferation ultimately leading to malignancy [32]. Therefore, these seemingly unrelated two diseases involving different tissue and cell types might be connected by the same molecular pathway activated by identical mutations in germline or in somatic cells. As described in the results section, two of these five mutations are also present in endometrium cancers, and they are largely mutual exclusive from the most frequently mutated oncogenes (Figure 5), suggesting that deregulated activation of the BMP pathway in uterus epithelial cells is likely a key oncogenic mechanism in at least some cases of endometrium cancers. Interestingly, the ACVR1 mutations and their potential oncogenic roles in endometrium cancers were also discussed in a recent study [13].

We recognize the limitations in our study. Since the percentage of cancer somatic mutations overlapping with germline pathogenic variants is small (0.6% of somatic mutations, 1.6% of recurrent somatic mutations in COSMIC; Table 1), our approach will not be applicable to the majority of the somatic mutation data from cancer genomic sequencing. Furthermore, identification of putative oncogenes based on high recurrence and mutual exclusivity from known oncogenes yielded few candidates. This is partly due to the fact that very few HGMD/COSMIC overlapping mutations have high recurrence in cancers. In addition, lack of mutual exclusivity with known oncogenes does not necessarily preclude the mutations as cancer drivers. Our goal was only to identify potentially novel cancer genes with high confidence. As more cancer genomic sequencing data become available in COSMIC, our approach will likely lead to the identification of additional putative oncogenes. Another limitation is that most of the candidate cancer genes from our analysis lack apparent functional connection to cancer development. This is somewhat expected due the inherent nature of our approach using Mendelian diseases pathogenic variants to aid novel cancer gene discovery. Accordingly, our study demonstrates a powerful technique for hypothesis generation to identify associations that warrant further experimental validation.

Acknowledgments

We thank Robert Maki from Icahn School of Medicine at Mount Sinai for valuable discussions. This work was supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

References

1. Chmielecki J, Meyerson M. *Annual review of medicine* **65**, 63-79 (2014).
2. Garraway LA, Lander ES. *Cell* **153**, 17-37 (2013).
3. Hudson TJ*, et al. Nature* **464**, 993-998 (2010).
4. Tomczak K, Czerwinska P, Wiznerowicz M. *Contemporary oncology (Poznan, Poland)* **19**, A68-77 (2015).
5. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. *Science (New York, NY)* **339**, 1546-1558 (2013).
6. Lawrence MS*, et al. Nature* **499**, 214-218 (2013).

7. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. *BMC genomics* **14 Suppl 3**, S7 (2013).
8. Lawrence MS*, et al. Nature* **505**, 495-501 (2014).
9. Peterson TA, Doughty E, Kann MG. *Journal of molecular biology* **425**, 4047-4063 (2013).
10. Friend SH*, et al. Nature* **323**, 643-646 (1986).
11. Nagy R, Sweet K, Eng C. *Oncogene* **23**, 6445-6470 (2004).
12. Melamed RD, Emmett KJ, Madubata C, Rzhetsky A, Rabadan R. *Nature communications* **6**, 7033 (2015).
13. Zhao B, Pritchard JR. *PLoS genetics* **12**, e1006081 (2016).
14. Forbes SA*, et al. Nucleic acids research* **43**, D805-811 (2015).
15. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. *Human genetics* **133**, 1-9 (2014).
16. Futreal PA*, et al. Nature reviews Cancer* **4**, 177-183 (2004).
17. Than BL*, et al. Oncogene*, (2016).
18. Xie C*, et al. Oncogene* **32**, 2282-2291, 2291.e2281-2287 (2013).
19. Zhang JT*, et al. Biochimica et biophysica acta* **1833**, 2961-2969 (2013).
20. Martins VL*, et al. Journal of the National Cancer Institute* **108**, (2016).
21. Loeys B*, et al. Human mutation* **24**, 140-146 (2004).
22. Mizuguchi T*, et al. Nature genetics* **36**, 855-860 (2004).
23. The Cancer Genome Atlas. *Nature* **489**, 519-525 (2012).
24. Stracquadanio G*, et al. Nature reviews Cancer* **16**, 251-265 (2016).
25. Imielinski M*, et al. The Journal of clinical investigation* **124**, 1582-1586 (2014).
26. Tomasson MH*, et al. Blood* **111**, 4797-4808 (2008).
27. Ulaganathan VK, Sperl B, Rapp UR, Ullrich A. *Nature* **528**, 570-574 (2015).
28. Miosge LA*, et al. Proceedings of the National Academy of Sciences of the United States of America* **112**, E5189-5198 (2015).
29. Iyengar P, Tsao MS. *Surgical oncology* **11**, 167-179 (2002).
30. Kaplan FS*, et al. Human mutation* **30**, 379-390 (2009).
31. Zadeh G, Aldape K. *Nature genetics* **46**, 421-422 (2014).
32. Taylor KR, Vinci M, Bullock AN, Jones C. *Cancer research* **74**, 4565-4570 (2014).

IDENTIFYING CANCER SPECIFIC METABOLIC SIGNATURES USING CONSTRAINT-BASED MODELS

A. SCHULTZ[1], S. MEHTA[1], C.W. HU[1], F.W. HOFF[2], T.M. HORTON[3], S.M. KORNBLAU[2] and A.A. QUTUB[1*]

[1]*Department of Bioengineering, Rice University,*
Houston, Texas 77005, U.S.A
**E-mail: aminaq@rice.edu*

[2]*Department of Leukemia, The University of Texas M.D. Anderson Cancer Center,*
Houston, Texas 77030, U.S.A

[3]*Department of Pediatrics, Baylor College of Medicine and Texas Children's Hospital,*
Houston, Texas 77030, U.S.A

Cancer metabolism differs remarkably from the metabolism of healthy surrounding tissues, and it is extremely heterogeneous across cancer types. While these metabolic differences provide promising avenues for cancer treatments, much work remains to be done in understanding how metabolism is rewired in malignant tissues. To that end, constraint-based models provide a powerful computational tool for the study of metabolism at the genome scale. To generate meaningful predictions, however, these generalized human models must first be tailored for specific cell or tissue sub-types. Here we first present two improved algorithms for (1) the generation of these context-specific metabolic models based on omics data, and (2) Monte-Carlo sampling of the metabolic model flux space. By applying these methods to generate and analyze context-specific metabolic models of diverse solid cancer cell line data, and primary leukemia pediatric patient biopsies, we demonstrate how the methodology presented in this study can generate insights into the rewiring differences across solid tumors and blood cancers.

Keywords: Genome-scale metabolic reconstructions, constraint-based models, tissue-specific models, Flux Balance Analysis, cancer metabolism.

Introduction

Cancer tissues exhibits significant metabolic differences when compared to their healthy counterparts, such as the *Warburg effect*[1] and *glutamine addiction*.[2] In recent years it has been revealed that these metabolic transformations are largely driven by oncogenes and subdued by tumor suppressor genes.[3,4] This regulation suggests that cancer metabolism plays an important role in tumor progression, as opposed to being a consequence of the tumor microenvironment.[5] These findings have led to a renewed interest in the field of cancer metabolism,[6] with particular interest in exploiting metabolic differences as therapeutic targets.[7] Cancer metabolism, however, is also extremely heterogeneous across cancer types,[8] and treatments targeting metabolic pathways need to be carefully tailored to specific cancer phenotypes. Consequently, a better understanding of the metabolic differences across cancer sub-types, and between healthy and cancerous tissues will greatly assist the development of novel therapeutic strategies.[7,8]

Genome-Scale Models: To help elucidate the metabolic differences between cancer and healthy tissues, computational approaches can be extremely helpful. In particular, genome-scale models (GEMs) have proven extremely useful in studying human metabolism at the genome level,[9,10] with many studies dedicated specifically to cancer metabolism.[11–13] These studies have, for example, identified glycosaminoglycans as a marker for clear cell renal cell carcinoma,[14] identified carnitine palmitoyltransferase 1 as a potential target for hepatocellular carcinoma,[15] and identified MLYCD as a potential target for leukemia and kidney cancer.[16]

GEMs are defined at the core by a *stoichiometric matrix* S, where each row corresponds to a metabolite, each column to a metabolic reaction, and each entry to the stoichiometric coefficient of that particular metabolite in that particular reaction.[17] For any given *stoichiometric matrix*, flux distribution column vectors (v) can be defined where each element v_i gives the metabolic flux (e.g. rate of metabolite conversion) through each reaction i. The matrix multiplication $S \cdot v = m$ then yields a vector m where each element m_j gives the rate of change of concentration of metabolite j given the reaction fluxes defined by v. A steady-state flux distribution is one where $S \cdot v = 0$. A more detailed description of the constraint-based model formulation is available in the *supplemental information*.

Metabolic Model Analysis: Although a wide array of methods have been developed to study GEMs,[18] many of them are dependent on an *objective function*, which is most often assumed to be cellular growth.[19] Mammalian cells, however, do not have a well established objective, and do not seek to optimize biomass production. One prominent unbiased and objective-independent method for GEM analysis, suited for the study of mammalian cells, is Monte-Carlo sampling (MCS). This method finds normally distributed steady-state flux distributions inside the solution space of $S \cdot v = 0$ defined by lower (lb) and upper (ub) reaction bounds, such that $lb_i \leq v_i \leq ub_i$. Valuable insight into the metabolic capabilities of the model in question can be obtained by analyzing how different MCS conditions (e.g. different lower and upper bounds) affect the sampled reaction flux values. This approach has been used, for example, to model the metabolic exchange between *M. tuberculosis* and human macrophages,[20] and between different cell types in the human brain;[21] to study aspirin resistance in platelet cells;[22] and to characterize metabolic differences between healthy and cancerous tissues.[23]

Mammals also have a complex and compartmentalized metabolism, where not every metabolic reaction takes place in all cells of the body. In order to generate predictions specific to different cell types, cancer categories or patients, generalized human GEMs then need to be tailored to specific contexts.[24] We recently introduced the Cost Optimization Reaction Dependency Assessment (*CORDA*) tissue-specific algorithm,[23] which builds tissue-specific metabolic models based on omics data and a generalized human metabolic reconstruction. The algorithm is based on a *dependency assessment* (DA), where reactions associated with little experimental evidence, called negative confidence reactions (NC), are assigned an arbitrarily high cost. This cost is then minimized while enforcing a small flux through medium (MC) or high (HC) confidence reactions (i.e. reactions with medium or considerable experimental evidence) in order to identify which NC reactions are beneficial for MC or HC reactions to carry flux. This DA is then used to build a tissue-specific model including all HC reactions and as many MC reactions as possible, while minimizing the inclusion of NC reactions. For additional details on the original algorithm we refer readers to the original *CORDA* publication.[23]

Need for New Analyses: MCS of large metabolic networks is computationally expensive, and static approaches are only feasible for extremely small networks.[25] For MCS of higher dimensional networks, the Artificially Centered Hit and Run (ACHR) algorithm[26] is most frequently used. Given a set of points, or steady-state flux distributions, inside the solution space, ACHR calculates a center point as the average of all points, then moves each point i randomly along the directional vector defined by the trajectory between the center and another random point j. ACHR sampling of large networks can be extremely time consuming, however, and even small relative increments in computational efficiency can lead to fewer hours of computational time. Although alternatives to ACHR have been proposed, many of these methods are limited by sample distributions that are significantly different than ACHR outputs,[27–29] by their dependence on objective functions,[27] by long computational times,[30] or by lack of validation and parametrization in larger metabolic networks.[31]

Introduction of CORDA2 and mfACHR: Here we present two improved algorithms for the study of human GEMs. We first introduce an improved version of the *CORDA* algorithm to build tissue-specific metabolic models,[23] referred to here as *CORDA2*. *CORDA2* yields tissue models very similar to the ones given by the previous algorithm, but it is considerably faster than *CORDA* computationally. *CORDA2* is also noise-independent, thus providing unique model outputs for any given set of parameters, which facilitates the comparison of metabolic models across different modeling conditions (i.e. different cancer categories). We next introduce a new formulation of the ACHR algorithm,[26] referred to here as the matrix-form ACHR (mfACHR), which performs significantly faster than previous formulations.

Integrating the two new methods, we generate a panel of cell-line specific metabolic models using *CORDA2* and experimental data from the Human Protein Atlas[32] (HPA), and illustrate how flux samples generated using *mfACHR* can provide valuable insights into the metabolic profile of different cancer types, including pediatric leukemia. While we had previously shown that MCS of CORDA models can identify metabolic differences between healthy and cancerous tissues, here we show that this framework can also pinpoint metabolic differences between different cancer categories. The methods presented in this study provide significant advances in the generation and analysis of context-specific metabolic models.

Methods

Cost Optimization Reaction Dependency Assessment 2

In this work we present two modifications to *CORDA*, defining a new version of the algorithm referred to here as *CORDA2*. First, in the original algorithm, reversible reactions were split into forward and backward rates during every DA to ensure cost production regardless of directionality. That is, a reaction '$A \Leftrightarrow B$' was split into '$A \Rightarrow B + cost$' and '$B \Rightarrow A + cost$'. Since thousands of DAs are performed throughout the model building process, this modification was then repeated thousands of times during the algorithm. In *CORDA2*, this modification is performed at the beginning of the algorithm, and forward and backward rates are treated separately throughout the model building process, speeding the computational time. Furthermore, while in *CORDA* the reaction directionality in the tissue-model was imported from the generalized human reconstruction, *CORDA2* assigns directionality based on whether the forward, backward, or both reaction parts are included in the final tissue model.

Second, pathways with similar costs are captured in *CORDA* by adding a small amount of noise to reaction costs during every DA. This noise-driven approach leads to different

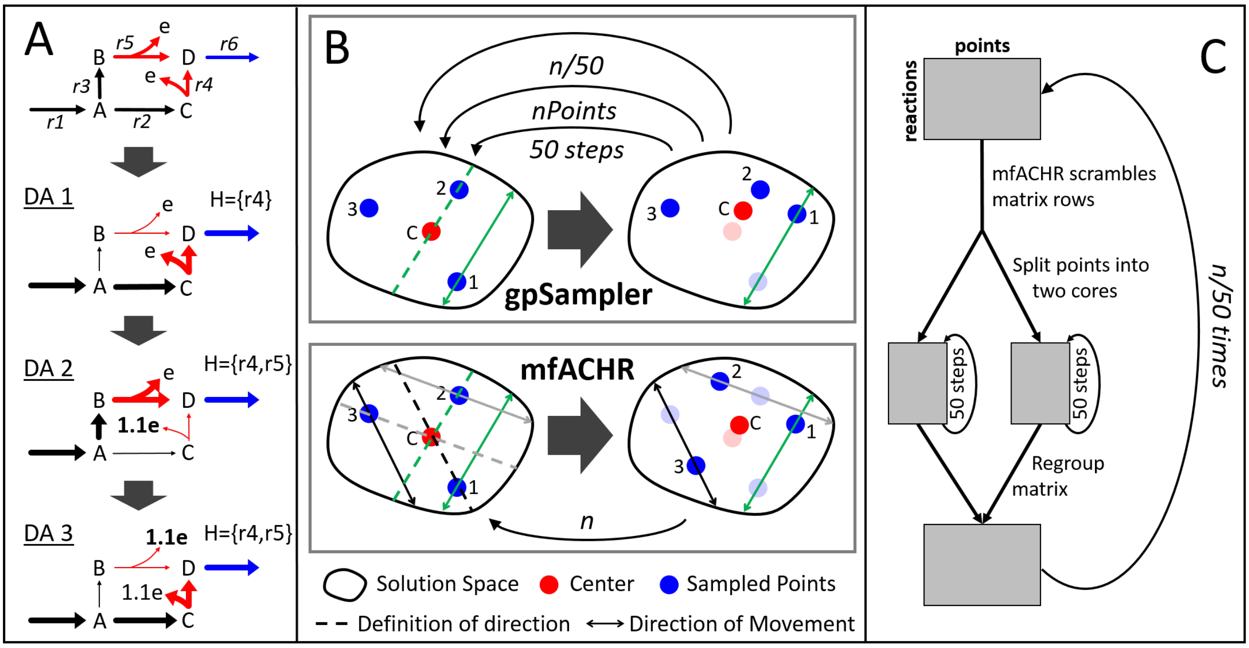


Fig. 1. **Representation of the CORDA2 and mfACHR algorithms**. (A) Identification of undesirable reactions (red) beneficial for the desirable reaction (blue) to carry flux through three DAs. Pathways taken during each DA are highlighted, and H represents the set of undesirable reactions taken up to that point. After an undesirable reaction is used, its cost (e) is increased. The process is repeated until H is unchanged. (B) gpSampler moves one point at a time, 50 steps at a time. The mfACHR algorithm identifies all possible directions of movement at once and moves all points simultaneously. Vectors defining the trajectory of movement, taken as the difference between j and the center point, and the corresponding path of movement of i are color-coded. (C) During parallelization of the MCS process, the matrix of sampled points is divided into 2 cores, which are sampled for 50 steps, then re-combined.

reconstructions after every run of the algorithm, and it is not guaranteed to include every alternative pathway. This approach is also inefficient since the same pathway can be sampled multiple times. In *CORDA2*, only undesirable reactions are assigned an arbitrarily high cost (while in *CORDA* all reactions received a basal cost value). This cost is then minimized during the DA, and the high cost reactions used are saved in a set H. The cost associated with the reactions in H is then increased, and the DA is performed again (**Fig. 1A**). This process is repeated iteratively until H is unchanged. This way, once a pathway is used, its cost is increased and another pathway with similar but now slightly lower cost is identified in the next DA. Additional details of the *CORDA2* formulation, as well as the MATLAB code for its implementation, can be found in the *supplemental information*.

Matrix-Form Artificially Centered Hit and Run

One of the most widely implemented ACHR formalisms is *gpSampler*.[33] *GpSampler* starts by moving a given point 50 steps as described by the ACHR algorithm, then repeats the process for each point being sampled. This whole process is then repeated $\frac{n}{50}$ times for a total of n ACHR steps (**Fig. 1B**). Here we propose a slightly different ACHR formulation, termed matrix-form ACHR (mfACHR). In *mfACHR*, all possible directions of movement are first calculated as the directional vectors defined by each sampled point and the center (dashed lines in **Fig. 1B**). These trajectories are then randomly assigned to each point, and each point is moved randomly along its assigned direction of movement (solid lines in **Fig. 1B**) within the bounds of the solution space. This whole process is repeated a total of n times for a desired number of steps. Both *gpSampler* and *mfACHR* can also be implemented in multiple cores. For that, the points being sampled are first divided into i groups, i being the number of cores used. Each group is then assigned to a core and mixed for 50 steps. All points are then re-combined and the process is repeated $\frac{n}{50}$ times for a total of n steps (**Fig. 1C**).

Cancer Cell Proteomics and Model Generation

Cell line gene and protein expression data were obtained from the HPA[32] in order to build the cell-line specific models. Gene expression data was measured using RNA-seq and protein expression was measured by immunohistochemistry using an extensive library of well validated antibodies. Forty-four models were generated using gene expression data and fifty-two models were generated using the proteomics data. Protein expression was available for 523 (35.0%) gene products, and gene expression data was available for 1,474 (98.7%) of the 1,494 unique genes in the generalized human reconstruction Recon1.[34] All gene and protein expression values were categorized into not detected, low/medium, and high expression in line with threshold values from the HPA, then used to categorize reaction confidence values used in the *CORDA2* algorithm. Following the reconstruction all models were sampled using *mfACHR*. Details of how these models were generated and sampled can be found in the *supplemental information*. For additional details on how the dataset was collected we refer readers to the HPA.[32]

Leukemia Patient Samples: Pediatric leukemia data was obtained from bone marrow biopsies of 95 acute myeloid leukemia (AML), 57 B-cell acute lymphoblastic leukemia (B-ALL), and 16 T-cell acute lymphoblastic leukemia (T-ALL) pediatric patients, and were collected at the Texas Children's Hospital. Protein expression level was measured using reverse phase protein array (RPPA) using 194 strictly validated antibodies.[35] Additional information on the pediatric leukemia data is available in the *supplemental information*.

Results and Discussion

Results of our study demonstrate the robustness of the *CORDA2* and *mfACHR* methods, and their utility in analyzing diverse cell line and primary leukemia cancer metabolism. A summary of the *CORDA2* and *mfACHR* validation is provided below, while a complete description of the algorithm validation and analysis is provided in the *supplemental information*.

CORDA2 Validation

In order to validate the *CORDA2* algorithm, outputs of this formulation were compared to 108 tissue-specific metabolic models generated using *CORDA* and similar model parameters (e.g. same dataset and overlapping algorithm parameters). Overall, at least 99.7% of MC reactions, 88.9% of NC reactions, and 93% of unclassified reactions included in each of the previous 108 models are also included in the *CORDA2* model, showing significant overlap between the output of both algorithms. Furthermore, *CORDA2* was approximately 2.5 times faster than *CORDA* when the later was performed with five DAs for every reaction tested. Although performing fewer DAs in *CORDA* led to computational times comparable to *CORDA2*, the reconstructions returned in that case are not as comprehensive. In the original *CORDA* publication, models reconstructed using one DA were on average 2.3% smaller than models built using multiple DAs. The *CORDA2* algorithm also showed very similar results across multiple metabolic tests when compared to the previous formulation. This analysis shows that *CORDA2* yields models similar to *CORDA* in composition and behavior, while being faster and noise independent.

mfACHR *Validation*

To assess the performance of *mfACHR* when compared to *gpSampler*, flux distributions and convergence speed of both formulations were compared for three different metabolic models: a red blood cell (RBC) model,[36] a platelet model,[22] and the generalized human reconstruction Recon1.[34] These models have 453, 1,008, and 2,473 active reactions respectively, and were sampled for $3{\cdot}10^4$, $7{\cdot}10^4$, $3{\cdot}10^5$ steps respectively. As an initial step in this validation, MCS outputs of four algorithm formulations (*mfACHR*, *mfACHR* parallel, *gpSampler*, and *gpSampler* parallel) were compared, and all four formulations were shown to converge to similar steady states (*supplemental information*).

Next, convergence speed was assessed by computational time and number of algorithm steps. Convergence based on number of steps was measured as the percentage of reactions at any given point with a Kullback-Leibler divergence (KLD) of sampled flux values below 0.05 of the final distribution. KLD represents the expected logarithmic difference between two probability distributions, and it has been previously used with a similarity threshold of 0.05 to compare sets of sampled flux distributions in metabolic models.[31] The four tested formula-

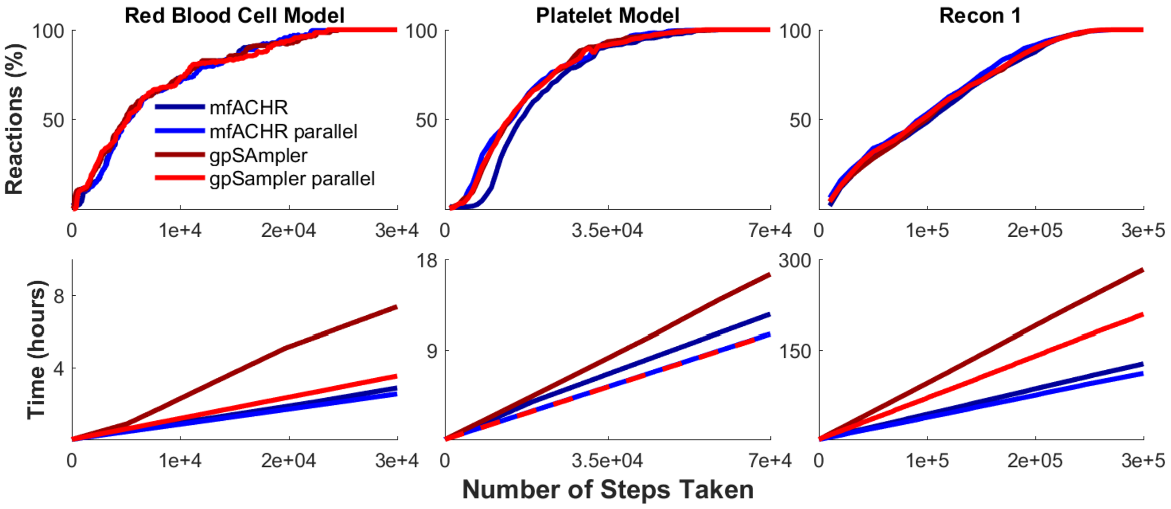


Fig. 2. **Conversion speed of *mfACHR* and *gpSampler*.** (Top) Percentage of reactions in the model with a KLD below 0.05 when compared to the final set of sampled points. (Bottom) Computational running time per number of algorithm steps.

tions showed nearly identical conversion curves when considering the number of steps taken (**Fig. 2**). When considering computational times, *mfACHR* performed significantly better than *gpSampler* when both methods were performed without parallelization. When considering parallelization, *mfACHR* showed very similar computational times in the platelet model, slightly better times in the RBC model, and significantly better times in Recon1. Differences in computational time can be partially attributed to the fact that matrix operations performed by *mfACHR* are automatically parallelized in MATLAB, while the *for* loops performed by *gpSampler* are not. This allows for *mfACHR* to perform significantly faster than *gpSampler* even when the latter is performed with parallelization, and explains the low relative increase in efficiency when explicit parallelization is implemented in *mfACHR*. Overall, *mfACHR* showed consistently faster computational times when compared to *gpSampler*, often in the order of hours, while converging at the same speed in terms of number of algorithm steps.

Cancer Cell Models

Following the validation of both algorithms, a series of cell-line specific models were generated using *CORDA2* and sampled using *mfACHR*, as described in the methods section. Twenty-six of the cancer metabolic models were combined into four tissue categories as presented in the HPA: myeloid, lymphoid, brain, and female reproductive system (FRS) cancer cell lines. These cancer types were chosen since they had the most number of cell lines. We then identified metabolic reactions that have significantly different sampled flux distributions between the four cancer categories (**Fig. 3**). MCS of *CORDA* models previously highlighted metabolic differences between healthy and cancerous tissues.[23] That is, using *CORDA* we correlated high sampled flux values with metabolic pathways known to take place in healthy or cancerous phenotypes. Analogously, in this study we demonstrate that *mfACHR* sampling of *CORDA2* models generated using HPA expression data can also highlight metabolic characteristics between different cancer categories. These characteristics include:

Brain tumors produce high levels of triglyceride: Lipid synthesis is an important factor for cancer survival and progression, and it has been previously suggested as a therapeutic target.[37–40] However, while most cancer types divert fatty-acids predominantly towards the production of phospholipids, not triglycerides,[39,41] glioma cells have been shown to synthesize triacylglycerol at high rates for membrane complex lipids.[42,43] Glioma cells, as well as healthy astrocytes and neurons, can also produce fatty acids from ketone-bodies,[44,45] a metabolic characteristic of brain cells which can further explain the high rate of fatty acid production in glioma cells. In the MCS results presented here, brain tumors present a significantly higher flux through glycerol-3-phosphate acyltransferase (**Fig. 3**) and 1-acylglycerol-3-phosphate O-acyltransferase, enzymes responsible for triacylglycerol synthesis.

Brain and lymphoid tumors have highly active glutamine metabolism: Glutamine plays an essential role in cancer metabolism,[46,47] and different tumors have been shown to utilize glutamine differently.[47] Brain tumors, in particular, have been shown to accumulate glutamine both *in vitro* and *in vivo*.[48,49] Glutamine metabolism has also been shown to play an important role in lymphoid tissues.[50] The role of this pathway in breast cancer, on the other hand, is not well defined, since basal but not luminal breast cancer cells show glutamine-

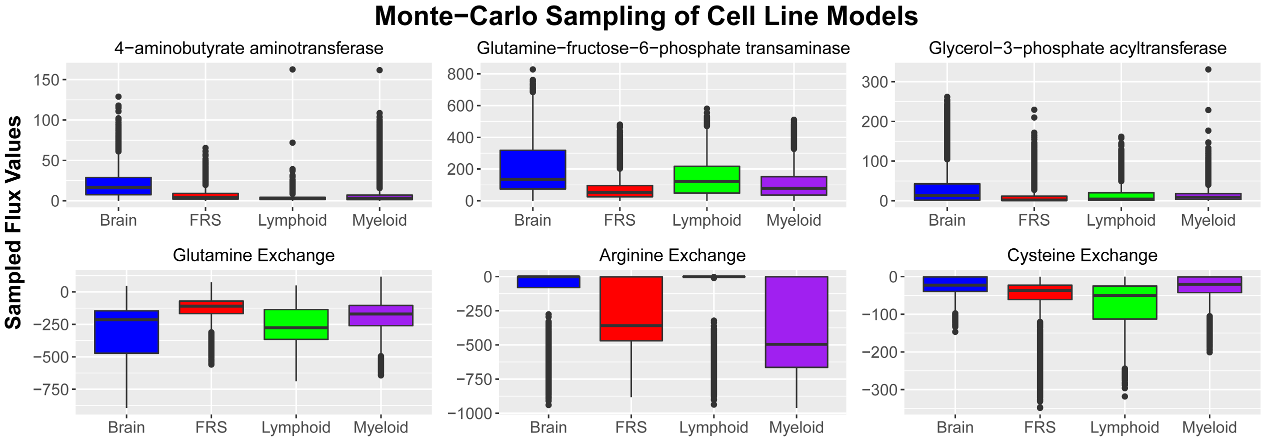


Fig. 3. **MCS results**. Sampled flux values for six different reactions across four model categories. Boxplots represent combined flux values for a particular reaction in all models in that cancer category. For exchange reactions, negative values represent uptake of the particular metabolite, while positive values represent secretion. Colored boxes represent values within the interquartile range (IQR), ranging from the 25th to the 75th percentile. Horizontal line represented the median value (50th percentile), and vertical lines indicate values within 1.5 IQR of the 25th and 75th percentiles. Outliers are represented by dots.

dependence.[51] In the results presented here, brain and lymphoid cell lines show high levels of glutamine uptake, while cell lines of the FRS show relatively low levels (**Fig. 3**).

Lymphoid tissues are cysteine dependent: While cysteine is not considered an essential amino-acid, lymphoid tumors have been shown to contain much lower levels of cystathionase, the last enzyme in the cysteine production pathway, when compared to healthy lymphoid tissues, and are dependent on cysteine for growth.[52] Targeting cysteine transporters has also been shown to selectively target lymphoma cells,[53] and cysteine uptake has been associate with malignant progression in lymphoma cells.[54] In this study, lymphoid models presented much higher levels of cysteine uptake (**Fig. 3**).

Tumors show different levels of arginine dependence: Different types of cancer respond differently to arginine deprivation.[55] A study performed on 26 healthy and cancerous cell lines found that tumor cells are much more sensitive to arginine deprivation than healthy cells.[56] Furthermore, while premyelocytic and lymphoblastic leukaemia cell lines die in about two days of arginine deprivation, cell lines of the FRS died largely in three to four days, and glioma cell lines died in four to five days.[56] Interestingly, levels of arginine dependence presented in the study by Scott et. al.[56] correspond to sampled flux values of arginine uptake in the present study. Myeloid cancers, the most arginine dependent, were predicted to uptake the largest amounts of arginine, followed by models of the FRS, then brain tumors, the least arginine dependent. Acute myeloid leukemia tumors have also been shown to be dependent on arginine for proliferation.[57]

Brain tumors were also predicted to have higher fluxes through the enzyme glutamine-fructose-6-phosphate transaminase (GF6PTA) (**Fig 3**), the rate limiting step in the hexosamines synthesis pathway (HSP), a nutrient sensor pathway.[58,59] When excess nutrients such as glucose and free fatty-acids are available, the HSP prevents cells from uptaking excess amounts from the bloodstream.[60] Furthermore, overweight and obese patients, which have excess amounts of nutrients in the bloodstream, are at an overall increased risk of mortality due to cancer.[61] Interestingly, sampled flux values through the HSP presented here are anti-correlated with the increase in risk of mortality in cancer patients. According to a study of over 57,000 cancer patients, obese patients with brain tumors have a modest increase in mortality compared to non-obese glioma patients, while patients with cancer of the FRS have a high increase in risk, and patients with Non-Hodgkins lymphoma, multiple myeloma, and leukemia have a medium increase.[62] Accordingly, brain tumor models in this study present high GF6PTA flux values, while tumors of the FRS present low fluxes, and lymphoid and myeloid tumors present intermediate values (**Fig. 3**). One possible explanation for this correlation is that higher fluxes through the HSP can prevent cells from uptaking excess amounts of nutrients, which in turn leads to a lower relative increase in malignancy. Further work should help elucidate these observations in context.[63,64]

Sampled flux values also predict a high flux through the enzyme 4-aminobutyrate aminotransferase in brain cancer cells. This result is expected since this enzyme is responsible for GABA production, a pathway highly active in brain tissues. In brain cancer cells, however, this enzyme can help produce acetyl-CoA for energy production, since larger amounts of nutrients are diverted away from glycolysis and into the HSP. A diagram of this proposed mechanism is presented in **Fig 4A**.

Primary pediatric leukemia models: We next analyzed sampled flux values in three different types of leukemia blood sample models (AML, T-ALL, and B-ALL) to clinical pro-

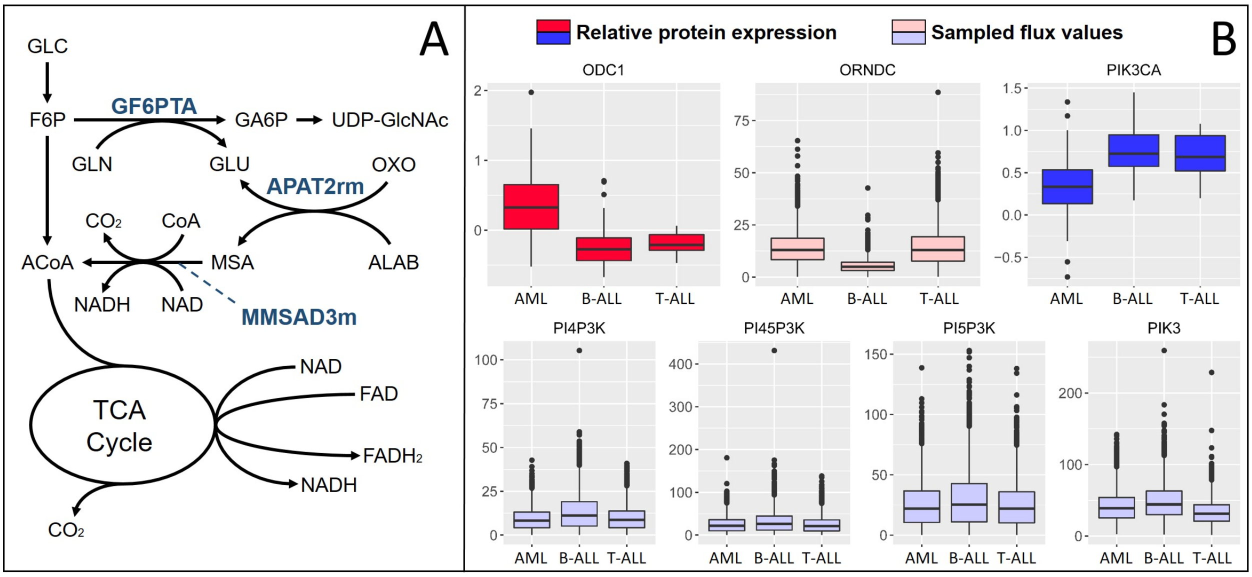


Fig. 4. **Model Predictions**. (A) Pathways with increased activity in brain tumors. Metabolites are glucose (GLC), fructose-6-phosphate (F6P), acetyl-CoA (ACoA), glutamine (GLN), glutamate (GLU), glucosamine-6-phosphate (GA6P), Uridine diphosphate N-acetylglucosamine (UDP-GlcNAc), oxoglutarate (OXO), beta-alanine (ALAB), and malonate semialdehyde (MSA). (B) Relative protein expression and sampled flux values for proteins differentially expressed between AML and ALL pediatric patients. ODC1 participates in the reaction Ornithine Decarboxylase (ORNDC), and PIK3CA participates in reactions PI4P3K, PI45P3K, PI5P3K, and PIK3. All reactions are labeled as in the BiGG database.[65]

teomics data collected from 168 pediatric leukemia patients as described in the methods section. Seven proteins were present both in the leukemia blood sample models and the clinical dataset, of which two were significantly differentially expressed between AML and ALL patients. The relative protein expression of these two proteins, along with the sampled flux values of reactions associated with these proteins, are presented in **Fig 4B**.

Sampled flux values follow trends that correlate with protein expression in both the B-ALL and AML models. That is, while AML patients show significantly higher expression levels of ODC1, the AML model showed significantly higher fluxes through Ornithine Decarboxylase (ORNDC), an ODC1 participating reaction, when compared to the B-ALL model. Likewise, while AML patients showed significantly lower expression of PIK3CA, the AML model also showed significantly lower sampled flux values through the PIK3CA reactions (**Fig 4B**). Sampled flux values between the AML and T-ALL model did not seem to match the differential protein expression, however. One possible explanation for this is the fact that there were considerably fewer T-ALL patients in the clinical dataset, and fewer T-ALL samples were used to generate the proteomics data used in the models building process (2 compared to 3 B-ALL and 4 AML). For instance, in the HPA, T-ALL ODC1 and PIK3CA protein scores are in between B-ALL and AML values, as opposed to much closer to B-ALL values like we see in the pediatric clinical data. This first example application to integrating RPPA leukemia data with metabolic pathway analysis demonstrates how *CORDA2* and *mfACHR* can also be used to analyze clinical data and provide insight into patient-specific metabolic behaviors.

Conclusion

This work illustrates how Monte-Carlo sampling of metabolic models generated using *CORDA2* can generate valuable predictions about context specific cancer metabolism. In applying these new optimized methods to different cancer systems, we show how this work goes beyond the identification of metabolic differences between healthy and cancerous tissues. It identifies differences in metabolism between different cancer types, paving the way to patient-specific metabolic models of cancer. In sum, the *CORDA2* platform elucidates metabolic differences across cancers and provides valuable knowledge of context-specific metabolic behavior that can help guide future directed cancer therapies.

Acknowledgments

This work was funded by NSF grant numbers 1354390 and 1150645, and NIH grant numbers GM106027 and CA164024.

Author Contributions

AS and AAQ developed the computational methods. AS, SM, and AAQ applied and validated the methods. TMH and SMK collected the AML and ALL clinical data. FWH and CWH processed the clinical data.

Supplemental Information

Supplemental files are available at www.qutublab.org/psb

References

1. O. Warburg *et al.*, *Science* **123**, 309 (1956).
2. H. Eagle, *Science* **122**, 501 (1955).
3. R. J. DeBerardinis, N. Sayed, D. Ditsworth and C. B. Thompson, *Current opinion in genetics & development* **18**, 54 (2008).
4. R. D. Michalek and J. C. Rathmell, *Immunological reviews* **236**, 190 (2010).
5. C. Munoz-Pinedo, N. El Mjiyad and J. Ricci, *Cell death & disease* **3**, p. e248 (2012).
6. R. A. Cairns, I. S. Harris and T. W. Mak, *Nature Reviews Cancer* **11**, 85 (2011).
7. M. G. Vander Heiden, *Nature reviews Drug discovery* **10**, 671 (2011).
8. J. R. Cantor and D. M. Sabatini, *Cancer discovery* **2**, 881 (2012).
9. A. Bordbar and B. O. Palsson, *Journal of internal medicine* **271**, 131 (2012).
10. A. Mardinoglu and J. Nielsen, *Journal of internal medicine* **271**, 142 (2012).
11. K. Yizhak, B. Chaneton, E. Gottlieb and E. Ruppin, *Molecular systems biology* **11**, p. 817 (2015).
12. I. Goldstein, K. Yizhak, S. Madar, N. Goldfinger, E. Ruppin and V. Rotter, *Cancer Metab* **1**, 10 (2013).
13. C. Frezza, L. Zheng, O. Folger, K. N. Rajagopalan, E. D. MacKenzie, L. Jerby, M. Micaroni, B. Chaneton, J. Adam, A. Hedley *et al.*, *Nature* **477**, 225 (2011).
14. F. Gatto, N. Volpi, H. Nilsson, I. Nookaew, M. Maruzzo, A. Roma, M. E. Johansson, U. Stierner, S. Lundstam, U. Basso *et al.*, *Cell reports* **15**, 1822 (2016).
15. R. Agren, A. Mardinoglu, A. Asplund, C. Kampf, M. Uhlen and J. Nielsen, *Molecular systems biology* **10**, p. 721 (2014).
16. K. Yizhak, E. Gaude, S. Le Dévédec, Y. Y. Waldman, G. Y. Stein, B. van de Water, C. Frezza and E. Ruppin, *Elife* **3**, p. e03641 (2014).
17. J. D. Orth, I. Thiele and B. Ø. Palsson, *Nature biotechnology* **28**, 245 (2010).
18. N. E. Lewis, H. Nagarajan and B. O. Palsson, *Nature Reviews Microbiology* **10**, 291 (2012).
19. A. M. Feist and B. O. Palsson, *Current opinion in microbiology* **13**, 344 (2010).
20. A. Bordbar, N. E. Lewis, J. Schellenberger, B. Ø. Palsson and N. Jamshidi, *Molecular systems biology* **6**, p. 422 (2010).
21. N. E. Lewis, G. Schramm, A. Bordbar, J. Schellenberger, M. P. Andersen, J. K. Cheng, N. Patel, A. Yee, R. A. Lewis, R. Eils *et al.*, *Nature biotechnology* **28**, 1279 (2010).
22. A. Thomas, S. Rahmanian, A. Bordbar, B. Ø. Palsson and N. Jamshidi, *Scientific reports* **4** (2014).
23. A. Schultz and A. A. Qutub, *PLoS Comput Biol* **12**, p. e1004808 (2016).
24. S. R. Estévez and Z. Nikoloski, *Front. Plant Sci* **5**, 10 (2014).
25. N. D. Price, J. Schellenberger and B. O. Palsson, *Biophysical journal* **87**, 2172 (2004).
26. D. E. Kaufman and R. L. Smith, *Operations Research* **46**, 84 (1998).
27. N. Chaudhary, K. Tøndel, J. Puchałka, V. A. M. dos Santos and R. Bhatnagar, *Molecular BioSystems* (2016).
28. W. Megchelenbrink, M. Huynen and E. Marchiori, *PloS one* **9**, p. e86587 (2014).
29. S. Bordel, R. Agren and J. Nielsen, *PLoS Comput Biol* **6**, p. e1000859 (2010).
30. P. A. Saa and L. K. Nielsen, *Bioinformatics* , p. btw132 (2016).
31. D. De Martino, M. Mori and V. Parisi, *PloS one* **10**, p. e0122670 (2015).
32. M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund *et al.*, *Science* **347**, p. 1260419 (2015).
33. J. Schellenberger, R. Que, R. M. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, S. Rahmanian *et al.*, *Nature protocols* **6**, 1290 (2011).
34. N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas and B. Ø. Palsson, *Proceedings of the National Academy of Sciences* **104**, 1777 (2007).

35. T. M. Horton, Y. Qiu, G. Jenkins and S. M. Kornblau, *Blood* **124**, 3784 (2014).
36. A. Bordbar, N. Jamshidi and B. O. Palsson, *BMC systems biology* **5**, p. 110 (2011).
37. T. Mashima, H. Seimiya and T. Tsuruo, *British journal of cancer* **100**, 1369 (2009).
38. J. A. Menendez and R. Lupu, *Nature Reviews Cancer* **7**, 763 (2007).
39. F. P. Kuhajda, *Cancer research* **66**, 5977 (2006).
40. T. Migita, S. Okabe, K. Ikeda, S. Igarashi, S. Sugawara, A. Tomida, R. Taguchi, T. Soga and H. Seimiya, *The American journal of pathology* **182**, 1800 (2013).
41. F. P. Kuhajda, K. Jenner, F. D. Wood, R. A. Hennigar, L. B. Jacobs, J. D. Dick and G. R. Pasternack, *Proceedings of the National Academy of Sciences* **91**, 6379 (1994).
42. H. Cook and M. Spence, *Canadian Journal of Biochemistry and Cell Biology* **63**, 919 (1985).
43. H. W. Cook and M. W. Spence, *Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism* **918**, 217 (1987).
44. M. S. Patel, J. J. Russell and H. Gershman, *Proceedings of the National Academy of Sciences* **78**, 7214 (1981).
45. L. M. Roeder, S. E. Poduslo and J. T. Tildon, *Journal of neuroscience research* **8**, 671 (1982).
46. R. J. DeBerardinis and T. Cheng, *Oncogene* **29**, 313 (2010).
47. C. T. Hensley, A. T. Wasti and R. J. DeBerardinis, *The Journal of clinical investigation* **123**, 3678 (2013).
48. E. A. Maher, I. Marin-Valencia, R. M. Bachoo, T. Mashimo, J. Raisanen, K. J. Hatanpaa, A. Jindal, F. M. Jeffrey, C. Choi, C. Madden *et al.*, *NMR in biomedicine* **25**, 1234 (2012).
49. I. Marin-Valencia, C. Yang, T. Mashimo, S. Cho, H. Baek, X.-L. Yang, K. N. Rajagopalan, M. Maddie, V. Vemireddy, Z. Zhao *et al.*, *Cell metabolism* **15**, 827 (2012).
50. M. Ardawi and E. Newsholme, Glutamine metabolism in lymphoid tissues, in *Glutamine metabolism in mammalian tissues*, (Springer, 1984) pp. 235–246.
51. H.-N. Kung, J. R. Marks and J.-T. Chi, *PLoS Genet* **7**, p. e1002229 (2011).
52. J. Iglehart, R. M. York, A. P. Modest, H. Lazarus and D. Livingston, *Journal of Biological Chemistry* **252**, 7184 (1977).
53. P. Gout, A. Buckley, C. Simms and N. Bruchovsky, *Leukemia (08876924)* **15** (2001).
54. P. Gout, Y. Kang, D. Buckley, N. Bruchovsky and A. Buckley, *Leukemia* **11**, 1329 (1997).
55. D. N. Wheatley, *Seminars in Cancer Biology* **15**, 247 (2005).
56. L. Scott, J. Lamb, S. Smith and D. Wheatley, *British journal of cancer* **83**, p. 800 (2000).
57. F. Mussai, S. Egan, J. Higginbotham-Jones, T. Perry, A. Beggs, E. Odintsova, J. Loke, G. Pratt, A. Lo, M. Ng *et al.*, *Blood* **125**, 2386 (2015).
58. M. G. Buse, *American Journal of Physiology-Endocrinology And Metabolism* **290**, E1 (2006).
59. M.-J. J. Pouwels, C. J. Tack, P. N. Span, A. J. Olthaar, C. Sweep, F. C. Huvers, J. A. Lutterman and A. R. Hermus, *The Journal of Clinical Endocrinology & Metabolism* **89**, 5132 (2004).
60. L. Wells, K. Vosseller and G. Hart, *Cellular and Molecular Life Sciences CMLS* **60**, 222 (2003).
61. E. E. Calle and R. Kaaks, *Nature Reviews Cancer* **4**, 579 (2004).
62. E. E. Calle, C. Rodriguez, K. Walker-Thurmond and M. J. Thun, *New England Journal of Medicine* **348**, 1625 (2003).
63. T. N. Sergentanis, G. Tsivgoulis, C. Perlepe, I. Ntanasis-Stathopoulos, I.-G. Tzanninis, I. N. Sergentanis and T. Psaltopoulou, *PloS one* **10**, p. e0136974 (2015).
64. T. Niedermaier, G. Behrens, D. Schmid, I. Schlecht, B. Fischer and M. F. Leitzmann, *Neurology* **85**, 1342 (2015).
65. Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson and N. E. Lewis, *Nucleic acids research* **44**, D515 (2016).

DIFFERENTIAL PATHWAY DEPENDENCY DISCOVERY ASSOCIATED WITH DRUG RESPONSE ACROSS CANCER CELL LINES [*]

GIL SPEYER, DIVYA MAHENDRA[†], HAI J. TRAN[†], JEFF KIEFER

The Translational Genomics Research Institute
Phoenix, AZ 85004, U.S.A.
Email: gspeyer@tgen.org, mahendradivya@gmail.com, hjtran@brown.edu, jkiefer@tgen.org

STUART L. SCHREIBER, PAUL A. CLEMONS

Broad Institute of Harvard and MIT
Cambridge MA 02142, U.S.A.
Email: stuart_schreiber@harvard.edu, pclemons@broadinstitute.org

HARSHIL DHRUV, MICHAEL BERENS, SEUNGCHAN KIM

The Translational Genomics Research Institute
Phoenix, AZ 85004, U.S.A.
Email: hdhruv@tgen.org, mberens@tgen.org, skim@tgen.org

The effort to personalize treatment plans for cancer patients involves the identification of drug treatments that can effectively target the disease while minimizing the likelihood of adverse reactions. In this study, the gene-expression profile of 810 cancer cell lines and their response data to 368 small molecules from the Cancer Therapeutics Research Portal (CTRP) are analyzed to identify pathways with significant rewiring between genes, or differential gene dependency, between sensitive and non-sensitive cell lines. Identified pathways and their corresponding differential dependency networks are further analyzed to discover essentiality and specificity mediators of cell line response to drugs/compounds. For analysis we use the previously published method EDDY (Evaluation of Differential DependencY). EDDY first constructs likelihood distributions of gene-dependency networks, aided by known gene-gene interaction, for two given conditions, for example, sensitive cell lines vs. non-sensitive cell lines. These sets of networks yield a divergence value between two distributions of network likelihoods that can be assessed for significance using permutation tests. Resulting differential dependency networks are then further analyzed to identify genes, termed *mediators*, which may play important roles in biological signaling in certain cell lines that are sensitive or non-sensitive to the drugs. Establishing statistical correspondence between compounds and mediators can improve understanding of known gene dependencies associated with drug response while also discovering new dependencies. Millions of compute hours resulted in thousands of these statistical discoveries. EDDY identified 8,811 statistically significant pathways leading to 26,822 compound-pathway-mediator triplets. By incorporating STITCH and STRING databases, we could construct evidence networks for 14,415 compound-pathway-mediator triplets for support. The results of this analysis are presented in a searchable website to aid researchers in studying potential molecular mechanisms underlying cells' drug response as well as in designing experiments for the purpose of personalized treatment regimens.

[*] This work was supported in part by the National Cancer Institute, National Institutes of Health [1U01CA168397] and a grant from Dell, Inc. via its *Legacy of Good* program that seeks to put technology and expertise to work where it can do the most for people and the planet.

[†] D. Mahendra and H. Tran were supported by the Helios Education Foundation through the Helios Scholars at TGen summer internship program in biomedical research at the Translational Genomics Research Institute in Phoenix, AZ.

1. Introduction

The effort to personalize treatment plans for patients involves the identification of drug treatments that can effectively target the disease while minimizing the likelihood of adverse reactions. The advent of high-throughput –omics and drug-screening data has given rise to the development of complex analytical approaches to identify biomarkers and drug-targets) [1]. Considering complex molecular mechanisms underlying complex diseases such as cancer, the discovery of such biomarkers and subtype-specific drug targets must be based on activities of multiple genes rather than individual genes. Gene Set Enrichment Analysis (GSEA) [2] is one popular method of testing for differential expression of gene sets between conditions. As pathways are capable of complex rewiring between conditions, network-based analyses have become increasingly attractive for extraction of biological hypotheses from big data [3]. For example, the approaches to identify individual differential dependencies[‡] [4-8] or condition-specific sub-networks from genome-wide dependency networks such as a protein-protein interaction networks have gained much interest [9-11] for the determination of biomarkers and subtype-specific therapeutic vulnerabilities.

Recently, we developed a novel computational method *Evaluation of Differential DependencY* (EDDY) that identifies pathways enriched with differential dependencies and that discovers mediators as potential therapeutic targets. The method has been further improved by incorporating known gene interactions as prior knowledge. The method has been successfully applied to the study of glioblastoma (GBM) [12, 13] and adrenocortical carcinoma (ACC) [14].

In this study, we present results from an integrated analysis of large-scale transcriptomic data of 810 cancer cell lines and large-scale high-throughput screening data of the same cancer cell lines across 368 compounds using EDDY algorithm. The analysis not only identified the pathways enriched with differential dependencies between sensitive and non-sensitive cancer cell lines to each compound, but also discovered mediators as potential novel targets of the compound via graphical analysis of differential dependency networks. Identified compound-pathway-mediator triplets were further queried across known drug-gene database as well as a known gene-gene interaction database to identify corroborating evidence to support newly discovered compound-pathway-mediator triplets. We also developed a searchable website to aid researchers in studying potential molecular mechanisms underlying cells' drug response and in designing experiments for the purpose of personalized treatment regimens, publicly available at http://biocomputing.tgen.org/software/EDDY/CTRP.

2. Methods

2.1. *High-Throughput Drug Screening of Cancer Cell Lines*

The Cancer Cell Line Encyclopedia (CCLE) project is an effort to conduct detailed genetic characterization of a large panel of human cancer cell lines. The CCLE provides public access to DNA copy number, mRNA expression, and mutation data for 1,000 cancer cell lines,

[‡] In this manuscript, we use 'dependency' to denote statistical dependencies derived from data such as co-expression, conditional dependencies, and 'interaction' to denote *known* relationships between genes or related molecules.

encompassing 36 different tumor types [15].

The Center for the Science of Therapeutics at Broad Institute performed analysis of sensitivity of CCLE cell lines using ~500 small molecules as perturbagens, and made the data available at the Cancer Therapeutics Response Portal (CTRP; http://www.broadinstitute.org/ctrp/). The "Informer Set" consists of 481 small compounds, including 70 FDA approved drugs, 100 clinical candidates and 311 small-molecule probes. In this study, we used the transcriptomic profile and CTRP drug-response data to identify pathways with condition-specific rewiring of gene dependencies in the context of drug sensitivity [16, 17]. All of these aforementioned processed data is publicly available on the CTD^2 data portal (https://ctd2.nci.nih.gov/dataPortal/).

2.2. *EDDY: Evaluation of Differential Dependency*

EDDY is a statistical approach that combines pathway-guided and differential dependency analyses in a probabilistic framework [12, 13]. The algorithm queries each pathway (gene set) in a database such as BioCarta (http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) or REACTOME [18] to test for differential dependencies across the set of genes between two or more conditions, by comparing gene-dependency networks constructed for each condition. In evaluating differential dependency, EDDY uses a network likelihood distribution over multiple networks constructed via resampling for each condition and compares the distributions between the conditions, instead of just using the single, most probable network from each condition. The statistical significance of

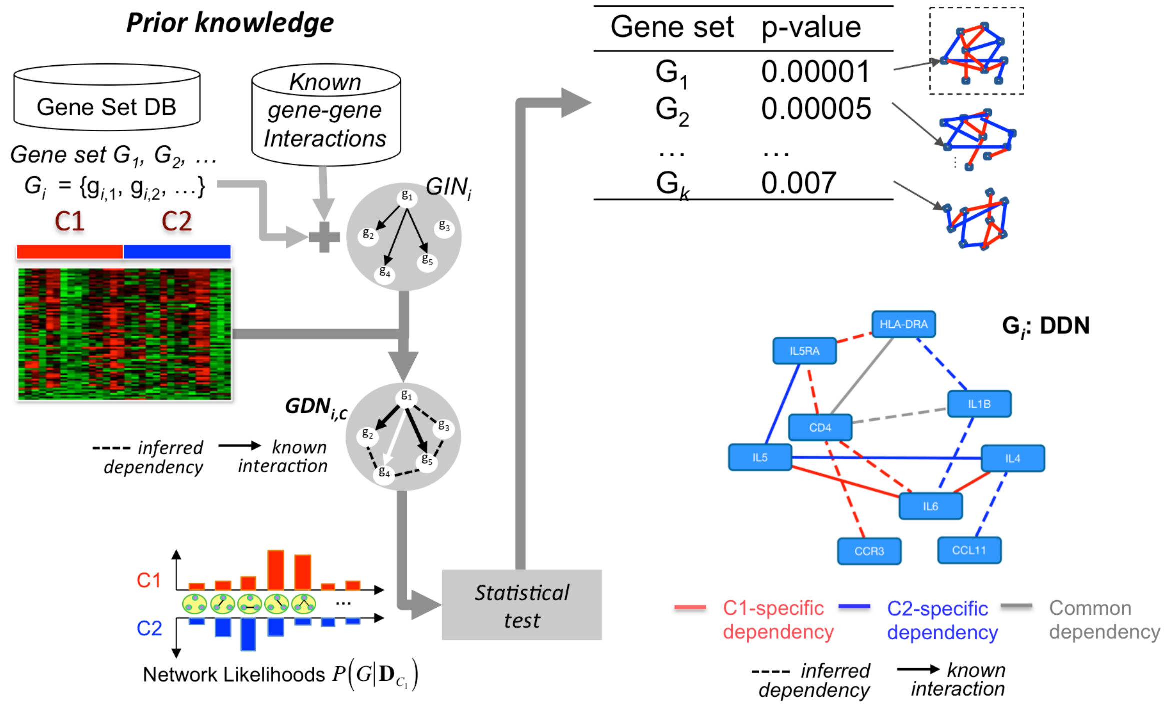


Figure 1. **Knowledge-assisted EDDY Workflow.** GDN$_{i,C}$ is a gene-dependency network constructed for a gene set G$_i$, for condition C, aided by gene interaction network GIN$_i$. A network likelihood distribution over multiple networks is constructed via resampling for each condition and the network score distributions between the conditions are compared. Permutation testing assesses the significance of the divergence between the distributions of scores. Differential dependency networks can then be constructed for statistically significant gene sets.

the divergence is then estimated using asymptotic approximation of Jensen-Shannon divergence based on a beta distribution whose parameters are estimated using a permutation test. Probabilistic and gene-set assisted approaches together contribute to significantly higher sensitivity and specificity of EDDY, compared to other methods, such as GSEA and Gene Set Co-expression Analysis (GSCA) [12].

Incorporation of Prior Knowledge into EDDY: Known interactions from the Pathway Commons 2 (http://www.pathwaycommons.org) database are integrated into EDDY as prior knowledge (Figure 1). This integration has been shown to improve the interpretability of results from EDDY. Prior weight (W_p) is specified to determine the degree of weight that is given to the prior knowledge in evaluating new edges to be included in the proposed dependency structure. Since prior knowledge is not condition-specific, large prior weight could decrease EDDY's sensitivity to detect differential dependency while reducing discovery of false-positive dependencies. For this analysis, a prior weight of $W_p = 0.5$ was used, meaning that any edges with half the support from data were included in the dependency network. The choice was based on extensive analysis of various data sets where $W_p = 0.5$ seemed to give the best compromise between sensitivity and false discovery rate when varying prior weight, as reported in Speyer et. al. [13].

2.3. *Input Data*

Transcriptomic data: BAM files of 935 CCLE cell lines downloaded from the Cancer Genomics Hub (https://cghub.ucsc.edu) were converted to a FASTQ format and transcript quantification was performed using Salmon [19] to obtain quantitative estimate of mRNA expression in TPM (transcripts per million). These mRNA expression values were $\log_2$ transformed and quantized to values -1 (under-expressed), 0 (intermediate), and 1 (over-expressed). For each gene, median average deviation (MAD) was computed and used to determine under-expression (MAD < -1), over-expression (MAD > 1), and intermediate.

Drug sensitivity: The cell lines were grouped into sensitive and non-sensitive classes using the Small-Molecule Cancer Cell Line Sensitivity Profiling CTRP 2.0 2015 Dataset, acquired from CTD^2 (Cancer Target Discovery and Development). CTRP summarizes drug sensitivity between each cell line and drug pair using the area-under-percent-viability-curve (AUC) values [16, 17]. We used the 'extremevalues' R package to identify outliers in AUC values and group the cell lines into sensitive (-1; lower-end outliers), non-sensitive (1; upper-end outliers), and intermediate (0; non-outliers) groups for each compound.

In order to conduct a statistically meaningful analysis using EDDY, only those drugs that had at least 50 samples in each sensitive and non-sensitive class were analyzed. This reduced the number of drugs that could be analyzed to 368 drugs.

2.4. *Identification of Mediators*

For each compound, the results from EDDY analysis (Figure 2) are summarized into 1) a list of pathways enriched with differential dependency of statistical significance, and 2) a differential dependency network (DDN) that captures how gene dependency changes between sensitive and non-sensitive cell lines. We identified those genes that seemed to play a significantly different role (based on statistical dependencies) between cell lines that were sensitive to a drug and cell lines that were non-sensitive, and termed them as *mediators*.

Essentiality mediators: Each DDN is split into condition-specific dependency networks (CDNs) where each CDN is composed of dependencies manifested in each condition. We then compute between-ness centrality for each gene in both CDNs and compute the difference of the betweenness centrality. The genes with the most differential betweenness centrality are termed *essentiality mediators*, as the genes with highest betweenness centrality in gene regulatory network are often interpreted as essential genes [20].

Specificity mediators: We also analyzed how many dependencies for each gene change between the CDN from sensitive cell lines and the CDN from non-sensitive cell lines. Formally, Let $P_C = E_C/(E_C + E_S)$, a proportion of condition-specific edges (E_C) across the overall number of edges ($E_C + E_S$), and E_{C_i} be the number of condition-specific edges and E_{S_i} be number of shared edges, of a gene i. Note $E_C = \sum_i E_{C_i}$ and $E_S = \sum_i E_{S_i}$. We can then compute the probability, $\Pr(k \geq E_{C_i})$, that a gene i can have E_{C_i} or more condition-specific edges by random chance, via binomial probability $B(k, E_{C_i} + E_{S_i}, P_C)$. If this probability, $\Pr(k \geq E_{C_i}) < 0.05$, we termed gene i as *specificity mediator*.

2.5. *Evidence Networks*

However, uncertainty in interpreting these drug-pathway-mediator triplets hinders prioritization of hypotheses or experimental design to explore these potentially valuable results. We address this challenge by constructing evidence networks built with protein and drug interactions from the STRING and STITCH interaction databases. STITCH and STRING are sister knowledge-bases that store scored drug-protein interactions and protein-protein interactions, respectively [21, 22]. As compounds can have multiple names, from commercial and generic labels to chemical formula

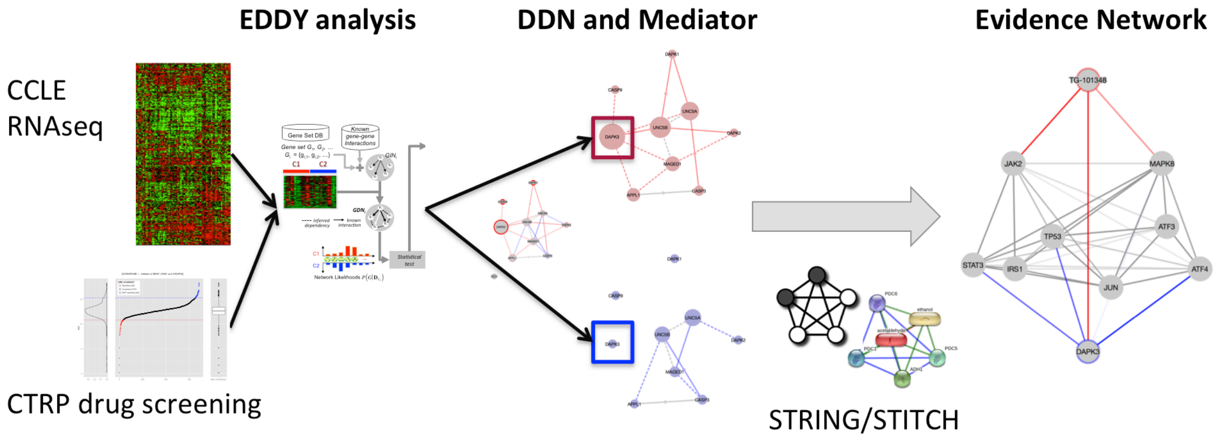


Figure 2. Overall workflow of EDDY analysis of CCLE and CTRP data. EDDY identifies significant pathways from RNA expression and compound-response categorization of cancer cell lines. Graphical analysis of output networks (edge color indicating condition) identifies important genes, termed mediators. Mining knowledge bases yields evidence networks for compound-mediator pairings (edge color here indicating evidence type).

and IUPAC ID, the database employed a unifying InChIKey to maximize comprehensiveness and to avoid false negatives.

Evidence networks were generated using a modified Yen's *K*-shortest paths algorithm [23] with a weight function of W(EDGE) = 1 – EDGE.SCORE, so that edges with higher scores would be preferred over edges with lower scores (all scores are within the interval [0,1] and are based on how compelling the supporting evidence is). To generate the evidence networks, shortest paths were continually found and added to the network until there were no more paths from the drug to the gene or there were at least *N* distinct nodes in the sub-network, where *N* is some arbitrary threshold. *N* was not a strict floor as sometimes the last path added to the sub-network would add two or more distinct nodes pushing the total number of distinct nodes over the threshold. Instead, *N* was used simply as a stopping condition and was chosen in order to prevent generation of evidence networks that would be too overwhelming for users to interpret. Choosing *N = 5* yielded abundant evidence nets without excessive density. Dijkstra's shortest-path algorithm with a Fibonacci heap was used as the supporting shortest-path algorithm in the modified Yen's *K*-shortest-paths algorithm [24, 25].

3. Results

3.1. *Pathway and Mediator Analysis*

EDDY analysis identified a total of 8,811 statistically significant pathways and 26,822 compound-pathway-mediator triplets. Of these, 534 pathways out of 685 BIOCARTA and REACTOME pathways were identified for at least one compound, and 2,401 genes out of 4,298 unique BIOCARTA and REACTOME genes were identified as mediators for at least one compound. On average each compound identified about 24 pathways and 73 mediators. We found that for 125 compounds, EDDY identified pathways that had the compound's intended target in their DDN, and 29 mediators were identified as intended targets. Only 248 out of the 368 compounds had intended targets that EDDY could potentially identify within the REACTOME and BIOCARTA pathways. Hence, EDDY identified pathways that included the intended target for 125 out of 248 compounds (50.4%). We tabulated (Table 1 & Table 2) the top 10 statistically significant pathways and mediators, respectively, which were identified by the largest number of compounds. We can see that the top two pathways that were statistically significant were ERYTH (erythrocyte

Table 1. Top 10 most commonly identified statistically significant pathways that were statistically significant

Pathway	# Compounds	Database
Erythrocyte differentiation (ERYTH)	78	BIOCARTA
Cells and molecules involved in local acute inflammatory response (LAIR)	61	BIOCARTA
CBL mediated ligand-induced downregulation of EGF receptors (CBL)	55	BIOCARTA
TERMINATION OF O GLYCAN BIOSYNTHESIS	52	REACTOME
SIGNALING BY HIPPO	49	REACTOME
NUCLEOTIDE LIKE PURINERGIC RECEPTORS	48	REACTOME
ZINC TRANSPORTERS	46	REACTOME
GRANULOCYTES	45	BIOCARTA
SYNTHESIS OF SUBSTRATES IN N GLYCAN BIOSYTHESIS	44	REACTOME
PURINE CATABOLISM	43	REACTOME

differentiation pathway) and LAIR (pathway for cells and molecules involved in local acute inflammatory response) from BIOCARTA. The erythrocyte differentiation pathway is the pathway responsible for the formation of red blood cells from the bone marrow. It is expected that this pathway would be altered in hematopoietic cancers and that its alteration would be involved in immune responses. The genes found in this pathway include TGFB2 and cytokines IL1A, IL3, IL6, IL9, and IL11. Cytokines are involved in various immune responses and inflammatory processes. The LAIR pathway includes mechanisms associated with the releases of cytokines IL1A and IL6. The genes IL1A and IL6 are among the top fourteen mediators identified by compounds in EDDY and they are also intended targets for the ERYTH and LAIR pathways. IL1A gene is a cytokine involved in various immune responses, inflammatory processes, and hematopoiesis. This protein is released in response to cell injury. IL6 is also a cytokine that functions in inflammation and maturation of B cells [26]. Indeed, upon further examination of the response data for the compounds differentially dependent for the ERYTH and LAIR pathways, hematopoietic cell lines were on average six times more prevalent in the sensitive versus the non-sensitive groups.

Table 2. The top 10 most commonly identified mediators

Pathway	# Compounds
MAPK1	185
MAPK3	171
GRB2	168
NUP210	158
HRAS	136
NUP37	125
AKT1	120
ORC4	114
MAPK14	114
CDK1	114

The MAPK signaling pathway is an important signaling pathway in cancer studies because it is altered in many different cancer types and regulates processes such as cell proliferation, cell differentiation, and cell death. MAPK1, MAPK3 and MAPK14 are mitogen-activated protein kinases and are members of the MAP kinase family. These genes act in signaling pathways (MAPK signaling, immune response) and various other cellular processes such as proliferation, differentiation, and cell cycle progression. MAPK14 is activated by environmental stresses and cytokines associated to inflammatory responses. MAP kinases play important roles in cascades of cellular responses and lead to direct activation of transcription factors [27].

3.2. *Evidence Network Analysis*

EDDY-CTRP analysis identified 26,822 drug-pathway-mediator triplets. Among these pairs, 19,222 of them consisted of a drug or a gene that is contained within the STRING and STITCH databases. Mining STITCH and STRING for each of 19,222 unique compound-pathway-mediator triplets yielded 14,415 evidence networks (~75%) of a path with 3 or fewer intermediate genes. These evidence networks are integrated into the main EDDY-CTRP portal as searchable tables (Table 3).

Table 3. Distribution of the number of intermediate genes in shortest path between drug and mediator pair.

	Direct targets	Indirect targets		
		# of intermediate genes in shortest path		
		1	2	3
# of pairs	102	988	3,410	9,915

We note that 102 evidence networks indeed were direct compound and mediator relations, among which only 34 were intended targets defined in the CTRP data and annotation. This indicates

STITCH/STRING contain drug-target relations that were not included in the CTRP database, but EDDY-CTRP analysis was able to discover those relations. Most of these evidence networks were for drug-pathway-mediator triplets where mediators were not direct targets of drug but had some known functional association to the drug (based on STITCH/STRING database). Note that known “hub” genes such as TP53 turned out to have high prevalence in the constructed evidence networks. In future development, the algorithm will introduce weighting to counter this bias.

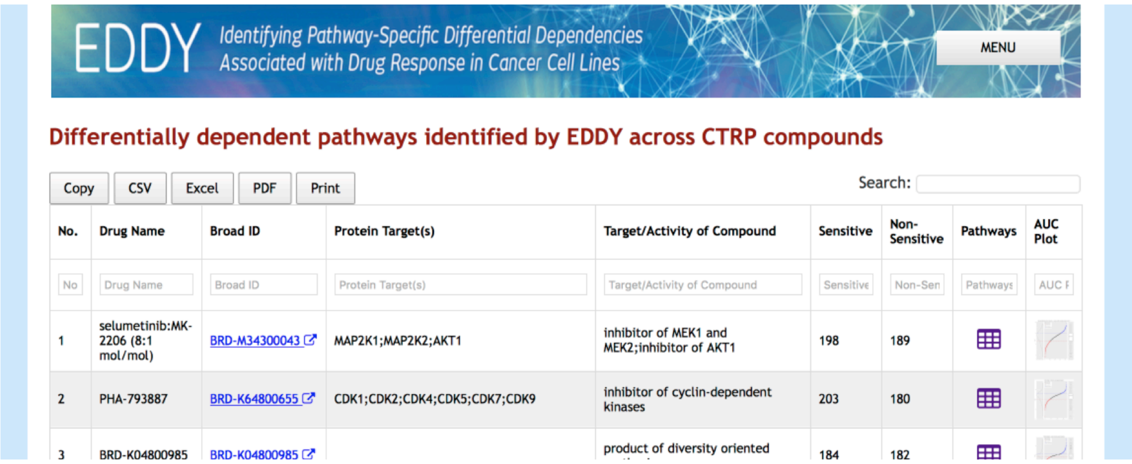
3.3. *Interactive and Searchable Web-Portal for EDDY-CTRP Results*

The web-portal of the CTRP analysis (http://biocomputing.tgen.org/software/EDDY/CTRP) consists of two main views: CTRP compound-centric and mediator-centric. These views provide alternate perspectives on hypothesis-testing data from the EDDY analysis. CTRP compound-centric view (Figure 3) provides pathways enriched with differential dependencies for each of 368 compounds uncovered by EDDY. For each compound, a user can explore each identified pathway, corresponding DDNs, and mediators. Mediator-centric view (Figure 4) lists all compound-pathway-mediator triplets uncovered across all compounds and all identified pathways. For each triplet, a user can also explore evidence networks as well as corresponding DDNs and pathways.

4. Case Studies: Potential Alternative Drug Targets

4.1. *DAPK3 as an Alternative Target for TG-101348*

TG-101348 was developed as a selective inhibitor of JAK2 kinase for the treatment of myeloproliferative disorder [28]. EDDY identified 29 pathways significantly enriched with differential dependency, and 66 mediators. One of the pathways is the EPONFKB pathway, which has JAK2 as an identified mediator, and, examining this DDN, JAK2 has exclusively sensitive-specific edges. We obtained the evidence networks for 59 of 66 mediators, and one of those mediators with evidence network is DAPK3 which is identified as a direct target of TG-101348, based on STITCH database. DAPK3 was identified as a mediator for the "ROLE OF DCC IN REGULATING APOPTOSIS" pathway which has an altered differential dependency

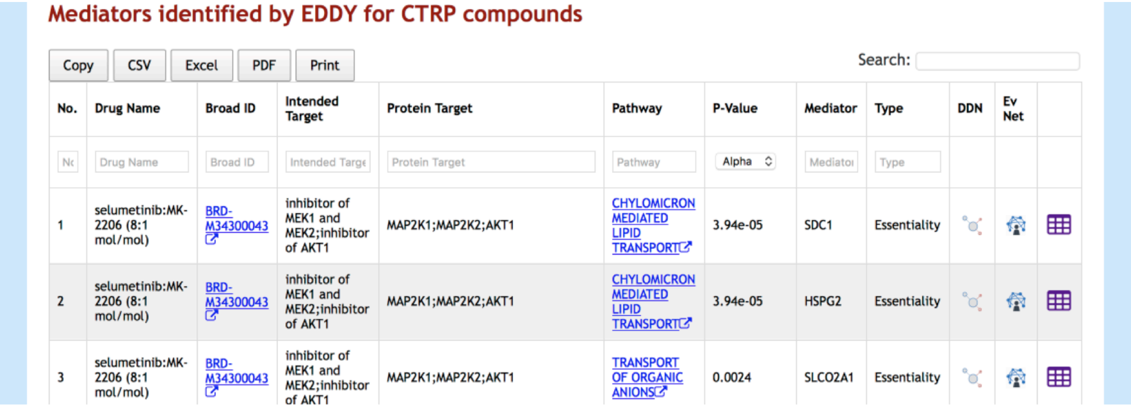

EDDY
Identifying Pathway-Specific Differential Dependencies
Associated with Drug Response in Cancer Cell Lines
MENU

Differentially dependent pathways identified by EDDY across CTRP compounds

Copy
CSV
Excel
PDF
Print
Search:

No.	Drug Name	Broad ID	Protein Target(s)	Target/Activity of Compound	Sensitive	Non-Sensitive	Pathways	AUC Plot
No	Drug Name	Broad ID	Protein Target(s)	Target/Activity of Compound	Sensitive	Non-Sen	Pathways	AUC F
1	selumetinib:MK-2206 (8:1 mol/mol)	BRD-M34300043	MAP2K1;MAP2K2;AKT1	inhibitor of MEK1 and MEK2;inhibitor of AKT1	198	189	▦	
2	PHA-793887	BRD-K64800655	CDK1;CDK2;CDK4;CDK5;CDK7;CDK9	inhibitor of cyclin-dependent kinases	203	180	▦	
3	BRD-K04800985	BRD-K04800985		product of diversity oriented	184	182	▦	

Figure 3. CTRP compound-centric view



Mediators identified by EDDY for CTRP compounds

Copy CSV Excel PDF Print

Search:

No.	Drug Name	Broad ID	Intended Target	Protein Target	Pathway	P-Value	Mediator	Type	DDN	Ev Net	
No	Drug Name	Broad ID	Intended Targe	Protein Target	Pathway	Alpha	Mediatoı	Type			
1	selumetinib:MK-2206 (8:1 mol/mol)	BRD-M34300043	inhibitor of MEK1 and MEK2;inhibitor of AKT1	MAP2K1;MAP2K2;AKT1	CHYLOMICRON MEDIATED LIPID TRANSPORT	3.94e-05	SDC1	Essentiality			
2	selumetinib:MK-2206 (8:1 mol/mol)	BRD-M34300043	inhibitor of MEK1 and MEK2;inhibitor of AKT1	MAP2K1;MAP2K2;AKT1	CHYLOMICRON MEDIATED LIPID TRANSPORT	3.94e-05	HSPG2	Essentiality			
3	selumetinib:MK-2206 (8:1 mol/mol)	BRD-M34300043	inhibitor of MEK1 and MEK2;inhibitor of AKT1	MAP2K1;MAP2K2;AKT1	TRANSPORT OF ORGANIC ANIONS	0.0024	SLCO2A1	Essentiality			

Figure 4. CTRP mediator-centric view

network for TG-101348. The gene product of DAPK3 was a mediator in this pathway due to high change of essentiality (betweenness centrality) between the condition-specific dependency networks TG-101348 sensitive cancer cell lines and non-sensitive cancer cell lines. In TG-101348-sensitive cell lines, DAPK3 is highly connected in the network (Figure 5a), consistent with DAPK3 playing a central role in a functioning apoptotic network. In the non-sensitive cell lines, however, DAPK3 is not connected to the rest of the network (Figure 5b), corroborating the indication that disconnected DAPK3 may confer insensitivity to TG-101348 sensitivity.

The evidence network built for TG-101348 - DAPK3 supports this hypothesis by showing a direct association between TG-101348 and DAPK3, discovered from the STITCH database (Figure 5c). Indeed, the evidence link was from a study that showed TG-101348 can inhibit the kinase activity of DAPK3, indicating that TG-101348 actually does target DAPK3 in addition to JAK2. Additionally, an association between the downstream JAK2 modulator and DAPK3 was revealed suggesting further signaling interactions targeted by TG-101348 [29]. So, while this target was not annotated in CTRP annotation for known targets of TG-101348, EDDY-CTRP

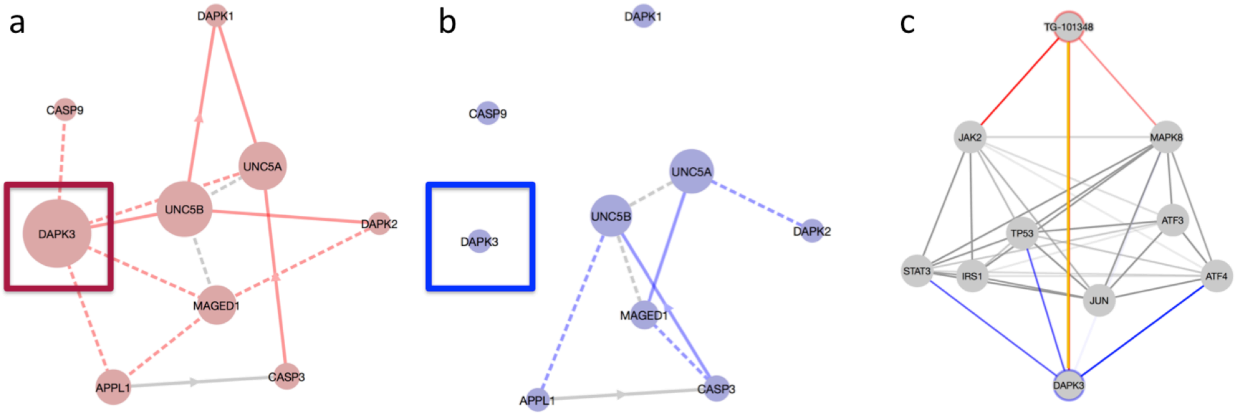


Figure 5. **(a)** Condition-specific dependency network (CDN) for TG-101348-sensitive cell lines. Dashed lines represent statistical dependencies while solid lines known interactions. Size of nodes represents node essentiality. **(b)** CDN for TG-101348-insensitive cell lines. **(c)** Evidence network for the TG-101348 – DAPK3 drug-mediator pair. All edges represent a known association based from the STRING/STITCH databases. Blue edges represent mediator-gene associations, red edges drug-gene associations, and yellow edge a direct drug-mediator association.

analysis was able to detect this relationship. This example illustrates EDDY can discover potentially novel targets of a compound and how the evidence network provides further contextual information regarding the possible mechanisms of how mediators selected in the EDDY analysis function to alter individual drug responses.

4.2. *HIF1A as an Alternative Target for Indisulam*

Indisulam is a carbonic anhydrase IX (CA9) inhibitor [30]. CA9 activity in cancer is associated with an acidic microenvironment that favors tumor cell survival and growth [31]. EDDY identified the HIF pathway as a DDN associated with indisulam response. The HIF pathway is important for cancer-cell survival in hypoxic conditions often seen in tumors [32]. In the non-responsive HIF pathway DDN two genes, HIF1A and JUN exhibit high essentiality compared to the responsive HIF DDN (Figure 6a). HIF1A is a major gene that signals for cell survival in hypoxic conditions [32]. The evidence network for indisulam and HIF1A reveals a direct link between CA9 and HIF1A (Figure 6c). This would not be evident if investigator had only HIF pathway DDN evidence. Inspection of the evidence from STRING shows that HIF1A positively regulates CA9 expression. Cancer cells may be non-responsive to indisulam because HIF1A increases CA9 levels such that the drug is not effective at tested concentration in fully inhibiting CA9. This example shows how the evidence network is able to mechanistically link EDDY DDNs to drug targets and expand understanding of signaling events associated with drug response.

5. Conclusions

While the current CTRP dataset allows the study of the correlations between genetic features with sensitivity to compounds, and while there are previous studies associating genes with compound sensitivity [33], this paper presents an unprecedented identification of pathways with differential dependency networks across a large number of cancer cell lines with drug-screening data. Additionally we have created a web repository to allow clinicians and researchers to view the results of our analysis. The web repository provides an interactive method to view the results for

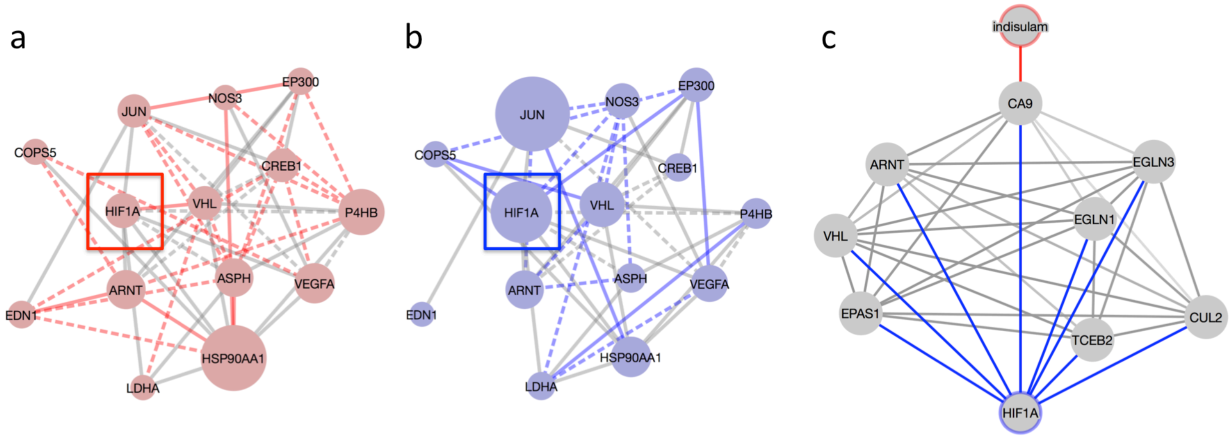


Figure 6. **(a)** Condition-specific dependency network (CDN) for indisulam for drug-sensitive cell lines. **(b)** CDN for indisulam for drug-insensitive cell lines. **(c)** Evidence network for the indisulam – HIF1A drug-mediator pair.

specific drugs. Researchers can query the intended targets, genes, or pathways to identify types of drugs, known targets, and to discover hitherto unknown mediators. We integrated quick unique links to the CTRP database, MSigDB Database, and Gene Cards, for each of the compounds, pathways, and genes. These links allow users to view the analysis and information about the drug, pathway, or gene seamlessly. We also provide links to the interactive DDN and condition-specific CDNs so that users can move around the nodes and edges to better analyze the results. In addition we provide links to generate the Oncoprints for the sensitive and non-sensitive cell lines for each DDN. These links allow the users to look at the mutation data used to generate the DDN.

This resource can be valuable for researchers to explore potential targets of their interest and allow them to look at differential dependencies across a large number of cell lines and compounds. It may aid in studying potential molecular mechanisms underlying cells' response to drug as well as designing experiments for the purpose of personalized treatment regimens.

Computational methods that can efficiently predict the effectiveness of drugs based on the genetic makeup of tumors would provide a major breakthrough towards personalized therapy for cancer patients based on their tumor's molecular markers. To strengthen the validity of our analysis and resource, experimental validation of the pathways identified by EDDY is warranted. We anticipate that this web repository will be a living resource for clinicians and researchers to use for designing experiments and identifying potential personalized treatment regimens.

References

1. Roden, D.M. and A.L. George, Jr., *The genetic basis of variability in drug responses.* Nat Rev Drug Discov, 2002. **1**(1): p. 37-44.
2. 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. 15545-50.
3. Califano, A., *Rewiring makes the difference.* Mol Syst Biol, 2011. **7**: p. 463.
4. Lai, Y., et al., *A statistical method for identifying differential gene-gene co-expression patterns.* Bioinformatics, 2004. **20**(17): p. 3146-3155.
5. Hu, R., et al., *Detecting intergene correlation changes in microarray analysis: a new approach to gene selection.* BMC Bioinformatics, 2009. **10**(1): p. 20.
6. Mentzen, W., M. Floris, and A. de la Fuente, *Dissecting the dynamics of dysregulation of cellular processes in mouse mammary gland tumor.* BMC Genomics, 2009. **10**(1): p. 601.
7. Zhang, B., et al., *Differential dependency network analysis to identify condition-specific topological changes in biological networks.* Bioinformatics, 2009. **25**(4): p. 526-32.
8. Zhang, B., et al., *DDN: a caBIG(R) analytical tool for differential network analysis.* Bioinformatics, 2011. **27**(7): p. 1036-8.
9. Hwang, T. and T. Park, *Identification of differentially expressed subnetworks based on multivariate ANOVA.* BMC Bioinformatics, 2009. **10**(1): p. 128.
10. Kim, Y., et al., *Principal network analysis: Identification of subnetworks representing major dynamics using gene expression data.* Bioinformatics, 2010.
11. Ma, H., et al., *COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method.* Bioinformatics, 2011.
12. Jung, S. and S. Kim, *EDDY: a novel statistical gene set test method to detect differential genetic dependencies.* Nucleic Acids Res, 2014. **42**(7): p. e60.

13. Speyer, G., et al., *Knowledge-Assisted Approach to Identify Pathways with Differential Dependencies.* Pac Symp Biocomput, 2016. **21**: p. 33-44.
14. Zheng, S., et al., *Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma.* Cancer Cell, 2016. **29**(5): p. 723-36.
15. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
16. Seashore-Ludlow, B., et al., *Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset.* Cancer Discov, 2015. **5**(11): p. 1210-23.
17. Rees, M.G., et al., *Correlating chemical sensitivity and basal gene expression reveals mechanism of action.* Nat Chem Biol, 2016. **12**(2): p. 109-16.
18. Fabregat, A., et al., *The Reactome pathway Knowledgebase.* Nucleic Acids Res, 2016. **44**(D1): p. D481-7.
19. Patro, R., G. Duggal, and C. Kingsford, *Accurate, fast, and model-aware transcript expression quantification with Salmon.* biorxiv, 2015.
20. Khuri, S. and S. Wuchty, *Essentiality and centrality in protein interaction networks revisited.* BMC Bioinformatics, 2015. **16**: p. 109.
21. Kuhn, M., et al., *STITCH 4: integration of protein-chemical interactions with user data.* Nucleic Acids Res, 2014. **42**(Database issue): p. D401-7.
22. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life.* Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
23. Yen, J., *Finding the K Shortest Loopless Paths in a Network* Management Science, 1971. **17**(11): p. 712-716.
24. Dijkstra, E.W., *A note on two problems in connexion with graphs.* Numerische Mathematik, 1959. **1**(4): p. 269-271.
25. Fredman, M.L. and R.E. Tarjan, *Fibonacci heaps and their uses in improved network optimization algorithms.* Journal of the Association for Computing Machinery, 1987. **34**(3): p. 596-615.
26. Seruga, B., et al., *Cytokines and their relationship to the symptoms and outcome of cancer.* Nat Rev Cancer, 2008. **8**(11): p. 887-99.
27. Lei, Y.Y., et al., *Mitogen-activated protein kinase signal transduction in solid tumors.* Asian Pac J Cancer Prev, 2014. **15**(20): p. 8539-48.
28. Wernig, G., et al., *Efficacy of TG101348, a selective JAK2 inhibitor, in treatment of a murine model of JAK2V617F-induced polycythemia vera.* Cancer Cell, 2008. **13**(4): p. 311-20.
29. Sato, N., et al., *Physical and functional interactions between STAT3 and ZIP kinase.* Int Immunol, 2005. **17**(12): p. 1543-52.
30. Supuran, C.T., *Indisulam: an anticancer sulfonamide in clinical development.* Expert Opin Investig Drugs, 2003. **12**(2): p. 283-7.
31. Swietach, P., et al., *New insights into the physiological role of carbonic anhydrase IX in tumour pH regulation.* Oncogene, 2010. **29**(50): p. 6509-21.
32. Masson, N. and P.J. Ratcliffe, *Hypoxia signaling pathways in cancer metabolism: the importance of co-selecting interconnected physiological pathways.* Cancer Metab, 2014. **2**(1): p. 3.
33. Gottlieb, A. and R.B. Altman, *Integrating systems biology sources illuminates drug action.* Clin Pharmacol Ther, 2014. **95**(6): p. 663-9.

A METHYLATION-TO-EXPRESSION FEATURE MODEL FOR GENERATING ACCURATE PROGNOSTIC RISK SCORES AND IDENTIFYING DISEASE TARGETS IN CLEAR CELL KIDNEY CANCER

JEFFREY A. THOMPSON[1] and CARMEN J. MARSIT[2]

[1]*Program in Quantitative Biomedical Science, Geisel Medical School at Dartmouth College,*
Lebanon, NH 03756, USA

[2]*Department of Environmental Health, Rollins School of Public Health at Emory Unveristy,*
Atlanta, GA 30322, USA

E-mail: carmen.j.marsit@emory.edu

Many researchers now have available multiple high-dimensional molecular and clinical datasets when studying a disease. As we enter this multi-omic era of data analysis, new approaches that combine different levels of data (e.g. at the genomic and epigenomic levels) are required to fully capitalize on this opportunity. In this work, we outline a new approach to multi-omic data integration, which combines molecular and clinical predictors as part of a single analysis to create a prognostic risk score for clear cell renal cell carcinoma. The approach integrates data in multiple ways and yet creates models that are relatively straightforward to interpret and with a high level of performance. Furthermore, the proposed process of data integration captures relationships in the data that represent highly disease-relevant functions.

Keywords: prognostic; survival; cancer; data integration; eQTL; m2eQTL; m2eGene

1. Introduction

The recent abundance of large datasets of diverse molecular features have vastly increased our knowledge of cellular processes disrupted in disease; yet, these datasets, taken individually, have frequently failed to reveal useful biomarkers for complex diseases, such as cancer [1, 2].

Despite the clear utility of individual 'omic' datasets, such as gene expression, DNA methylation, copy number alteration, etc., in better understanding disease etiology and in some cases providing useful prognostic or predictive value [3], it is equally clear that each of these data types can only capture part of the disease signature in a cell. Therefore, interest has been growing in more holistic methods, which integrate data of different types. As of yet, these approaches have met with mixed success. For example, a study of long-term survival in patients with glioblastoma multiforme (an aggressive form of brain cancer), found that joint regression of different types of data did not improve predictive accuracy [4]. Another study across five different cancer types came to a similar conclusion [5]. Nevertheless, a more nuanced approach, based on integrating separate models built from individual datatypes for ovarian cancer outcomes did show a higher predictive accuracy for integration across datatypes [6].

A recent review of data integration approaches classified them as falling into one of two broad categories: multi-stage and meta-dimensional integration [7]. Multi-stage integration techniques are currently the most developed and wide-spread. These involve using separate analyses of multiple types of data, with the results from one data type used to filter, and presumably increase the power of, another. The most commonly used example of multi-stage integration is expression-quantitative trait loci (eQTL) analysis, wherein single nucleotide polymorphisms (SNPs) are associated with changes in gene expression, which in turn are associated with disease [8, 9]. Meta-dimensional techniques consist of integrated models, in which all data are used as part of a joint model or analysis, which might involve joint regression, or integration at the level of individual models [10, 11].

Important prognostic information may in some cases be obscured by noise. However, it is much less likely that noise will obscure that information from different types of data for the same features. For example, it may be that repression of a gene promoter through DNA methylation represents a disease state. Nevertheless, that gene's expression may be altered in healthy individuals through alternative regulation. Therefore, it may not be enough to capture the gene expression data alone. Furthermore, one type of data may capture nascent information of disease progression that is not yet apparent in other data types. In some cases, there may not be one superior type of data for predicting prognosis. Finally, there may be informative interactions between data types that are not possible to assess when using only one type of data. For these reasons, we hypothesize that an appropriate multi-omic data-integrated approach will create superior prognostics to those using only a single data type.

In this work, we developed a data integration approach for combining gene expression data with DNA-methylation to create prognostic models for clear cell renal cell carcinoma (the most common form of kidney cancer). Our data integration approach is a hybrid method combining both multi-stage and meta-dimensional elements but results in a model that is easily interpreted by those familiar with traditional statistical approaches. Furthermore, it is amenable to extremely high dimensional data but runs quickly compared to other methods. We demonstrate the viability of our approach in the context of creating prognostic markers for kidney cancer and compare it to two other methods that have proven successful in this context: random survival forests and penalized Cox regression [12–14].

We chose to integrate DNA methylation and gene expression because they have proven to be prognostically useful data sources for a number of cancers [12] and are highly related. DNA methylation controls tissue specific expression of genes. Therefore, if we can exploit this redundant information, we may be able to create a more informative prognostic model. Furthermore, it has long been suspected that aberrant DNA methylation itself is related to carcinogenesis [15], although only recently has evidence begun to mount for a causative role [4, 16]. Given that DNA methylation tends to be a more stable mark than gene expression [17, 18], in certain cases it may be informative where gene expression is not. In cancer, hypermethylation of gene promoters silences tumor suppressors and other genes throughout the genome [19]. Hypomethylation of other regions is associated with genomic instability [20]. Thus, disruption of DNA methylation patterns may be a potentially relevant etiological factor, which could increase the utility of our approach.

2. Methods

We used M2EFM, and two other approaches, to model overall survival in clear cell renal cell carcinoma. For the main analysis, gene expression and DNA methylation profiles from untreated, resected tumors for patients with clear cell renal cell carcinoma were created by The Cancer Genome Atlas (TCGA) project [21] on the Illumina HiSeq 2000 sequencing and

Illumina Infinium HumanMethylation450 platforms respectively. RNA-seq data normalization was performed by TCGA and normalized data were downloaded from the UCSC Cancer Genomics Browser [22] (Table 1). The RSEM normalized read counts were log_2 transformed by the UCSC, and we left them in that form. DNA methylation data were obtained from the National Cancer Institute's Genomic Data Commons. These were functionally normalized using the `minfi` package [23, 24] for the R statistical environment [25].

A separate smaller dataset of methylation profiles (from the same platform) was also used by our method to identify differentially methylated loci between 46 paired tumor and tumor-adjacent normal clear cell kidney cancer samples obtained through the National Center for Biotechnology Information's Gene Expression Omnibus (GSE61441) [26]. Again, we used functional normalization for these data.

Table 1. Distribution of Samples in TCGA Clear Cell Renal Cell Carcinoma (Clear Cell Kidney Cancer) Data

	RNA-seq (%)	450k (%)	Overlap (%)
Samples w/ overall survival data	525	311	310
Male	341 (64.95)	201 (64.63)	201 (64.84)
Female	184 (35.05)	110 (35.37)	109 (35.16)
Stage I	262 (49.90)	150 (48.23)	150 (48.39)
Stage II	56 (10.67)	30 (9.65)	30 (9.68)
Stage III	126 (24.00)	75 (24.16)	74 (23.87)
Stage IV	81 (15.43)	56 (18.01)	56 (18.06)
Grade 1	12 (2.29)	7 (2.25)	7 (2.26)
Grade 2	228 (43.43)	132 (42.44)	132 (42.58)
Grade 3	202 (38.48)	119 (38.26)	119 (38.39)
Grade 4	75 (14.29)	49 (15.76)	48 (15.48)
Grade X	5 (0.95)	2 (0.64)	2 (0.65)
Missing Grade	3 (0.57)	2 (0.64)	2 (0.65)
Deaths	166 (31.62)	99 (31.83)	98 (31.61)
Mean Age	60.65	61.43	61.48

There was no evidence of significant differences in the distribution of staging or tumor grade for cases in the RNA-seq and DNA-methylation data (χ^2 test, p = 7.77e-01 and p = 9.54e-01 respectively). For all data types, there were 8 cases missing survival data, with 5 having no clinical annotation at all. The remaining 3 were female, had a mean age of 70.33 years, and contained 2 stage I and 1 stage II tumors. Other than the 5 with no clinical annotation, there were no samples missing on clinical predictors, therefore we decided to remove the 8 samples missing outcomes from the analysis.

Beta values were transformed into M-values [27], and we removed probes on the X or Y chromosomes, containing SNPs [28, 29], or with cross-hybridization issues [30]. Finally, probes with values missing for greater than 50% of samples were removed and the remaining values were imputed using the k-nearest neighbors method, with k=10, from the `impute` package [31, 32] for R.

2.1. *M2EFM*

We developed a data-integrated modeling approach we call Methylation-to-Expression Feature Model (M2EFM). The basis of this approach is to find loci that are differentially methylated between matched pathologic and non-pathologic data and to associate those loci with significant differences in gene expression in the disease state. The process is analogous to expression quantitative trait loci (eQTL) analysis, except that instead of associating SNPs with changes in gene expression, we associate differentially methylated loci. The loci are then called m2eQTLs (for methylation-to-expression QTLs) and the genes are called m2eGenes.

The approach consists of five primary steps (summarized in Fig. 1):

(1) **Filtering probes and genes for variability.** Gene expression values were filtered to remove very low variability genes (usually genes with no expression) by removing genes with a median absolute deviation of .05 or less, leaving 16907 genes. Methylation probes were filtered to remove those with a median absolute deviation of less than 0.8 (after transformation to M-values). This left 27700 probes for the kidney cancer data.

(2) **Identifying differentially methylated loci.** Differential methylation was identified using the empirical Bayes method from the `limma` package [33] for R. We used 46 paired tumor and tumor-adjacent normal samples from a separate dataset than used in the rest of the analysis. This initial step was used to identify which loci to focus on. We passed the 500 CpG loci with the lowest adjusted p-values (Benjamini-Hochberg) for differential methylation on to the the next step.

(3) **Identifying methylation-to-expression quantitative trait loci (m2eQTLs).** m2eQTL analysis involves associating methylation levels at the loci identified in the previous step with gene expression levels genome-wide. In terms of an eQTL analysis, the proportion of methylated alleles for a particular loci is equivalent to the genotype at a single nuncleotide polymorphism (SNP), although it is a continuous, rather than discrete value. Identification of m2eQTLs was performed using the `MatrixEQTL` package [34] for R, which builds linear models to test association in a computationally efficient manner. In this way, the M-value of probes in the training data that were found to be differentially methylated in the first step were tested for their association with gene expression patterns in both *cis* and *trans* in a manner analogous to that used in typical eQTL analysis. An m2eQTL was defined to act in *cis* if it was associated with a gene within 10000bp, otherwise it was defined to act in *trans*. The top 150 *cis* and *trans*-m2eQTLs (by effect size) and their associated m2eGenes were passed on to the next step. This number was simply chosen to identify around 200 relevant genes and may not be optimal.

(4) **Building integrated models from m2eQTLs and m2eGenes.** From the previous results we built a joint regression model across both probes and genes involved in the m2eQTLs. Given that these were bound to have collinearity, to prevent overfitting we used Cox regression with Ridge penalty [35]. The linear predictor from the Cox model was used as a molecular risk score for all training samples (see Supplementary File 1, http://dx.doi.org/10.5061/dryad.b1t61).

(5) **Integrating clinical variables.** M2EFM uses a second regression to integrate clinical variables. For this step, we performed an unpenalized Cox regression on the molecular risk

score from the previous step and the values of clinical variables. This allows the hazards in the model to be more interpretable and keeps the clinical covariates from being penalized. In a typical Cox proportional hazards model, there is a rule of thumb that there should be no more than about 10 events in the data per variable in the model. Each training dataset in our data will have about 69 events (depending on the split of the data), meaning the model should have only about 7 variables. Clinical variables used for cancer prognosis vary but can include TNM staging, tumor grade, AJCC stage, patient sex, and age at diagnosis. We tried a few alternative clinical models on the training data only and picked the one with the highest discrimination (measured by concordance index, Table S1). Although the results were close for TNM staging and AJCC stage (the difference was significant at p = 1.04e-05), TNM staging would add 17 variables to the model and AJCC stage only 4, so our final model includes patient age at diagnosis, sex, tumor stage, and risk score. Although this is 8 variables, relaxing the rule to 9 events per variable has been shown to be acceptable [36] and can moreover be judged to some degree from our results.

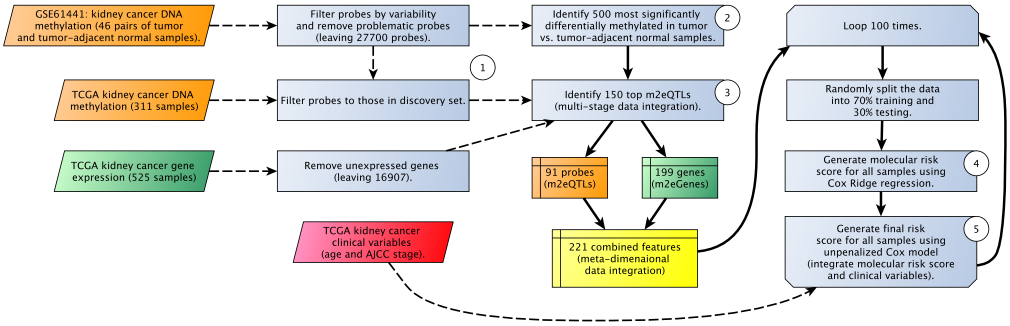

GSE61441: kidney cancer DNA methylation (46 pairs of tumor and tumor-adjacent normal samples).
Filter probes by variability and remove problematic probes (leaving 27700 probes).
Identify 500 most significantly differentially methylated in tumor vs. tumor-adjacent normal samples.
2
Loop 100 times.
1
TCGA kidney cancer DNA methylation (311 samples)
Filter probes to those in discovery set.
Identify 150 top m2eQTLs (multi-stage data integration).
3
Randomly split the data into 70% training and 30% testing.
TCGA kidney cancer gene expression (525 samples)
Remove unexpressed genes (leaving 16907).
91 probes (m2eQTLs)
199 genes (m2eGenes)
Generate molecular risk score for all samples using Cox Ridge regression.
4
TCGA kidney cancer clinical variables (age and AJCC stage).
221 combined features (meta-dimenaional data integration)
Generate final risk score for all samples using unpenalized Cox model (integrate molecular risk score and clinical variables).
5

Fig. 1. Workflow for M2EFM analysis of clear cell renal cell carcinoma.

2.2. *Experimental Design*

We built 100 different M2EFM models of overall survival in clear cell kidney cancer for 100 different random splits of the data, using 70% training and 30% testing data sets. This process was repeated for different combinations of data (clinical variables only, gene expression only, methylation only, expression and clinical variables, and methylation and clinical variables).

The results of our approach were compared with two other methods that have previously been shown to successfully integrate molecular and clinical data to generate prognostic markers: penalized Cox regression and random survival forest [12] (although that work did not attempt molecular data integration). We used Cox Ridge regression, rather than LASSO (which was used in [12]), because it generally has better predictive performance. The model was built using the `glmnet` package [37] for R, and the lambda parameter was found using 10-fold cross validation for each split of the data. The random survival forest was built using the `randomForestSRC` package [38] for R. The run time of the random forest prevented cross validation of the parameters, so these were left at the defaults, as in [12]. The performance of the models was evaluated using concordance or C-index, a commonly used measure of discrimination in prognostic models. The C-index is a measure of how likely it is, in any given pair of individuals, that the individual with the higher risk score has the event first.

2.3. *Functional Analysis Approach*

Although it is not a requirement that the genes used in a prognostic model are functionally related to the disease, models built from functional relationships can reveal important insight into why one patient might have a better prognosis than another, which can lead to improved treatment decisions and a higher probability of model validation. Therefore, we performed a functional analysis of the gene set used in our model. The m2eQTL genes were used to perform a gene set network enrichment analysis using the online tool WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) [39] to identify genes in our gene set that were enriched in sub-networks of protein-protein interactions that were, in turn, enriched for biological functions. We also used it to perform enrichment analysis for GO biological process terms. For both of these analyses we required at least 5 genes to overlap the gene module or pathway.

Our goal with this work was to demonstrate a method by which a biomarker can be identified. We do not identify a specific gene and DNA methylation probe set, in part because an independent validation dataset would be required.

3. Results

3.1. *M2EFM Prognostics*

The m2eQTL phase of M2EFM identifies differentially methylated loci that are associated with changes in gene expression throughout the genome. An example is shown in Fig. S1.

For the M2EFM-based risk score, the median C-index over 100 random splits of the data of the score from combined clinical and molecular variables (M2EFM Exp+Meth+Clin) reflects the highest prognostic accuracy of any method or data type used at .792. The median C-index of the risk score from clinical variables alone (M2EFM Clin) was .776 and the median C-index of the risk score from molecular variables alone (M2EFM Exp+Meth) was .702 (Fig. 2). The improvement in C-index for the combined clinical and molecular model over the clinical variables alone was significant at p = 4.25e-06 by two tailed Wilcoxon signed-rank test.

The M2EFM expression without methylation models had only slightly lower accuracy than models built using both data types. For these models, the median C-index for the combined clinical and expression models (M2EFM Exp+Clin) was .791 and for the expression only models (M2EFM Exp) was .703. The improvement in C-index for M2EFM Exp+Clin over the clinical variables alone was significant at p = 1.50e-08.

The M2EFM methylation without expression models were not as accurate as the other M2EFM models. The median C-index for the combined clinical and methylation models (M2EFM Meth+Clin) was .755 and for the methylation only models (M2EFM Meth) was .643. In this case, the clinical variable model had generally stronger C-index values than M2EFM Meth+Clin at p = 2.068e-08.

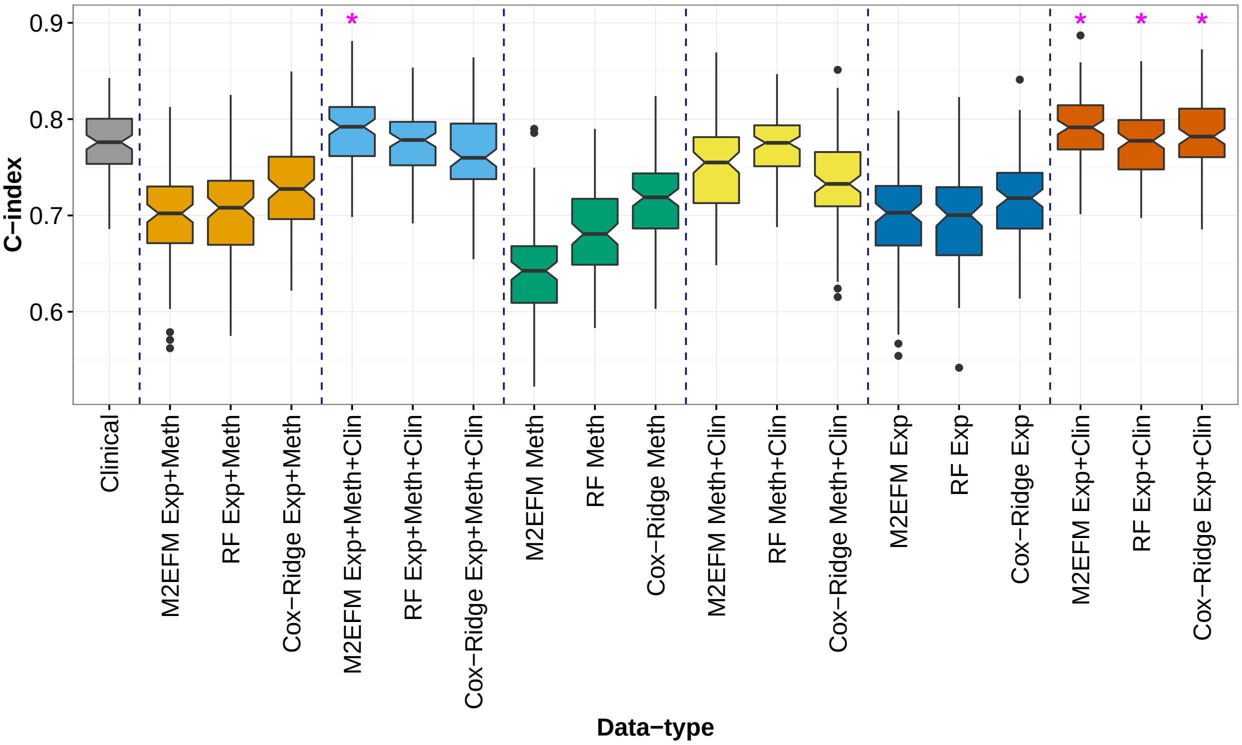


Fig. 2. C-index across 100 random splits into training and testing data of the various approaches. If one method was significantly better than another than the notches in the box plots will not overlap. For convenience, if a method resulted in significantly better results than clinical data alone, it is marked with "*".

3.2. *Random Survival Forest Prognostics*

Random survival forest was not as effective at exploiting the integrated expression and methylation data as our guided M2EFM approach. The median C-index for the combined clinical and molecular features (RF Exp+Meth+Clin) over the same 100 random splits of the data was .776 and the median C-index of the models built from the molecular data alone (RF Exp+Meth) was .696. The addition of the molecular data using random survival forest models was no more discriminatory than the clinical variables alone.

The performance of the expression without methylation random survival forest model was similar to the model with both data types. The median C-index for RF Exp+Clin model was .777, which was very slightly but significantly stronger than the clinical only model (p = 7.16e-03) while the median C-index for the RF Exp model was .694.

The performance of the methylation without expression random survival forest model was slightly worse than when both data types were used. The median C-index for the RF Meth+Clin model was .776, but the RF Meth model was a significant improvement over M2EFM Meth (p = 8.05e-13) and had a median C-index of .682.

3.3. *Cox-Ridge Prognostics*

Cox regression with ridge penalty [40] outperformed M2EFM when it came to the molecular data alone, but its molecular risk score was less independent of the clinical variables, thus its accuracy for the full model was less than that of M2EFM. The median C-index for the combined clinical and molecular features (Cox-Ridge Exp+Meth+Clin) of the same 100 random splits of the data was .760 and the median C-index of the models built from molecular data alone (Cox-Ridge Exp+Meth) was .727 and was improved over M2EFM Exp+Meth (p = 3.10e-05).

The performance of Cox-Ridge Exp+Clin model was slightly worse than the M2EFM Exp+Clin model (p = 9.14e-13) with a median C-index of .782. Again, the performance of the molecular data only model, Cox-Ridge Exp, was somewhat better than M2EFM Exp (p = 2.35e-06) with a median C-index of .718.

Finally, the Cox-Ridge Meth+Clin model did not perform as well as the M2EFM model. It achieved a median C-index of .735, which was significantly worse than the M2EFM Meth+Clin model (p = 6.37e-15). Nevertheless, the Cox-Ridge Meth model, with a median C-index of .705, performed better than the M2EFM Meth model (p = 1.62e-12).

3.4. *Comparison to Yuan et al.*

A direct comparison of our approach to that used in [12] on the same data was not possible, because the data they deposited included only the pre-filtered DNA methylation values, which did not include the same probes we identified in our discovery set. Nevertheless, we attempted to run our method on this subset of probes (which necessarily created different models than those used above). The highest mean C-index of any method listed in [12] on the kidney cancer data as .767 for a model including microRNA and clinical variables. On the same data (normalized by Yuan et al.), we achieved a mean C-index of .775 for the M2EFM Meth+Exp+Clin model and a mean C-index of .773 for the M2EFM Exp+Clin.

3.5. *Functional Analysis*

3.5.1. *Gene Set Network Enrichment*

Next we performed gene set network enrichment analysis using the online tool WebGestalt, requiring a minimum of 5 genes to overlap a gene module. All significant results (after multiple testing correction) are shown in Table 2. The full list of genes found in each pathway is given in Supplementary File 2. This approach revealed enrichment for gene modules associated with immune response, proliferation, and other functions. As an example, a portion of the largest sub-network our model was enriched in (which is enriched for the JAK-STAT Cascade) is shown in Fig. 3 (visualized using Cytoscape [41]). The genes from our gene set are shown in green and are highly connected nodes in the network.

3.5.2. *Biological Process Enrichment*

We further tested the straight enrichment for biological process terms in the Gene Ontology using our gene set (without network enrichment), again requiring a minimum of 5 genes to overlap a pathway. The results in Table 3 show the top 5 most enriched GO terms, with a clear enrichment for immune system related genes. The full list of genes enriched in each pathway is given in Supplementary File 3.

Table 2. Protein Interaction Network Module Enrichment

Pathway	Observed	Expected	Adj. p
T Cell Costimulation	7	.33	2.69e-07
Regulation of Defense Response to Virus by Host	11	1.18	2.69e-07
JAK-STAT Cascade Involved in Growth Hormone Signaling Pathway	34	19.64	3.80e-03
Complement Activation	6	1.82	3.43e-02

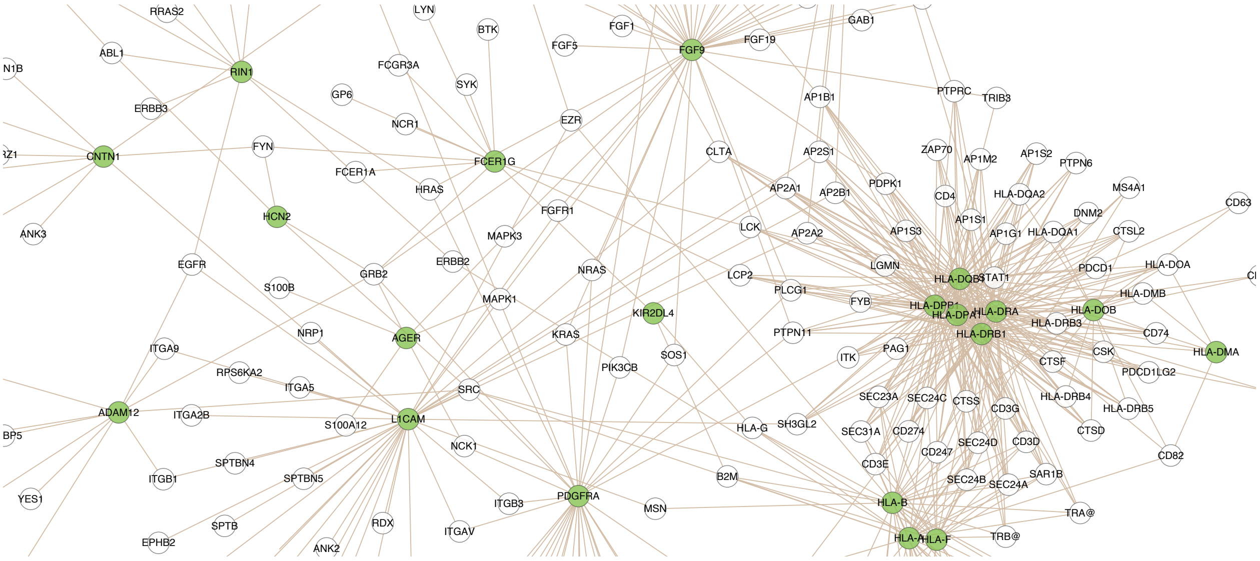


Fig. 3. Portion of the gene module enriched for the JAK-STAT cascade. Genes from our gene set are shown in green.

Table 3. Enriched for GO Biological Process

Pathway	Observed	Expected	Adj. p
Antigen Processing and Presentation of Exogenous Antigen	13	2.04	1.60e-05
Antigen Processing and Presentation of Exogenous Peptide Antigen	13	2.01	1.60e-05
Antigen Processing and Presentation	15	2.6	1.60e-05
Response to Interferon-Gamma	11	1.30	1.60e-05
Cellular Response to Interferon-Gamma	10	1.07	1.60e-05
Exogen	13	2.15	2.09e-05
Interferon-Gamma-Mediated Signaling Pathway	9	.89	2.09e-05
Antigen Processing and Presentation of Peptide Antigen	13	2.23	2.87e-05
Immune Response	32	12.24	2.95e-05

4. Discussion

All of the approaches described show it is possible to attain a meaningful level of prognostic discrimination using a joint regression on both gene expression and DNA-methylation values, if collinearity is properly accounted for. However, our approach, which first identifies dysregulation of DNA methylation in cancer, then associates that dysregulation to differences in gene expression, and finally builds prognostic markers from genes and CpG loci that are associated with this loss of regulation, was able to build models with a higher level of prognostic discrimination than either a random survival forest approach or Cox regression with Ridge penalty as well as a model built from traditional clinical variables. The results for our joint molecular regression with M2EFM were about the same as using expression data alone, leaving it unclear if this form of meta-dimensional integration is helpful on top of the multi-stage integration, which selected the features, thus more work on this part of the approach is needed.

The median C-index of .792 achieved by our M2EFM Exp+Meth+Clin model was the most accurate predictor of overall survival achieved by any approach in this study. This result was achieved through three data integrations, including different types of molecular data, as well as clinical variables. Notably, we showed that M2EFM's combination of a molecular risk score with clinical variables was a significant improvement over the clinical variables alone. Furthermore, our m2eQTL analysis identified 199 genes with high relevance to clear cell kidney cancer, without *a priori* knowledge of those genes' association to the disease. In fact, one of the top results from our gene set network enrichment analysis was for the JAK-STAT Cascade pathway, which is a known factor in kidney cancer progression [42]. That we identified this pathway by associating differentially methylated CpGs with differences in gene expression may suggest a role for dysregulation of methylation in the development of the disease, although caution in this interpretation is warranted, due to the cross-sectional nature of our study. An additional limitation was our lack of an independent dataset containing samples with gene expression and DNA methylation profiles as well as clinical data for validation.

The high enrichment we observed for genes involved in the immune system may indicate the utility of our approach in identifying survival differences based on dysregulation of immune functions. Given that immunotherapy has emerged over the last several years as an important component of kidney cancer treatment [43] and the pressing need for biomarkers that can identify the patients that will benefit from treatment [43], further development of this approach may be warranted in this regard. Another interesting result was our identification of *CA9*, which is currently of interest as a possible serum biomarker for kidney cancer [44], as a potential target for radioimaging [45], and as a potential therapeutic target [46]. Taken together, our results suggest that our approach is able to identify functionally relevant, and not just prognostic, genes. This is promising in terms of eventual validation of our approach.

Most of our results were better than those in a recent study including kidney cancer prognostics [12], but in a couple of cases, either the random forest or the Cox-Ridge approach did not perform as well as the methods in that work. However, they used fewer samples in that study and included inferred cancer subtypes from non-negative matrix factorization (NMF), in addition to gene and probe level measurements. Using only the DNA methylation and gene expression data from that study, which handicapped our method in discovery, M2EFM still showed slightly higher discrimination than any other approach. However, our goal was to develop a method based primarily on feature selection, rather than transformative dimensionality reduction techniques, in order to reduce the complexity of the models. Although interpretability is still limited by our use of Cox Ridge regression in generating the molecular risk score, it is over a limited number of genes that appear to be functionally related, mitigating this issue. It is notable that our m2eQTL-based approach creates models that outperform those using NMF, through a motivated feature selection technique that selects for putative regulatory relationships. We also note that Cox-Ridge in most cases outperformed the Cox-LASSO approach used in [12], and in some subsets of the data performed slightly better than M2EFM for prognostic accuracy. However, this accuracy comes at the cost of interpretability. The Cox-Ridge models contain thousands of genes or probes, telling us little in terms of the function of prognostic genes and creating unwieldy biomarkers in terms of real world use.

5. Conclusions

We developed a new data-integrated approach to modeling cancer prognostics and applied it to clear cell renal cell carcinoma data. M2EFM uses both a multi-stage data integration that links changes in methylation between tumor and normal tissues to levels of gene expression, and a meta-dimensional data integration that combines DNA methylation and gene expression values as part of a joint regression for outcome prediction. M2EFM was shown to identify not only prognostic, but functionally relevant features that may be associated with therapeutic response and that were highly connected in relevant protein-protein interaction networks.

6. Acknowledgements

We would like to acknowledge our funding sources: NIH-NIMH R01MH094609, NIH-NIEHS R01ES022223, and NIH-NIEHS P01 ES022832/EPA RD83544201 and a Hopeman award from the Norris Cotton Cancer Center.

References

[1] M. Huang, A. Shen, J. Ding and M. Geng, *Trends Pharmacol Sci* **35**, 41 (2014).
[2] S. E. Kern, *Cancer Res* **72**, 6097 (2012).
[3] A. S. Coates, E. P. Winer, A. Goldhirsch, R. D. Gelber, M. Gnant, M. Piccart-Gebhart, B. Thürlimann, H.-J. Senn, F. André, J. Baselga *et al.*, *Ann Oncol* **26**, 1533 (2015).
[4] J. Lu, M. C. Cowperthwaite, M. G. Burnett and M. Shpak, *PloS ONE* **11**, p. e0154313 (2016).
[5] L. Xu, L. Fengji, L. Changning, Z. Liangcai, L. Yinghui, L. Yu, C. Shanguang and X. Jianghui, *PloS ONE* **10**, p. e0142433 (2015).
[6] D. Kim, R. Li, S. M. Dudek and M. D. Ritchie, *BioData Min* **6**, p. 23 (2013).
[7] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass and D. Kim, *Nat Rev Genet* **16**, 85 (2015).
[8] A. C. Nica and E. T. Dermitzakis, *Phil Trans R Soc B* **368**, p. 20120362 (2013).
[9] R. Breitling, Y. Li, B. M. Tesson, J. Fu, C. Wu, T. Wiltshire, A. Gerrits, L. V. Bystrykh, G. De Haan, A. I. Su *et al.*, *PLoS Genet* **4**, p. e1000232 (2008).
[10] E. R. Holzinger, S. M. Dudek, A. T. Frase, S. A. Pendergrass and M. D. Ritchie, *Bioinformatics* , p. btt572 (2013).
[11] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine and C. Sander, *PLoS ONE* **6**, p. e24709 (2011).
[12] Y. Yuan, E. M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, L. A. Byers, Y. Xu, K. R. Hess, L. Diao *et al.*, *Nat Biotechnol* **32**, 644 (2014).
[13] G. Ambler, S. Seaman and R. Omar, *Stat Med* **31**, 1150 (2012).
[14] F. R. Datema, A. Moya, P. Krause, T. Bäck, L. Willmes, T. Langeveld, B. de Jong, J. Robert and H. M. Blom, *Head Neck-J Sci Spec* **34**, 50 (2012).

[15] M. Ehrlich *et al.*, *Oncogene* **21**, 5400 (2002).

[16] D.-H. Yu, R. A. Waterland, P. Zhang, D. Schady, M.-H. Chen, Y. Guan, M. Gadkari and L. Shen, *J Clin Invest* **124**, 3708 (2014).

[17] J.-P. Issa, *J Clin Oncol* **30**, 2566 (2012).

[18] P. W. Laird, *Nat Rev Cancer* **3**, 253 (2003).

[19] M. Esteller *et al.*, *Oncogene* **21**, 5427 (2002).

[20] K. L. Sheaffer, E. N. Elliott and K. H. Kaestner, *Cancer Prev Res (Phila)* **9**, 534 (2016).

[21] Cancer Genome Atlas Research Network *et al.*, *Nature* **499**, 43 (2013).

[22] M. S. Cline, B. Craft, T. Swatloski, M. Goldman, S. Ma, D. Haussler and J. Zhu, *Sci Rep* **3** (2013).

[23] J.-P. Fortin, A. Labbe, M. Lemire, B. W. Zanke, T. J. Hudson, E. J. Fertig, C. M. Greenwood and K. D. Hansen, *Genome Biol* **15**, p. 1 (2014).

[24] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen and R. A. Irizarry, *Bioinformatics* **30**, 1363 (2014).

[25] R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, (2016).

[26] J.-H. Wei, A. Haddad, K.-J. Wu, H.-W. Zhao, P. Kapur, Z.-L. Zhang, L.-Y. Zhao, Z.-H. Chen, Y.-Y. Zhou, J.-C. Zhou *et al.*, *Nat Commun* **6** (2015).

[27] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou and S. M. Lin, *BMC Bioinformatics* **11**, p. 587 (2010).

[28] 1000 Genomes Project Consortium *et al.*, *Nature* **491**, 56 (2012).

[29] L. Butcher, *Illumina450ProbeVariants.db: Annotation Package combining variant data from 1000 Genomes Project for Illumina HumanMethylation450 Bead Chip probes*, (2013). R package version 1.1.1.

[30] Y.-a. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson and R. Weksberg, *Epigenetics* **8**, 203 (2013).

[31] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. B. Altman, *Bioinformatics* **17**, 520 (2001).

[32] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown and D. Botstein, Imputing missing data for gene expression arrays (1999).

[33] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth, *Nucleic Acids Res* **43**, p. e47 (2015).

[34] A. A. Shabalin, *Bioinformatics* **28**, 1353 (2012).

[35] A. E. Hoerl and R. W. Kennard, *Technometrics* **12**, 55 (1970).

[36] E. Vittinghoff and C. E. McCulloch, *Am J Epidemiol* **165**, 710 (2007).

[37] N. Simon, J. Friedman, T. Hastie and R. Tibshirani, *J Stat Softw* **39**, p. 1 (2011).

[38] H. Ishwaran, U. B. Kogalur, E. H. Blackstone and M. S. Lauer, *Ann Appl Stat* , 841 (2008).

[39] J. Wang, D. Duncan, Z. Shi and B. Zhang, *Nucleic Acids Res* **41**, W77 (2013).

[40] H. Zou and T. Hastie, *J R Stat Soc Series B Stat Methodol* **67**, 301 (2005).

[41] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, *Genome Res* **13**, 2498 (2003).

[42] S. Li, S. J. Priceman, H. Xin, W. Zhang, J. Deng, Y. Liu, J. Huang, W. Zhu, M. Chen, W. Hu *et al.*, *PLoS ONE* **8**, p. e81657 (2013).

[43] M. W. Ball, M. E. Allaf and C. G. Drake, *Discov Med* **21**, 305 (2016).

[44] M. Takacova, M. Bartosova, L. Skvarkova, M. Zatovicova, I. Vidlickova, L. Csaderova, M. Barathova, J. Breza, P. Bujdak, J. Pastorek *et al.*, *Oncology Lett* **5**, 191 (2013).

[45] P.-C. Lv, J. Roy, K. S. Putt and P. S. Low, *Mol Pharm* **13**, 1618 (2016).

[46] J. Tostain, G. Li, A. Gentil-Perret and M. Gigante, *Eur J Cancer* **46**, 3141 (2010).

DE NOVO MUTATIONS IN AUTISM IMPLICATE THE SYNAPTIC ELIMINATION NETWORK[*]

GUHAN RAM VENKATARAMAN

Department of Bioengineering, Stanford University, 318 Campus Drive
Stanford, CA 94305, USA
Email: guhan@stanford.edu

CHLOE O'CONNELL, FUMIKO EGAWA, DORNA KASHEF-HAGHIGHI, DENNIS P. WALL

Department of Pediatrics and Biomedical Data Science, 1265 Welch Road
Stanford, CA 94305, USA
Email: dpwall@stanford.edu

Autism has been shown to have a major genetic risk component; the architecture of documented autism in families has been over and again shown to be passed down for generations. While inherited risk plays an important role in the autistic nature of children, *de novo* (germline) mutations have also been implicated in autism risk. Here we find that autism *de novo* variants verified and published in the literature are Bonferroni-significantly enriched in a gene set implicated in synaptic elimination. Additionally, several of the genes in this synaptic elimination set that were enriched in protein-protein interactions (CACNA1C, SHANK2, SYNGAP1, NLGN3, NRXN1, and PTEN) have been previously confirmed as genes that confer risk for the disorder. The results demonstrate that autism-associated *de novos* are linked to proper synaptic pruning and density, hinting at the etiology of autism and suggesting pathophysiology for downstream correction and treatment.

[*] This work was supported in part by the Hartwell Foundation's Autism Research and Technology Initiative.

1. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that impairs social skills, communication, and normal behavior. About 1% of the world's population has ASD, and this number is rapidly rising: the prevalence of autism more than doubled between 2002 and 2012[1]. ASD-linked impairment leads to higher lifespan costs[2] and a significant reduction in ability to procure both postgraduate education and jobs[3].

Both inherited (present in mother or father) and *de novo* (germline) mutations have been shown to contribute to the disease[4]. Although several of each type appear to contribute to ASD risk, there is still not a clear picture or full map of what leads to ASD[5]. Several hypotheses exist for the genetic etiology of autism; one of note is referred to as the "synaptic elimination hypothesis," the exploration of which is the focus of this paper.

Synaptic elimination is a normal neurodevelopmental process, starting in the fifth week of development and continuing throughout life. The process occurs in parallel with synaptic formation, which relies on input from both presynaptic and postsynaptic neurons. Elimination eventually outpaces formation in adolescence and adulthood[6,7]. During the development of the central nervous system, neurons form multiple synapses in excess of functional need. These redundant synapses are later eliminated through various means: 1) loss of signals necessary from either presynaptic or postsynaptic neurons to maintain synaptic stability[8]; 2) apoptosis of synapses; 3) ubiquitination of synaptic proteins for proteosomal degradation[9]; 4) macro-autophagy[7]; or 5) phagocytosis of synapses as a result of opsonization by synaptic elimination-mediating complement factors, microglia, and astrocytes[10-13].

Previous work has shown that faulty synaptic formation and maturation contribute to ASD[6]. However, given that increases in both dendritic spine density and brain weight (both of which are characteristic of autism) can be caused by mutations in genes regulating synaptic elimination, the hypothesis developed that autism could *also* be a disease of abnormal synaptic elimination[8,14-16].

Currently, the major pathway and ontology databases (KEGG, GO, Panther, and Reactome) do not contain any gene sets that pertain to synaptic elimination or synaptic pruning. As part of this study, we endeavored to create a robust and manually curated list of genes contributing to synaptic elimination; our goal was to test the hypothesis that the curated gene set would be enriched for *de novo* mutations (see Supplemental Materials 2 and 3 for list of genes and references used to generate this list, respectively). We hypothesized that increased burden of mutations in synaptic elimination genes would lead to the synaptic pruning abnormalities observed in autism, such as increased dendritic spine density and increased brain weight.

We used the *dnenrich* package[17], a network burden analysis tool, to test for enrichment in the synaptic elimination gene set on a comprehensive set of exomes from family-based trios having one child with autism. The package has been shown to be particularly powerful for identifying *de novo* mutations with small individual association to phenotype, but large effect in combination. We used dnenrich on previously documented autism-associated gene sets and autism-associated *de novos* as a pilot. This was done to verify that the program was suitable for use with autism *de novos* and that our list of *de novos* was large enough to provide sufficient power to detect enrichment of certain gene sets. After doing so, we tested the hypothesis that our list of genes involved in synaptic elimination would have a higher burden of autism *de novos* than would be expected by chance.

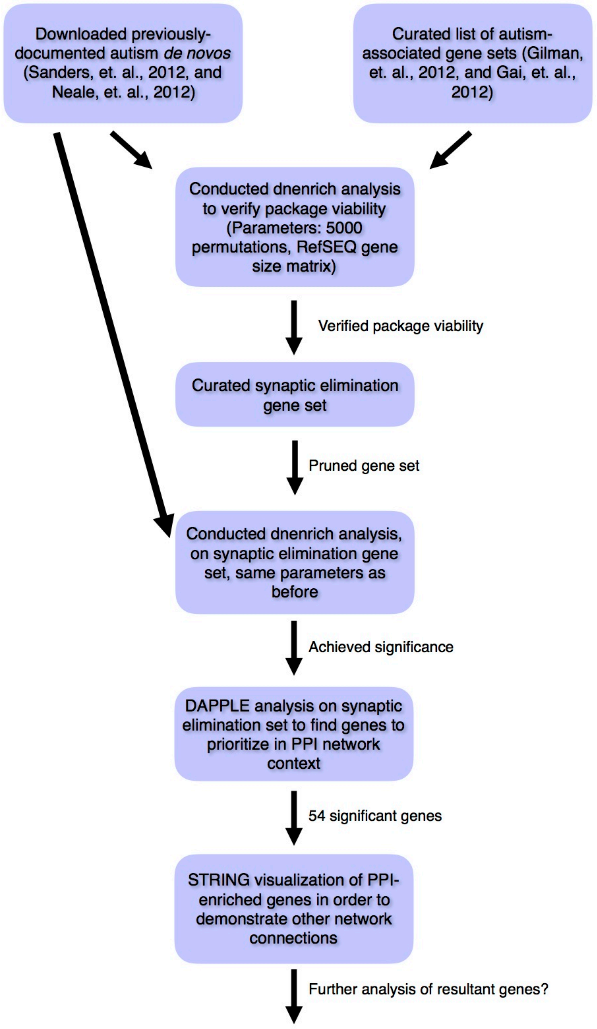

Downloaded previously-documented autism *de novos* (Sanders, et. al., 2012, and Neale, et. al., 2012)
Curated list of autism-associated gene sets (Gilman, et. al., 2012, and Gai, et. al., 2012)
Conducted dnenrich analysis to verify package viability (Parameters: 5000 permutations, RefSEQ gene size matrix)
Verified package viability
Curated synaptic elimination gene set
Pruned gene set
Conducted dnenrich analysis, on synaptic elimination gene set, same parameters as before
Achieved significance
DAPPLE analysis on synaptic elimination set to find genes to prioritize in PPI network context
54 significant genes
STRING visualization of PPI-enriched genes in order to demonstrate other network connections
Further analysis of resultant genes?

Fig 1. Schematic describing the overall flow of our experiment.

2. Methods

2.1. Autism *de novo* variants

We downloaded genomes of 3982 autism family trios from the Autism Sequencing Consortium (ASC) and the Simons Simplex Collection (SSC)[18-20]. These cohorts have been studied from a single-variant perspective, but have not yet been examined for their potential relationship to the synaptic elimination network.

Focusing on previously published full exome data, we built a comprehensive database of genomic variants to test for enrichment of synaptic elimination[18-21]. Specifically, we selected 189 autism trios and 31 unaffected siblings from SSC and then filtered out samples known to carry large *de novo* CNVs. Whole-exome sequencing was completed for 238 families selected from SSC, 200 of which included an unaffected sibling[23]. 15,480 DNA samples in 16 sample sets were analyzed, integrating *de novo*, inherited, and case-control loss-of-function counts and *de novo* missense variants predicted to be damaging. *De novos* were called using enhancements of previously published methods[18].

The full variant list from this collection, which also includes ASC cohorts, were compared to a larger set of 1,779 other exomes to confirm their putative roles in autism, and all *de novo* events were validated via PCR amplification and Sanger sequencing[22]. Family quads selected from SSC were sequenced with enrichment for higher functioning probands[19,20]. These *de novos* were interpreted using pipeline tools at each respective participating data center.

2.2. *Dnenrich Pilot Study*

Dnenrich simulates peppering the genome with random *de novos* by taking into account tri-nucleotide contexts, gene sizes, sequencing coverage, and functional effects of mutations. After permuting this process for a user-defined number of times, it then calculates one-sided P values, testing whether the observed number of mutations (in each gene set) is greater than the average simulated number of mutations (again, in each gene set).

We assembled 37 candidate gene sets to test their enrichment in our curated list of autism de novo variation. These 37 gene sets included Gene Ontology sets from previous autism network analyses[24,25], as well as genes shown to interact with FMRP (a mutation in which causes Fragile X syndrome, one of the most common causes of autism spectrum disorders)[21,26]. A full list of genes in each set tested for enrichment can be found in Supplemental Materials 1, along with their sizes.

We then performed an extensive process of literature mining and curation, through combined database search, hand-search, and related reference review, to assemble the synaptic elimination gene set. Our Pubmed search (conducted between May and June., 2016) included use of the terms “synapse,” “synaptic,” “elimination,” “pruning,” and “gene.” We then performed additional hand-searches of *Nature* and *Cell* using the same terms. References of included studies, review articles, and related references were screened for additional relevant studies based on title and abstract review. Our screening criteria for inclusion in the synaptic elimination set was the presence of the following terms: “synaptic elimination,” “synaptic pruning,” “synaptic stabilization,” “synaptic destabilization,” and “synaptic plasticity.” In all searches we excluded the following terms: “axon scaling,” “viral infection,” “axon repulsion,” “axon retraction,” and “neuromuscular junction.” Studies pertaining to synaptic formation, maintenance, and/or elimination within the peripheral nervous system were excluded on the basis of arising from separate embryologic origin than the central nervous system. Abstracts and unpublished data were excluded. The synaptic elimination gene set was curated through careful review of 120 selected studies and related reviews yielding 274 genes related to synaptic formation or elimination. Gene function was cross-referenced in ClinVar (accessed July 11, 2016), and 213 genes of interest were selected based on their role in synaptic elimination (see Supplemental Materials 2). After its curation, we tested the synaptic elimination gene set for enrichment for autism *de novos*.

Table 1. The 213 genes in the synaptic elimination gene set.

C1QA	PROS1	TYROBP	CD200	IGFBP4	TLR4	NFKB1
C1QB	CXC3L1	NFKB1	ITGAX	EDNRB	BDNF	NFKB2
C1QC	CX3CR1	CREB	ITGB2	TIMP2	C5	CAMK2G
C3	DAP12	MAPK14	GDNF	COL1Q2	H2-D	NCKAP5L
Mac-2	TREM2	NPTX2	CSF1	FN1	CCL7	NRXN2
CRK	CR-1	NGFR	CNTF	IRF8	CCL2	NRXN3
ELMO1	PGRN	APP	PTGER2	TGRBR2	CDC42	NRTK2
RAC1	CD68	PILRB	C1QBP	CFB/MHCIII	MBP	CRMP1
BAI1	CASP8	CD247	CALR	FCGR1B	CXCL13	CRK
MEGF10	CASP3	B2M	CR2	AIF1	Uba1	PLXA3
GULP1	CASP6	KLRA1	CD33	IL10BR	Mov34	PLXA4
ABCA1	CLU	TAP1	TNFRSF19	NOS1AP	Rpn6	TBR1
TYRO3	HLA-DR	C4	PDGFRA	MASP1	USP2	DPYSL2
AXL	HLA-C	CR-3	LEP	CD46	UFD2A	ADNP
MERTK	HLA-A	CD22	LEPR	CD55	MEF2	SPARC
GAS6	DR6	CD47	IGFBP3	TLR2	MEF2A	DYRK1A
EN2	GDA	TSPAN7	PAK3	CTNNB1	WNT2	FOXP1
MEF2B	REL	NLGN1	SEMA3A	BDNF	CHN2	RCAN1
MEF2C	RELA	NLGN2	SEMA3F	DHCR7	MAPK3	CHD8
MEF2D	RELB	NLGN3	NRP1	FMR1	MAPK1	RAC1
PARK2	SERPINA3	NLGN4	NRP2	AUTS10	TSC1	OPHN1
caspases	CUL3	SHANK1	RhoA	LAMC3	TSC2	FOXP2
hdc	ESCRT-I	SHANK2	ROCK1	MECP2	DOCK1	ARC
MIB1	shrub	SHANK3	OTX1	THBS1	EPHA4	CASK
UBE3A	ESCRT-III	CNTN4	DISC1	THBS2	EPHB3	DLG4
UBE3B	CHMP2B	CNTNAP2	KATNAL2	THBS4	EFNA4	HOMER1
PCDH10	mop	CNTNAP4	NTNG1	MAP2	EFNB3	PTEN
ATG5	Kat60L	CACNA1C	SYNGAP1	KALRN	NCK2/GRB4	
Atg7	IKBKG	SCN1A	Mek-1	KALRN	EB3	
LC3-II	Mical	SCN2A	Mek-2	CDC42	NGFR	
p62	NRXN1	RELN	SPARCL1	PPP1R9B	GRM5	

3. Results

We tested the 37 initial gene sets with dnenrich with the default gene size matrix provided on the dnenrich website (as adjusting for per-trio joint sequencing coverage “[does] not have a noticeable effect on results”[17]). We ran the simulation on the downloaded autism *de novos* for 5000 permutations without weighting any genes. Of the 37 gene sets tested, 10 were significantly enriched for *de novos* after Bonferroni adjustment for 37 hypotheses. These sets are listed in Table

2. Given the enrichment of *de novos* in known autism networks calculated by dnenrich, we felt confident in using both this set of previously-published *de novos* and dnenrich to test the single hypothesis that synaptic elimination genes would have an exceedingly high burden of *de novos*.

Table 2. Bonferroni-significant gene sets enriched for autism *de novos* using dnenrich. Unadjusted p-values were obtained directly from dnenrich; adjusted p-values were Bonferroni-corrected by the number of sets tested.

	p-value		Number of Mutations		Location	
Gene Set Name	Unadjusted	Adjusted	Observed	Expected	Reference to Autism	Source
Developmental Process	1.9996×10^{-4}	8.798×10^{-3}	731	648.659	Gai et. al. (2012)	GO
FMRP	1.9996×10^{-4}	8.798×10^{-3}	412	285.33	Darnell et. al. (2011)	Paper
Learning and/or Memory	1.9996×10^{-4}	8.798×10^{-3}	78	44.1032	Gilman et. al. (2011)	GO
M3	1.9996×10^{-4}	8.798×10^{-3}	206	151.955	Parikshak et. al. (2013)	Paper
Protein modification process	1.9996×10^{-4}	8.798×10^{-3}	577	496.748	Gai et. al. (2012)	GO
Synaptic transmission	1.9996×10^{-4}	8.798×10^{-3}	163	114.367	Gai et. al. (2012)	GO
Axonogenesis	3.9992×10^{-4}	1.7596×10^{-2}	136	93.8364	Gilman et. al. (2011)	GO
Cell-cell signaling	3.9992×10^{-4}	1.7596×10^{-2}	241	188.513	Gai et. al. (2012)	GO
Neuron development	5.9988×10^{-4}	2.6395×10^{-2}	253	207.273	Gilman et. al. (2011)	GO
Axon	9.998×10^{-4}	4.3991×10^{-2}	121	87.635	Gilman et. al. (2011)	GO

Consistent with the synaptic elimination hypothesis, the synaptic elimination set also proved to be significantly enriched for autism *de novo* mutations ($p = 1.9996 * 10^{-4}$). It exceeded the observed-to-expected mutation ratio of all other significantly enriched gene sets (Figure 2).

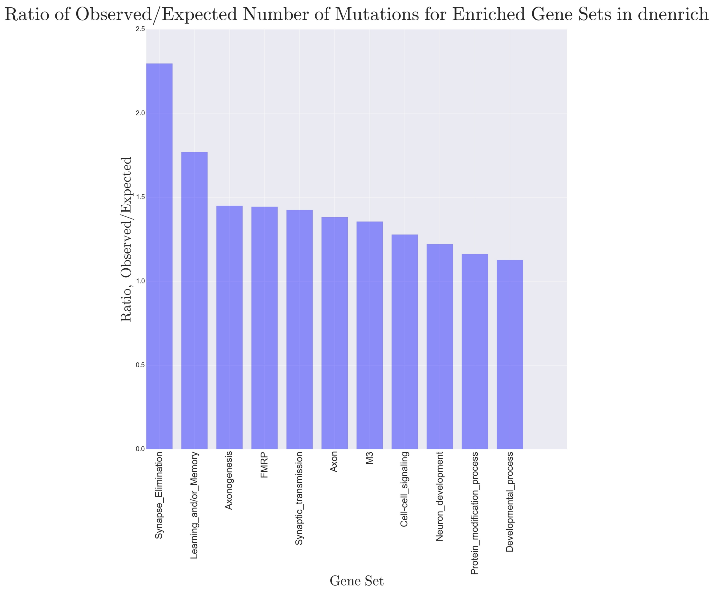


Fig 2. Ratio of observed-to-expected mutations per enriched gene set. The dnenrich software calculated expected number of mutations by simulating and averaging the number of *de novo* events in each gene set using information like tri-nucleotide context, gene size, etc. The systematically-generated synaptic elimination set has the highest ratio of observed-to-expected mutations by a significant margin.

To narrow the list of 213 genes in the synaptic elimination set down to a shorter list of genes to prioritize, we used DAPPLE. Developed by Elizabeth Rossin of the Broad Institute, the algorithm marks genes that are ripe for further study[28]. DAPPLE relies on protein-protein interaction databases such as InWeb (populated with hundreds of thousands of known protein-protein interactions). When researchers input a network, the algorithm compares the network it to what would be expected by pure probability by permuting proteins (linked to the inputted genes) many times. It determines if genes in all of the inputted regions could play a role in disease, and tests whether or not the network is more connected than would be expected by chance. For our purposes, this analysis would point to genes of interest within our synaptic elimination network that have higher levels of interconnectivity than expected.

Using DAPPLE (Figure 3) on the synaptic elimination gene set yielded fifty-four genes significantly enriched for protein-protein interactions (PPI), which are listed in Table 3. Six of these fifty-four (CACNA1C, SHANK2, SYNGAP1, NLGN3, NRXN1, and PTEN) have already been confirmed as genes associated with autism risk[27]. Those genes that were enriched were visualized using the STRING database (Figure 4) in order to examine other known (and predicted) gene interactions. Further inquiry into these resultant genes involved in synaptic elimination could elucidate etiology and shed light on related ASD risk.

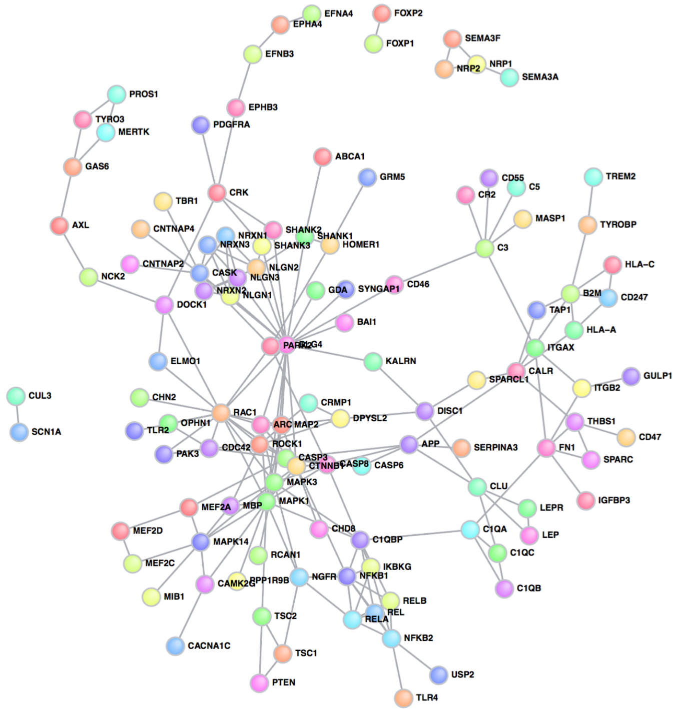


Fig 3. DAPPLE visualization of the synaptic elimination gene set. DAPPLE analyzes the protein-protein interaction network generated by the genes in the set; it marks the genes that are significantly more connected in the network than by chance (PPI-enriched). The nodes represent genes in the network, and the edges represent interactions between proteins downstream of the connected genes. The graphic is arbitrarily colored and is meant to show connectivity only.

Table 3. DAPPLE PPI-Enriched Genes in the synaptic elimination gene set. The table pairs genes with their DAPPLE significances.

Gene Name	*p*-value	Gene Name	*p*-value
DOCK1	0.005985024	CASK	0.001997004
HLA-A	0.045426102	TYRO3	0.005985024
CLU	0.00797604	HOMER1	0.025805363
SHANK1	0.00797604	SEMA3F	0.00797604
CD33	0.005985024	AXL	0.001997004
PARK2	0.003992012	SYNGAP1	0.00996506
ITGAX	0.005985024	CASP3	0.003992012
MERTK	0.045426102	TREM2	0.001997004
ELMO1	0.037601759	MAPK1	0.001997004
OPHN1	0.003992012	CACNA1C	0.035640683
CASP6	0.033677611	SHANK2	0.003992012
MAPK3	0.001997004	SHANK3	0.001997004
NRP1	0.01790118	CALR	0.013937112
NLGN1	0.001997004	NLGN2	0.001997004
GDA	0.011952084	NFKB2	0.013937112
NRXN3	0.001997004	DLG4	0.001997004
C3	0.001997004	TLR2	0.049326298
CTNNB1	0.001997004	ROCK1	0.00996506
CASP8	0.001997004	NRXN1	0.00797604
NGFR	0.001997004	MAP2	0.001997004
REL	0.023832312	EFNB3	0.037601759
ARC	0.001997004	MEF2A	0.01988022
RAC1	0.001997004	B2M	0.011952084
GAS6	0.001997004	NLGN3	0.001997004
NRP2	0.027776419	PTEN	0.00797604
GRM5	0.00996506	THBS1	0.039560839
NRXN2	0.001997004	NFKB1	0.001997004

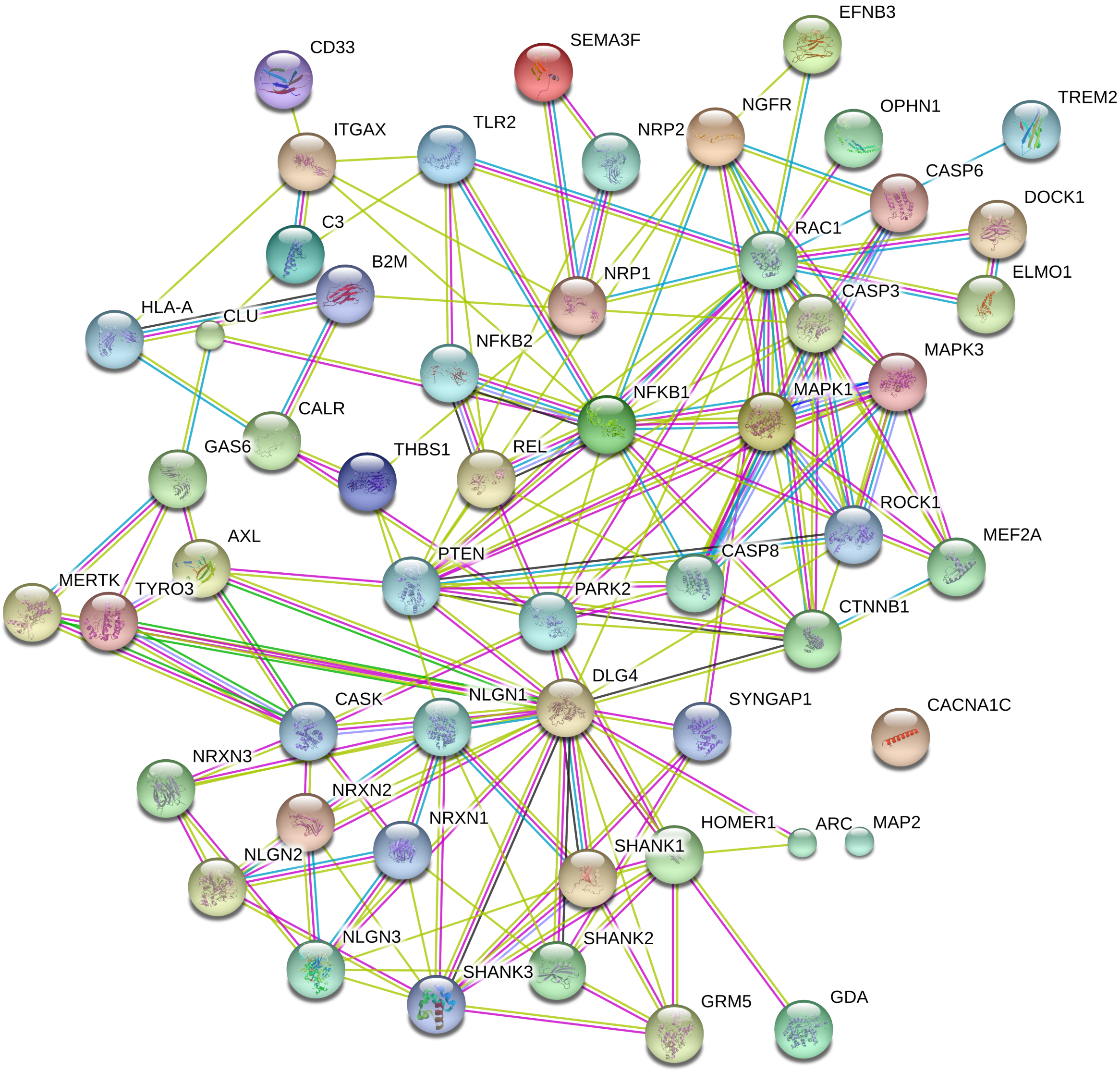


Fig 4. STRING visualization of DAPPLE PPI-enriched genes in the synaptic elimination gene set. Colored nodes represent query proteins and the first shell of interactors. White nodes represent the second shell of interactors. Cyan edges represent known interactions from curated databases; purple edges represent known interactions that are experimentally determined; green edges represent a gene neighborhood predicted interaction; red edges represent a gene fusion predicted interaction; blue edges represent a gene co-occurrence predicted interaction; yellow edges represent textmining; black edges represent co-expression; light blue edges represent protein homology.

4. Discussion

Overall, our results supported the hypothesis that genes involved in synaptic elimination are significantly enriched for autism *de novo* mutations, pointing to deregulation in synaptic elimination as a potential pathogenic mechanism for ASD. Synaptic elimination, as part of the larger synaptic homeostatic mechanism, contributes to higher structural and functional connectivity underlying cognitive functions through the removal of synaptic structures. Several of the genes that were PPI-enriched in the gene set were confirmed autism disease genes, suggesting that the genes central to the synaptic elimination network may play an important role in influencing genetic risk for autism. Given the biological plausibility of this pathway, along with the enrichment for *de novos* in known autism cases, the additional genes in this pathway may serve as candidate genes in the future investigation of the genetic etiology of autism spectrum disorder.

Within the context of the hypothesis that *de novo* mutations contribute to the risk of developing ASD in families with no previous history, previous gene enrichment studies have focused on identification of these *de novo* mutations and their interconnections as a multifaceted network without exploration of specific neurodevelopmental processes[22]. In the present study, we took advantage of a large collection of full exomes from trios with one affected child. This enabled us to explore the role of de novo mutations in synaptic density and pruning, confirming that there is a strong link and supporting the potential value of these de novos for use in increasing precision in early diagnosis/prognosis.

The lack of a validation cohort is a drawback of this study. A new set of *de novos* is currently undergoing quality control procedures; we will attempt to replicate this signal in a much larger collection of families. A consortium that includes our group has amassed over 5000 whole genomes (30x coverage) in multiplex families containing 2 or more children with autism. This is the largest database of its kind and valuable for determining whether the de novo signal seen replicates across siblings and families with varying levels of autism severity. In addition, it may be worthwhile to consider the genes involved in synaptic formation or maintenance in addition to those involved in elimination. Gene sets like the GO Neuron Development or Cell-cell Signaling sets, which showed significant *de novo* mutation enrichment, provide a good starting point for future studies, as neuronal activity and signaling play a definitive role in determining synapse strength and number.

More work is necessary to determine the biological implications of the association between synaptic elimination and autism. For the PPI-enriched genes in the synaptic elimination network, many of which have validated associations with autism, the exact process by which they affect brain development leading to behavioral change is unclear. The true role of these genes in the pathophysiology of autism must be elucidated by future science.

Network analyses like these have successfully been able to identify and validate gene sets that contribute risk to ASD and other neuropsychiatric disorders. The high likelihood that these findings are reproducible in the context of newer, more complete, and more specific datasets bolsters the hope of eventually having a more complete picture of ASD risk factors that impact precision care of this complex disorder. Such a map would be invaluable to both the diagnosis and subsequent treatment of ASD; synaptic elimination may play a key role in that map.

Supplemental Materials

Supplemental Materials 1 – Gene set sizes and gene/gene set mappings – https://drive.google.com/open?id=0B4nOSzAytcrBdlBLVWJWNzFpcGc

Supplemental Materials 2 – synaptic elimination genes and sources – https://drive.google.com/open?id=0B2UCU6mZg1CuSXNKaEJOODNLcmc

Supplemental Materials 3 – synaptic elimination curation references – https://drive.google.com/open?id=0B2UCU6mZg1CudVNiYzJmWGxTWjA

Acknowledgements

This work was supported in part by the Hartwell Foundation's Autism Research and Technology Initiative. It was also conducted with support from a TL1 Clinical Research Training Program of the Stanford Clinical and Translational Science Award to Spectrum (NIH TL1 TR 001084).

References

1 Keen, D. & Ward, S. Autistic spectrum disorder a child population profile. *Autism* **8**, 39-48 (2004).

2 Buescher, A. V., Cidav, Z., Knapp, M. & Mandell, D. S. Costs of autism spectrum disorders in the United Kingdom and the United States. *JAMA pediatrics* **168**, 721-728 (2014).

3 Shattuck, P. T. *et al.* Services for adults with an autism spectrum disorder. *The Canadian Journal of Psychiatry* **57**, 284-291 (2012).

4 Nord, A. S. *et al.* Reduced transcript expression of genes affected by inherited and de novo CNVs in autism. *European Journal of Human Genetics* **19**, 727-731 (2011).

5 Sung, Y. J. *et al.* Genetic investigation of quantitative traits related to autism: use of multivariate polygenic models with ascertainment adjustment. *The American Journal of Human Genetics* **76**, 68-81 (2005).

6 Bourgeron, T. From the genetic architecture to synaptic plasticity in autism spectrum disorder. *Nature Reviews Neuroscience* **16**, 551-563, doi:10.1038/nrn3992 (2015).

7 Tang, G. *et al.* Article Loss of mTOR-Dependent Macroautophagy Causes Autistic-like Synaptic Pruning Deficits. *Neuron* **83**, 1131-1143 (2014).

8 Ebert, D. H. & Greenberg, M. E. Activity-dependent neuronal signalling and autism spectrum disorder. *Nature* **493**, 327-337, doi:10.1038/nature11860.Activity-dependent (2013).

9 Toro, R. *et al.* Key role for gene dosage and synaptic homeostasis in autism spectrum disorders. *Trends in Genetics* **26**, 363-372, doi:10.1016/j.tig.2010.05.007 (2010).

10 Bialas, A. R. & Stevens, B. TGF-Beta Signaling Regulates Neuronal C1q Expression and Developmental Syanptic Refinement. *Nature Neuroscience* **16**, 1773-1782, doi:10.1038/nn.3560.TGF- (2013).

11 Stevens, B. *et al.* The Classical Complement Cascade Mediates CNS Synapse Elimination. *Cell* **131**, 1164-1178, doi:10.1016/j.cell.2007.10.036 (2007).

12 Paolicelli, R. C. & Gross, C. T. Microglia in development: linking brain wiring to brain environment. *Neuron glia biology* **7**, 77-83, doi:10.1017/S1740925X12000105 (2011).

13 Chung, W.-S., Allen, N. J. A. & Eroglu, C. Astrocytes Control Synapse Formation, Function, and Elimination. *Cold Spring Harb Perspect Biol* **7**, doi:10.1530/ERC-14-0411.Persistent (2015).

14 Siegel, A. & Sapru, H. N. *Essential neuroscience*. (Lippincott Williams & Wilkins, 2006).

15 Caglayan, A. O. Genetic causes of syndromic and non-syndromic autism. *Developmental Medicine and Child Neurology* **52**, 130-138, doi:10.1111/j.1469-8749.2009.03523.x (2010).

16 Geschwind, D. H. & Levitt, P. Autism spectrum disorders : developmental disconnection syndromes. *Current Opinion in Neurobiology* **17**, 103-111, doi:10.1016/j.conb.2007.01.009 (2007).

17 Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014).

18 De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).

19 Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-299 (2012).

20 Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014).

21 Parikshak, N. N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008-1021 (2013).

22 O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).

23 Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012).

24 Gilman, S. R. *et al.* Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70**, 898-907 (2011).

25 Gai, X. *et al.* Rare structural variation of synapse and neurotransmission genes in autism. *Molecular Psychiatry* **17**, 402-411, doi:10.1038/mp.2011.10 (2012).

26 Darnell, J. C. *et al.* FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. *Cell* **146**, 247-261, doi:10.1016/j.cell.2011.06.013 (2011).

27 Sanders, S. J. *et al.* Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215-1233 (2015).

28 Rossin, Elizabeth J., et al. "Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology." *PLoS Genet* 7.1 (2011): e1001273.

IDENTIFYING GENETIC ASSOCIATIONS WITH VARIABILITY IN METABOLIC HEALTH AND BLOOD COUNT LABORATORY VALUES: DIVING INTO THE QUANTITATIVE TRAITS BY LEVERAGING LONGITUDINAL DATA FROM AN EHR*

SHEFALI S. VERMA[1], ANASTASIA M. LUCAS[1], DANIEL R. LAVAGE[1], JOSEPH B. LEADER[1], RAGHU METPALLY[2], SARATHBABU KRISHNAMURTHY[1], FREDERICK DEWEY[3], INGRID BORECKI[3], ALEXANDER LOPEZ[3], JOHN OVERTON[3], JOHN PENN[3], JEFFREY REID[3], SARAH A PENDERGRASS[1], GERDA BREITWIESER[2], MARYLYN D. RITCHIE[1]

Department of Biomedical and Translational Informatics, Geisinger Health System, Danville, PA[1]
Department of Functional and Molecular Genomics, Geisinger Health System, Danville, PA[2]
Regeneron Genetics Center, Tarrytown, NY[3]

A wide range of patient health data is recorded in Electronic Health Records (EHR). This data includes diagnosis, surgical procedures, clinical laboratory measurements, and medication information. Together this information reflects the patient's medical history. Many studies have efficiently used this data from the EHR to find associations that are clinically relevant, either by utilizing International Classification of Diseases, version 9 (ICD-9) codes or laboratory measurements, or by designing phenotype algorithms to extract case and control status with accuracy from the EHR. Here we developed a strategy to utilize longitudinal quantitative trait data from the EHR at Geisinger Health System focusing on outpatient metabolic and complete blood panel data as a starting point. Comprehensive Metabolic Panel (CMP) as well as Complete Blood Counts (CBC) are parts of routine care and provide a comprehensive picture from high level screening of patients' overall health and disease. We randomly split our data into two datasets to allow for discovery and replication. We first conducted a genome-wide association study (GWAS) with median values of 25 different clinical laboratory measurements to identify variants from Human Omni Express Exome beadchip data that are associated with these measurements. We identified 687 variants that associated and replicated with the tested clinical measurements at $p<5x10^{-08}$. Since longitudinal data from the EHR provides a record of a patient's medical history, we utilized this information to further investigate the ICD-9 codes that might be associated with differences in variability of the measurements in the longitudinal dataset. We identified low and high variance patients by looking at changes within their individual longitudinal EHR laboratory results for each of the 25 clinical lab values (thus creating 50 groups – a high variance and a low variance for each lab variable). We then performed a PheWAS analysis with ICD-9 diagnosis codes, separately in the high variance group and the low variance group for each lab variable. We found 717 PheWAS associations that replicated at a p-value less than 0.001. Next, we evaluated the results of this study by comparing the association results between the high and low variance groups. For example, we found 39 SNPs (in multiple genes) associated with ICD-9 250.01 (Type-I diabetes) in patients with high variance of plasma glucose levels, but not in patients with low variance in plasma glucose levels. Another example is the association of 4 SNPs in *UMOD* with chronic kidney disease in patients with high variance for aspartate aminotransferase (discovery p-value: $8.71x10^{-09}$ and replication p-value: $2.03x10^{-06}$). In general, we see a pattern of many more statistically significant associations from patients with high variance in the quantitative lab variables, in comparison with the low variance group across all of the 25 laboratory measurements. This study is one of the first of its kind to utilize quantitative trait variance from longitudinal laboratory data to find associations among genetic variants and clinical phenotypes obtained from an EHR, integrating laboratory values and diagnosis codes to understand the genetic complexities of common diseases.

* This work is supported by funds from Geisinger Health System and the Regeneron Genetics Center. Supplementary material can be found at: http://ritchielab.psu.edu/publications/supplementary-data/psb-2017/CBC-Met-Labs.

1. Introduction

In this era of personalized medicine, emphasis is on preventive care facilitated by integration of a patient's medical and genomic information. De-identified electronic health records (EHR) and bio-repositories represent significant resources of information that have been widely used for association studies in past decade[1]. Electronic health record (EHR) data is primarily designed for clinical care and is represented in both structured (such as ICD-9 codes, medication information, clinical laboratory values) as well as unstructured (physician notes) forms. Many association studies have utilized ICD-9 codes as well as clinical lab variables (structured forms of EHR data) to identify variants associated with EHR-derived phenotypes that might be of clinical relevance[2–4]. The number of association studies using EHR-derived phenotypes (both structured and unstructured data) has been increasing rapidly[5].

The complete blood count (CBC) panel and comprehensive/basic metabolic panel (CMP/BMP) are part of routine medical care for all medical practices. These panels are comprised of tests that help clinical practitioners identify underlying causes for conditions like weakness and fatigue, as well as to identify chronic illnesses (e.g., kidney failure, heart disease). These tests are generally conducted on patients that show some signs of illness, but these routine measurements are conducted from time to time on healthy individuals as well. Thus, utilizing these panels can help us understand overall health of patients by comparing these measurements across all patients in an EHR. These tests are recorded as quantitative variables for which units of measurements can be standardized across multiple clinical practices. ICD-9 codes and clinical measurements go hand in hand for a patient's medical record as a diagnosis code may either initiate the lab test which confirms the code or the code may be entered as a result of the test. Thus, integrating both clinical laboratory measurements and diagnosis codes present powerful approaches for understanding genetic variants that show similar associations with both data types obtained from an EHR[3]. The majority of association studies that use quantitative traits derived from an EHR as phenotypes use either mean/median values[3,6] or most recent measurements[7]. While this approach has been successful, utilizing only mean/median values limits the understanding of these traits by neglecting the variability over time that may be present in an individual patient's clinical history. This can be captured for analysis by using unique longitudinal information from EHR. Longitudinal data provides a better picture of the patient's health by actually pinpointing the time of disease onset, or time in which the quantitative trait became out of the normal range, which is especially important for the diseases that are more heterogeneous in nature and progress over time/age. A strategy such as this has been applied to family-based studies, using a mixed effects model to find associations among candidate genes and longitudinal data[8]. Utilizing the longitudinal data in some way other than considering one value also provides the opportunity to consider not just the average, but also the variability in these traits over time. In this study, our goal was to develop a strategy to embrace the longitudinal data in a population-based dataset, using trait variance, rather than a measure of central tendency approach such as median values, by binning patients in high and low variance groups separately to then test for associations. This strategy allows for the integration of clinical lab measurements as quantitative traits, embracing the variability in the traits, along with ICD-9 code PheWAS associations as well as SNPs.

2. Materials and Methods

2.1 Genotype Data

The MyCode® Community Health Initiative is a research initiative to engage Geisinger Health System patients in research and integrate their clinical EHR data along with genetic information to make discoveries in health and disease[9]. Over 109,000 Geisinger patients have consented to participate in MyCode and approximately 50,000 participants have whole exome sequencing and genome-wide genotype data generated. For this study, we used participants that have been genotyped using the Illumina Human Omni Express plus Exome beadchip. This dataset contains 45,899 samples and ~600K variants after some initial quality control procedures. For this analysis, after sample QC (removing one sample from pairs of highly related samples up to 1st cousins and removing any samples that did not pass a sample call rate filter of 90%), we divided the total dataset into two random sets to perform discovery and replication analyses. We included only European American samples with age >18 years. Our discovery dataset consisted of 17,347 samples and our replication dataset consisted of 17,348 samples (see **Supplementary Table** 1 for demographic information on these samples). We also filtered the variants that did not pass a genotype call rate filter of 99% to keep only high quality SNP data. To test common variants only, we applied a minor allele frequency (MAF) filter of 1%. This resulted in a total of 629,274 variants that were considered for association testing in the discovery dataset and 629,016 variants tested in the replication dataset.

2.2 Phenotype Data

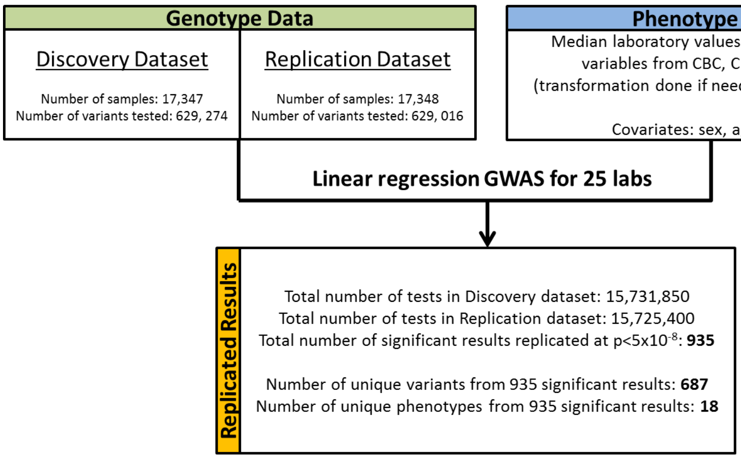
Twenty-five clinical laboratory variables were extracted from EHR outpatient data and checked for consistency of unit measurements. A list of all 25 variables is provided in Table 1, along with information on the panel from which they were obtained. The phenotype data is extracted from the EHR as longitudinal data for all patients across their clinical history. Thus, each sample has multiple entries for each variable. The first step in conducting our GWAS analysis was to obtain median values for all 25 variables across patients' longitudinal data. We wanted to be able to compare the GWAS on median values with the analyses in the high-variance and low-variance groups. We visually inspected the clinical lab variable distributions to determine which variables needed a natural log transformation. We also removed all outliers that were more than 2.5 standard deviations from the mean. While this could lose some very interesting data points, for this pilot analysis, we wanted to be sure to remove gross errors in lab variable coding/data entry. Supplementary Figure 1 and 2 show the distribution of discovery and replication datasets, respectively, after removing outliers and performing natural log transformation wherever necessary. Table 1 lists the name of the variable, how the sample is collected (i.e. Blood or Serum/Plasma), which panel the variable is obtained from (i.e. Complete Blood Count (CBC) or Comprehensive Metabolic Panel (CMP) or Basic Metabolic Panel (BMP)), the total sample size for each phenotype in both discovery and replication datasets, and whether or not the data were transformed.

Table 1. List of 25 clinical laboratory measurements that are used in the analysis.

Clinical Laboratory Measurement	Panel type	Discovery Sample Size	Replication Sample Size	Transformation
ALANINE AMINOTRANSFERASE (ALT) - SERUM/PLASMA	CMP	15527	15393	Yes
ALBUMIN - SERUM/PLASMA	CMP	15519	15439	Yes
ALKALINE PHOSPHATASE - SERUM/PLASMA	CMP	15189	15088	Yes
ANION GAP - SERUM/PLASMA	BMP/CMP	15954	15849	No
ASPARTATE AMINOTRANSFERASE (AST) - SERUM/PLASMA	CMP	15406	15310	Yes
BILIRUBIN - SERUM/PLASMA	CMP	15224	15141	Yes
CALCIUM (CA) - SERUM/PLASMA	BMP/CMP	16164	16098	No
CARBON DIOXIDE (CO2) - SERUM/PLASMA	BMP/CMP	16309	16203	No
CHLORIDE (CL) - SERUM/PLASMA	BMP/CMP	16235	16130	No
CREATININE - SERUM/PLASMA	BMP/CMP	16403	16323	Yes
Erythrocyte Distribution Width (RDW) - BLOOD	CBC	16032	15974	Yes
GLUCOSE - SERUM/PLASMA	BMP	16184	16137	Yes
Hematocrit (HCT) - BLOOD	CBC	16213	16184	No
HEMOGLOBIN - BLOOD	CBC	16234	16186	No
Mean Corpuscular Hemoglobin (MCH) - BLOOD	CBC	16175	16120	No
Mean Corpuscular Hemoglobin Concentration (MCHC) - BLOOD	CBC	16166	16114	No
Mean Corpuscular Volume (MCV) - BLOOD	CBC	16220	16161	No
PLATELET - BLOOD - COUNT	CBC	16122	16099	No
Platelet Mean Volume (MPV) - BLOOD	CBC	16281	16247	No
POTASSIUM (K) - SERUM/PLASMA	BMP/CMP	16255	16165	No
PROTEIN - SERUM/PLASMA	CMP	15002	14932	No
RBC-COUNT-BLOOD	CBC	16187	16142	No
SODIUM (NA) - SERUM/PLASMA	BMP/CMP	16222	16144	No
UREA NITROGEN - SERUM/PLASMA	BMP/CMP	16147	16049	No
WBC-COUNT-BLOOD	CBC	16478	16455	Yes

For the variance based analysis, we first calculated the variance for each sample across their longitudinal clinical data from EMR. For each clinical lab variable, we visually inspected scatterplots of the variance distribution and determined a threshold for discovery and replication datasets separately (**Supplementary Table 2**). Next, samples were divided into high and low variance groups. For the high-variance/low-variance PheWAS analyses, we extracted all ICD-9 codes from the EHR. Participants were defined as cases if they had 3 or more instances of a particular ICD-9 code; less than 3 instances per participant were set to missing; and for no occurrence of an ICD-9 code, participants were designated control status. This resulted in testing a total of 541 ICD-9 codes.

Figure 1. Flow chart describing the analyses for median lab variable linear regression GWAS on 25 clinical labs


Genotype Data
Discovery Dataset
Number of samples: 17,347
Number of variants tested: 629, 274
Replication Dataset
Number of samples: 17,348
Number of variants tested: 629, 016
Phenotype Data
Median laboratory values for 25 clinical lab variables from CBC, CMP/BMP labs
(transformation done if needed, refer to Table 1
Covariates: sex, age, PC1-4
Linear regression GWAS for 25 labs
Replicated Results
Total number of tests in Discovery dataset: 15,731,850
Total number of tests in Replication dataset: 15,725,400
Total number of significant results replicated at p<5x10-8: 935
Number of unique variants from 935 significant results: 687
Number of unique phenotypes from 935 significant results: 18

2.3 Analysis Methods

We performed the analysis for this study as a two-step process. First we performed a GWAS on median values for 25 different clinical lab variables (**Figure 1**). Next, we took the SNPs associated with the median trait values and performed an ICD-9 code PheWAS after grouping the participants into high-variance and low-variance groups for each clinical lab variable (Figure 2). Each of these analyses is described in more detail in the following sections.

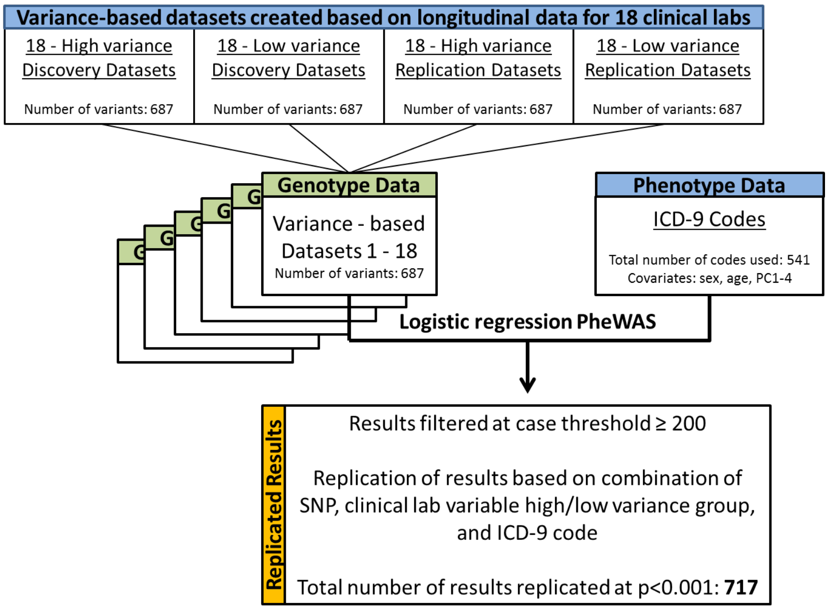
2.3.1 Genome wide association analysis for 25 median clinical laboratory measurement

We performed a genome-wide association study (GWAS) to identify associations among all variants from the data (after quality control data cleaning) with median lab values for each of the 25 phenotypes. Linear regression analysis was performed using PLATO[10] (http://ritchielab.psu.edu/software/plato-download). All models were adjusted for age, sex and first 4 principal components to control for confounding influences in the analysis. Approximately 15M (~600,000 SNPs and 25 variables) tests were performed for each patient for both discovery and replication datasets. This analysis was repeated for both discovery and replication datasets separately and then we identified p-values for all variant and clinical lab combinations that were below genome-wide significance (p-value $5\text{x}10^{-8}$) in both datasets (discovery and replication).

2.3.2 *Variance-based analysis to identify associations with ICD-9 codes*

For all phenotypes from the median lab GWAS that has statistically significant replicating results (18 out of the 25 clinical lab variables, see **Figure 1**), we obtained longitudinal data for each patient across the EHR and calculated the trait variance for each lab variable. Next, for each of the 18 variables, we created scatterplots of the variance to identify samples that can be categorized as high and low variance. Individual scatter plots for all of these variables are shown in **Supplementary Figure 3 and 4** for the discovery and replication datasets. For each variable, we created high variance and low variance groups based on a user-defined threshold to allow for PheWAS analyses separately in groups with high variability or low variability in each of the clinical lab variables. **Supplementary Table 2** lists the thresholds and samples sizes for low and high variance categories in both discovery and replication datasets. Participants below the chosen thresholds (based on looking at individual scatterplots) were categorized as low variance and above threshold were categorized as high variance.

Figure 2. Flow chart describing the PheWAS analyses for high/low variance based datasets

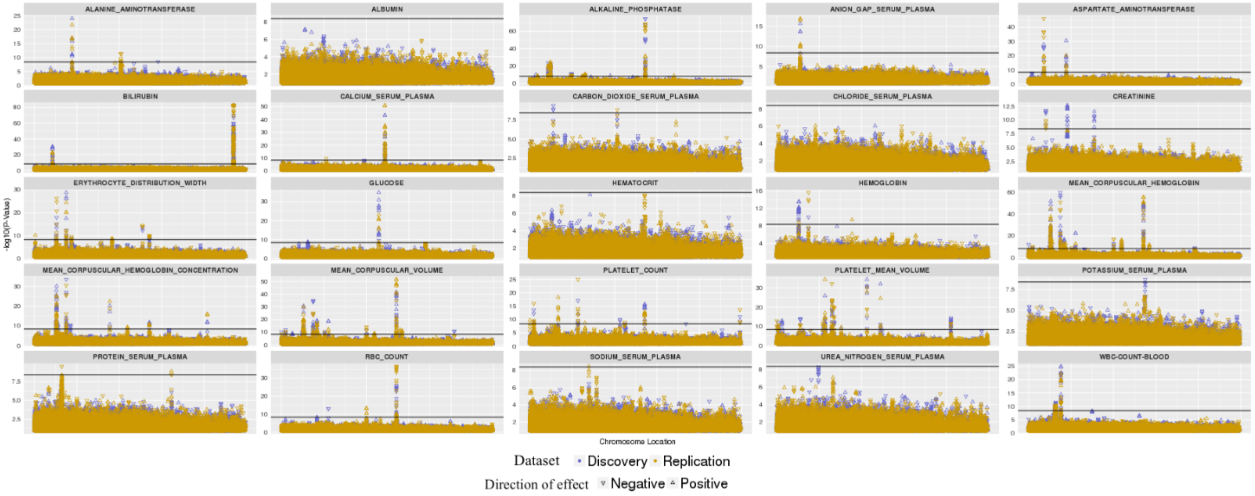

Variance-based datasets created based on longitudinal data for 18 clinical labs
18 - High variance Discovery Datasets
Number of variants: 687
18 - Low variance Discovery Datasets
Number of variants: 687
18 - High variance Replication Datasets
Number of variants: 687
18 - Low variance Replication Datasets
Number of variants: 687
Genotype Data
Variance - based Datasets 1 - 18
Number of variants: 687
Phenotype Data
ICD-9 Codes
Total number of codes used: 541
Covariates: sex, age, PC1-4
Logistic regression PheWAS
Replicated Results
Results filtered at case threshold ≥ 200
Replication of results based on combination of SNP, clinical lab variable high/low variance group, and ICD-9 code
Total number of results replicated at p<0.001: 717

The genotype data was filtered to include only those variants (687 SNPs) that were significantly associated in both the discovery and replication datasets for one or more clinical lab variables in the GWAS of median clinical lab values. Here, we are interested in the following question: Are genetic variants that are associated with a median clinical lab variable, also associated with diagnosis codes in patients with high variability or low variability in that lab variable? In other words, are there diseases that show association with that SNP in patients who are highly variable in their lab values or perhaps have low variability in their lab values? To investigate diagnosis codes that are associated with these variants, we performed logistic regression analysis for ICD-9 codes using PLATO (http://ritchielab.psu.edu/software/plato-download) by adjusting all models by age, sex and first 4 principal components. We only considered ICD-9 codes that had at least 200 or more cases with the code to reduce any false positive associations. Thus, for each sample 371,667 tests were erformed (687 SNPs and 541 ICD-9 codes). Lastly, we report the PheWAS results below a p-value threshold of 0.001 that replicate in low variance and/or high variance categories.

3. Results

Genome-wide association studies for median values from 25 clinical laboratory variables produced 935 SNP-phenotype associations that are present in discovery and replication sets at p-value less than 5x10^{-8}. Association results below p-value 0.1 are shown in **Figure 3** as Manhattan plots for both discovery and replication datasets. Among the top results are multiple variants in the *UGT1A* gene family associated with serum bilirubin levels, where p-values for both discovery and replication datasets is 3.29 x 10^{-83}. This association has been identified and extensively reported by candidate gene and genome-wide association studies[11]. Hyperbilirubinemia results from a mutation in the *UGT1A1* gene which causes the non- or slow elimination of bilirubin from the body. We also identified variants in *SLCO1B1* associated with bilirubin levels, as suggested by previous GWAS studies [12–14] (rs4149081, Discovery p-value: 8.18x10^{-31} Replication p-value:3.81x10^{-22}).
Another association we identified is between missense variant, rs855791, on chromosome 22 in gene *TMPRSS6* (Discovery p-value: 2.04x10^{-60} (beta=-0.27); Replication p-value: 1.73x10^{-51} (beta=-0.25)). This association was identified by previous GWAS studies with hemoglobin levels as well as hemoglobin concentration[15,16]. It has been suggested that *TMPRSS6* is essential for maintaining iron levels in blood as it is involved in the control of iron homeostasis [16,17]. In addition, our GWAS analyses also identified many more previously reported associations, including variants

Figure 3. Manhattan plots for GWAS performed on all 25 clinical lab variables. X-axis represents the chromosome and base pair location of each SNP and Y-axis represent the –log10 of p-value from association analysis. The two colors represent p-value for discovery and replication datasets. Direction of effect (positive or negative) is shown by the direction of arrows. Results at p-value <0.1 are shown in the plot. Black line indicates genome-wide significance (5e-08) threshold.



in the *ABO* gene with alkaline phosphatase[18] (rs505922, discovery p-value: 2.41x10^{-52}, replication p-value:8.48x10^{-65}), the *CASR* gene with calcium levels[19,20] (rs17251221, discovery p-value: 6.55x10^{-44}, replication p-value:2.31x10^{-51}), and the *TCF7L2* gene with glucose levels[21] (rs7903146, discovery p-value: 1.41x10^{-35}, replication p-value:6.23x10^{-24}).

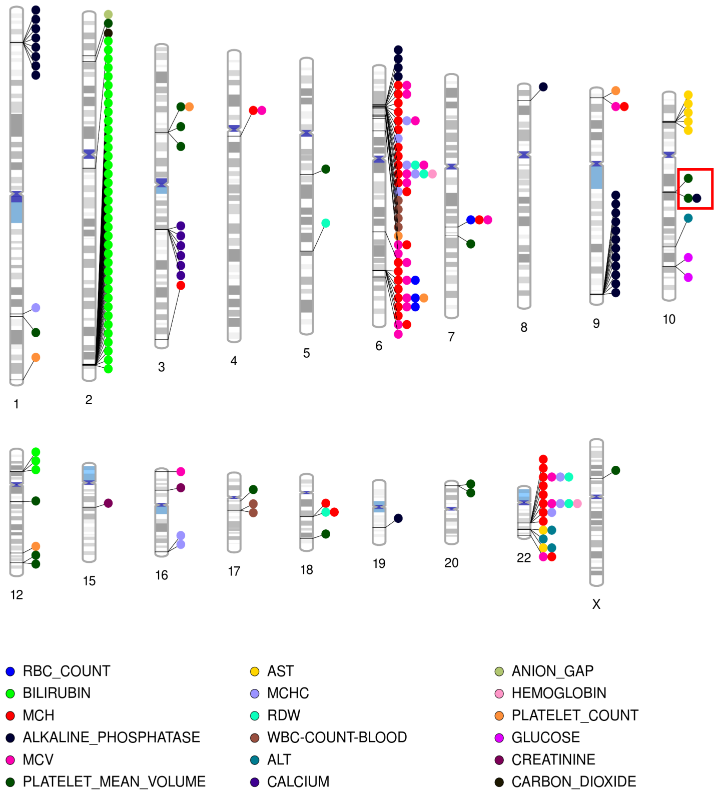


Figure 4. Phenogram plot representing pleiotropic associations. Here each colored circle is a SNP and its location is represented on the chromosome. SNPs are color coded based on the phenotype colors as shown in the legend. SNPs are also pruned to LD threshold of 0.4. Here MCH is Mean Corpuscular Hemoglobin; MCHC is MCH is Mean Corpuscular Hemoglobin Concentration; AST is Aspartate Aminotransferase; RDW is Erythrocyte Distribution Width; ALT is Alanine Aminotransferase.

To explore pleiotropic associations among variants where one SNP is associated with multiple phenotypes, we generated a phenogram plot[22] shown in **Figure 4.** This plot shows, for example, multiple associations on chromosome 10 in gene *JMJD1C* to be associated with platelet mean volume as well as alkaline phosphatase (red box on **Figure 4**). Different GWAS studies performed separately on blood and metabolic panels have identified these associations[23,24] and our study serves as confirmation for these associations when both panels are combined together and analysis is run on the same patients. In our analysis, we see opposite directions of effect for both of these associations, i.e. erythrocyte distribution width (discovery beta: -0.004 and replication beta: -0.004) and mean corpuscular hemoglobin (discovery beta: 0.09 and replication beta: 0.12) which confirms the relationship observed in anemic patients, where elevation in RDW and decrease in hemoglobin is observed.

Among our novel associations are intronic variant rs8095374 in gene *C18orf25* associated with erythrocyte distribution width known as RDW (discovery p-value: 8.79x10^{-10}, and replication p-value: 2.16x10^{-10}) and mean corpuscular hemoglobin (discovery p-value: 3.57x10^{-9}, and replication p-value: 1.84x10^{-13}). Both laboratory measurements are for red blood cells and could be useful in understanding the etiology of anemia.

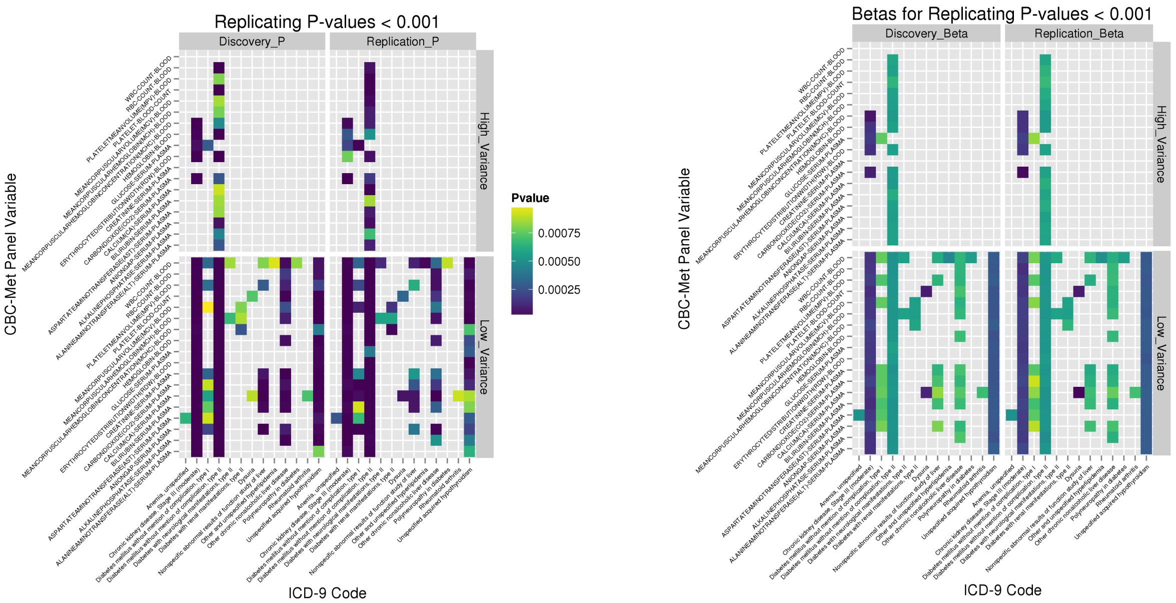


Figure 5. Heat map representing p-values (on left) and beta (on right) from variance based analysis for the combination of a SNP, ICD-9 code and clinical lab measurement in both high and low variance categories. Each point is the replicating SNP with the color gradient showing the range of p-value and beta. The results are only shown for replicating results at p-value<0.001 for both discovery and replication datasets in both high variance and low variance categories. X-axis lists all the ICD-9 codes and Y-axis lists the corresponding clinical lab variable for which replicating association is observed.

Our next approach was to integrate ICD-9 code data along with clinical lab variables to identify variants that we have found to be associated with median values of quantitative traits, and are *also* linked to diagnosis codes in the EHR. To perform this analysis, we wanted to utilize longitudinal data, rather than a measure from a single point in time. Hence, we divided patients into categories of high and low variance as described in *Methods*. Replication was observed based on the combination of SNP, clinical lab variable, ICD-9 code, and variance category (high or low). Replicated results are shown in form of a heat map in **Figure 5**. These heat maps show that in our study, the majority of our replicating associations occur in the low variance category. The primary reason for this is likely due to low sample size in the high variance groups gave us less statistical power to detect associations; although we would like to continue to explore this to determine whether there is a biological explanation for this. In total, this analysis resulted in 717 replicated associations.

We observed 39 SNPs on chromosome 6 that map to multiple genes (*C6orf10, FKBPL, BAT3, BAT2, EGFL2, RDBP, MSH5, TNXB*, C6orf27, *CSNK2B and BAT1*) are associated with Type 1 Diabetes (ICD-9 code 250.01) when the samples with high variance glucose levels were evaluated. These associations were not seen in samples in the low variance glucose category. One of the most interesting associations identified is between four SNPs in the *uromodulin (UMOD)* gene and ICD-9 code 585.3 (Chronic kidney disease) in patients with low variance for aspartate aminotransferase (discovery p-value: 8.71×10^{-9} and replication p-value: 2.03×10^{-6}). It has been observed by previous studies that patients with chronic kidney disease usually have low levels of aminotransferase in serum[25]. This association was not replicated in the high variance aspartate aminotransferase group. Association of variants in the *UMOD* gene with chronic kidney disease, kidney stones, and end stage renal diseases has been previously established[26,27] but an association with aspartate aminotransferase levels has not been identified by previous studies. Next, to integrate both the GWAS results and variance-based grouping PheWAS results, we generated networks of all genome-wide significant results from GWAS analysis and replicated results from variance based PheWAS analysis using Cytoscape[28] as shown in **Figure 6**. We explored the integrated results for SNP-Clinical lab variable- ICD-9 code in order to identify the three-way associations that are indicative of disease diagnosis. This figure shows the three top integrated networks from our analysis where both ICD-9 codes and clinical lab variables are linked via a SNP. One thing to note here is that all these networks resulted from the low variance groups only.

From the network visualization, we determined three variants in gene *TCF7L2* are associated with Type 2 Diabetes (T2D) and glucose levels. This association is expected because these variants have been reported by many previous studies to be associated with T2D[21,29,30]. Similarly, from this network analysis we also observed variants in the *UMOD* gene associated with chronic kidney disease and creatinine levels obtained from serum which has been previously reported by GWAS[26,27,31]. Lastly, a novel network obtained from this analysis is a link between rs3132941 (mapped to gene, *EGFL8*) with WBC count and Type I Diabetes. A high WBC has been observed in a few studies in T1D patients[32,33]. The *EGFL8* gene maps near the MHC region (Major-histocompatibility complex) on chromosome 6 and thus its association with T1D can be easily established[34,35] but its association with WBC has not been found in any previous studies. Our study presents this novel result which warrants further investigation.

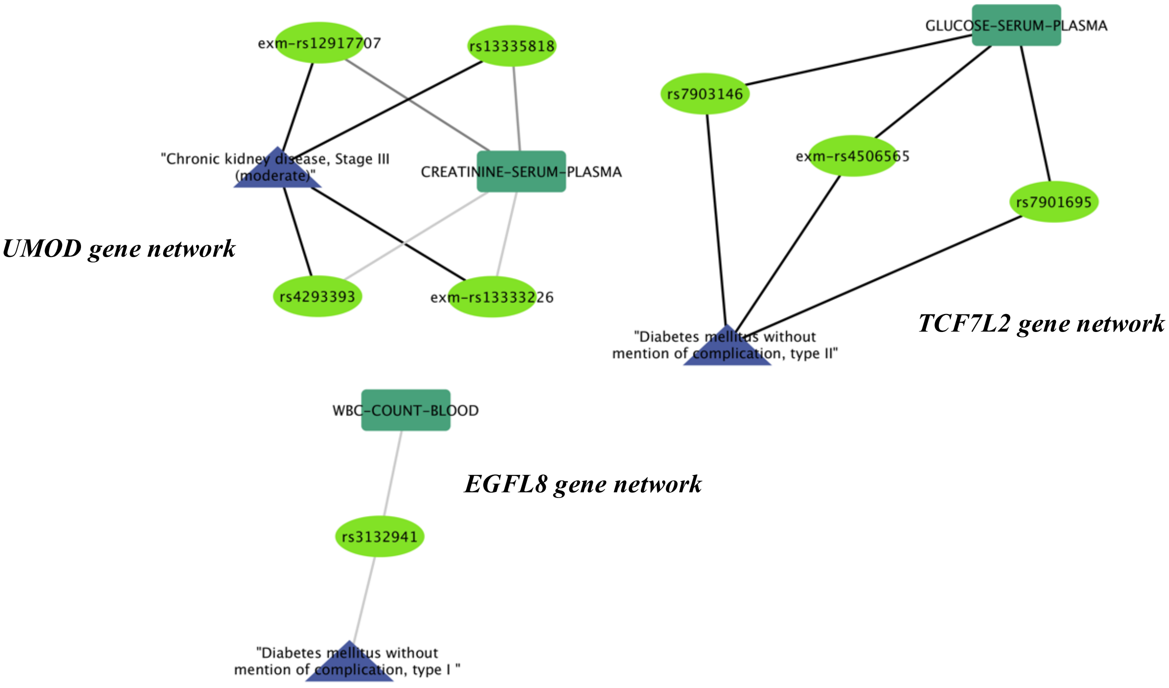


Figure 6. Network visualization generated by Cytoscape using replicated results from both GWAS and variance based analysis. Here, triangles represent ICD-9 code description, rectangles represent clinical lab variable, and ovals represent SNP. Darker edges represent more significant associations.

4. Discussion

Genome-wide association studies have been tremendously successful in unravelling the etiologies of common complex diseases and the use of EHR in conducting such genome-wide and phenome-wide studies has shown resounding progress. Many researchers are now working on approaches to incorporate longitudinal information from the EHR into these studies. As a proof of concept, in this study we aimed at advancing the use of longitudinal information from laboratory values by looking at the variance for each outpatient clinical lab value rather than just mean/median or most recent value. We first conducted a GWAS for 25 clinical lab median values and then, based on variance, we divided participants into high and low variance groups. Next, we conducted a PheWAS to identify which SNPs are associated with median clinical lab variable *and* ICD-9 codes. This study represents a proof-of concept approach for utilizing trait variance and the longitudinal data as we successfully identified and confirmed many previously known associations. We also described several novel associations observed from our study. Variance, rather than mean/median may better capture the richness of the longitudinal data. In this pilot analysis, we demonstrate that this approach can be used to identify networks which reveal trends of associations among SNPs, laboratory measurements, and diagnosis codes. In the future, we plan to replicate this analysis with a larger sample size and in an independent EHR system. We also plan to use variance as the outcome for an association study in all 50,000 patients from Geisinger MyCode dataset and replicate in an independent dataset. One limitation of our approach here is that the use of longitudinal data in the way shown in this study ignores the fact that in an EHR, the duration of longitudinal information varies from patient to patient. Future approaches should also focus on developing methods which adjust for the duration of longitudinal information. Developing approaches, such as the one described in this manuscript, to explore the longitudinal nature of EHR data will provide greater opportunities for discovery and understanding of the genetic and clinical architecture of common diseases.

5. References

1. Manolio, T. A. Biorepositories--at the bleeding edge. *Int J Epidemiol* **37,** 231–233 (2008).
2. Moore, C. B. *et al.* Phenome-wide Association Study Relating Pretreatment Laboratory Parameters With Human Genetic Variants in AIDS Clinical Trials Group Protocols. *Open Forum Infectious Diseases* **2,** ofu113–ofu113 (2015).
3. Verma, A. *et al.* INTEGRATING CLINICAL LABORATORY MEASURES AND ICD-9 CODE DIAGNOSES IN PHENOME-WIDE ASSOCIATION STUDIES. *Pac Symp Biocomput* **21,** 168–179 (2016).
4. Namjou, B. *et al.* Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. *Front Genet* **5,** (2014).
5. Wei, W.-Q. & Denny, J. C. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med* **7,** (2015).
6. Kullo, I. J., Ding, K., Jouni, H., Smith, C. Y. & Chute, C. G. A Genome-Wide Association Study of Red Blood Cell Traits Using the Electronic Medical Record. *PLoS One* **5,** (2010).
7. Namjou, B. *et al.* EMR-linked GWAS study: investigation of variation landscape of loci for body mass index in children. *Front Genet* **4,** (2013).
8. Luan, J. ’an *et al.* A multilevel linear mixed model of the association between candidate genes and weight and body mass index using the Framingham longitudinal family data. *BMC Proc* **3 Suppl 7,** S115 (2009).
9. Carey, D. J. *et al.* The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet Med* (2016). doi:10.1038/gim.2015.187

10. Grady, B. J. *et al.* Finding unique filter sets in PLATO: a precursor to efficient interaction analysis in GWAS data. *Pac Symp Biocomput* 315–326 (2010).
11. Lin, J.-P. *et al.* Association between the UGT1A1*28 allele, bilirubin levels, and coronary heart disease in the Framingham Heart Study. *Circulation* **114,** 1476–1481 (2006).
12. de Azevedo, L. A. *et al.* UGT1A1, SLCO1B1, and SLCO1B3 polymorphisms vs. neonatal hyperbilirubinemia: is there an association? *Pediatr Res* **72,** 169–173 (2012).
13. Kang, T.-W. *et al.* Genome-wide association of serum bilirubin levels in Korean population. *Hum. Mol. Genet.* **19,** 3672–3678 (2010).
14. Johnson, A. D. *et al.* Genome-wide association meta-analysis for total serum bilirubin levels. *Hum. Mol. Genet.* **18,** 2700–2710 (2009).
15. Chambers, J. C. *et al.* Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. *Nat Genet* **41,** 1170–1172 (2009).
16. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492,** 369–375 (2012).
17. Benyamin, B. *et al.* Common variants in TMPRSS6 are associated with iron status and erythrocyte volume. *Nat Genet* **41,** 1173–1175 (2009).
18. Li, J. *et al.* Genome-wide association study on serum alkaline phosphatase levels in a Chinese population. *BMC Genomics* **14,** 684 (2013).
19. Bonny, O. & Bochud, M. Genetics of calcium homeostasis in humans: continuum between monogenic diseases and continuous phenotypes. *Nephrol. Dial. Transplant.* **29,** iv55–iv62 (2014).
20. Kapur, K. *et al.* Genome-Wide Meta-Analysis for Serum Calcium Identifies Significantly Associated SNPs near the Calcium-Sensing Receptor (CASR) Gene. *PLOS Genet* **6,** e1001035 (2010).
21. Billings, L. K. & Florez, J. C. The genetics of type 2 diabetes: what have we learned from GWAS? *Ann N Y Acad Sci* **1212,** 59–77 (2010).
22. Wolfe, D., Dudek, S., Ritchie, M. D. & Pendergrass, S. A. Visualizing genomic information across chromosomes with PhenoGram. *BioData Min* **6,** 18 (2013).
23. Yuan, X. *et al.* Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am. J. Hum. Genet.* **83,** 520–528 (2008).
24. Qayyum, R. *et al.* A meta-analysis and genome-wide association study of platelet count and mean platelet volume in african americans. *PLoS Genet.* **8,** e1002491 (2012).
25. Ray, L., Nanda, S. K., Chatterjee, A., Sarangi, R. & Ganguly, S. A comparative study of serum aminotransferases in chronic kidney disease with and without end-stage renal disease: Need for new reference ranges. *Int J Appl Basic Med Res* **5,** 31–35 (2015).
26. Gudbjartsson, D. F. *et al.* Association of Variants at UMOD with Chronic Kidney Disease and Kidney Stones—Role of Age and Comorbid Diseases. *PLOS Genet* **6,** e1001039 (2010).
27. Reznichenko, A. *et al.* UMOD as a susceptibility gene for end-stage renal disease. *BMC Medical Genetics* **13,** 78 (2012).
28. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27,** 431–432 (2011).
29. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
30. Lyssenko, V. *et al.* Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J. Clin. Invest.* **117,** 2155–2163 (2007).
31. Pattaro, C. *et al.* A meta-analysis of genome-wide data from five European isolates reveals an association of COL22A1, SYT1, and GABRR2 with serum creatinine level. *BMC Med. Genet.* **11,** 41 (2010).
32. Xu, W. *et al.* Correlation between Peripheral White Blood Cell Counts and Hyperglycemic Emergencies. *Int J Med Sci* **10,** 758–765 (2013).
33. Twig, G. *et al.* White Blood Cells Count and Incidence of Type 2 Diabetes in Young Men. *Diabetes Care* **36,** 276–282 (2013).
34. Nejentsev, S. *et al.* Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature* **450,** 887–892 (2007).
35. Abraham, R. S., Wen, L., Marietta, E. V. & David, C. S. Type 1 Diabetes-Predisposing MHC Alleles Influence the Selection of Glutamic Acid Decarboxylase (GAD) 65-Specific T Cells in a Transgenic Model. *J Immunol* **166,** 1370–1379 (2001).

STRATEGIES FOR EQUITABLE PHARMACOGENOMIC-GUIDED WARFARIN DOSING AMONG EUROPEAN AND AFRICAN AMERICAN INDIVIDUALS IN A CLINICAL POPULATION

LAURA K. WILEY
Div. of Biomedical Informatics and Personalized Med., University of Colorado, 13001 E. 17th Pl. MS F-563
Aurora, CO 80045, USA
Email: laura.wiley@ucdenver.edu

JACOB P. VANHOUTEN
Dept. of Biomedical Informatics, Vanderbilt University, 2525 West End Ave Ste. 1475
Nashville, TN 37203, USA
Email: jacob.p.vanhouten@vanderbilt.edu

DAVID C. SAMUELS
Dept. of Mol. Physiology & Biophysics, Vanderbilt Genetics Inst., Vanderbilt University, 2215 Garland Ave.
Nashville, TN 37232, USA
Email: david.c.samuels@vanderbilt.Edu

MELINDA C. ALDRICH
Dept. of Thoracic Surgery, Div. of Epidemiology, Vanderbilt University Medical Center, 609 Oxford House
Nashville, TN 37232, USA
Email: melinda.aldrich@vanderbilt.edu

DAN M. RODEN
Dept. of Medicine, Vanderbilt University, 2215B Garland Ave
Nashville, TN 37203, USA
Email: dan.roden@vanderbilt.edu

JOSH F. PETERSON
Dept. of Biomedical Informatics, Dept. of Medicine, Vanderbilt University, 2525 West End Ave Ste.1050
Nashville, TN 37203, USA
Email: josh.peterson@vanderbilt.edu

JOSHUA C. DENNY
Dept. of Biomedical Informatics, Dept. of Medicine, Vanderbilt University, 2525 West End Ave Ste.1475
Nashville, TN 37203, USA
Email: josh.denny@vanderbilt.edu

The blood thinner warfarin has a narrow therapeutic range and high inter- and intra-patient variability in therapeutic doses. Several studies have shown that pharmacogenomic variants help predict stable warfarin dosing. However, retrospective and randomized controlled trials that employ dosing algorithms incorporating pharmacogenomic variants under perform in African Americans. This study sought to determine if: 1) including additional variants associated with warfarin dose in African Americans, 2) predicting within single ancestry groups rather than a combined population, or 3) using percentage African ancestry rather than observed race, would improve warfarin dosing algorithms in African Americans. Using BioVU, the Vanderbilt University Medical Center biobank linked to electronic medical records, we compared 25 modeling strategies to existing algorithms using a cohort of 2,181 warfarin users (1,928 whites, 253 blacks). We found that approaches incorporating additional variants increased model accuracy, but not in clinically significant ways. Race stratification increased model fidelity for African Americans, but the improvement was small and not likely to be clinically significant. Use of percent African ancestry improved model fit in the context of race misclassification.

1. Introduction

Warfarin is a commonly used anticoagulant with a narrow therapeutic index and high rate of significant adverse reactions from both over- and under-dosing.[1] A number of pharmacogenomic variants are associated with stable warfarin dose,[2] and many studies have developed dosing algorithms using these variants.[1,3] Genotype-guided dosing is part of the United States Food and Drug Association (FDA) product label for warfarin.

The two largest randomized controlled trials of pharmacogenomic-guided warfarin dosing, EU-PACT[4] and COAG[5], yielded discordant findings on the clinical utility of incorporating pharmacogenomics into current dosing strategies. The EU-PACT study showed significantly increased percent time in therapeutic range (PTTR) over 12 weeks for the pharmacogenomic group while the COAG trial did not see a significant difference in PTTR over a 4-week time period. One of the reasons highlighted for these inconsistent findings across trials was the higher frequency of African descent individuals in COAG (27%) compared to EU-PACT (0.9%).[6] In COAG, African Americans with genotype-guided dosing spent an average of 8% less time in therapeutic range than the clinical algorithm group. Studies have shown that the *CYP2C9**2/*3 variants used by both COAG and EU-PACT are less frequent among those of African descent.[7] There are also variants important for dosing among individuals of African descent alleles that were unaccounted for in these trials.[7–11] Drozda found that failing to take into account these expanded variants resulted in significantly worse dose predictions among African Americans.[12] Additionally, Limdi found that using a race stratified dosing approach resulted in significantly more dose variation explained in both whites and blacks compared to a race-combined dosing model.[13]

Although much work has been conducted in this area, there remain outstanding questions that need to be answered. For example, because the algorithm proposed by Drozda was developed only in African Americans, its generalizability to individuals of European descent is unknown. Additionally, clinical dosing algorithms using a stratified approach, as advocated by Limdi have not been robustly tested to determine clinical validity. Further, in other clinical predictive models, using percent African ancestry as a more nuanced and biologically accurate measure of race provided better predictive performance than categorical race.[14] This study seeks to expand on previous warfarin dosing algorithm development efforts within Vanderbilt's EMR-linked biobank[15] to account for new variants associated with warfarin dose in African Americans. Additionally, we investigate whether race-stratified models or models using percent African ancestry result in clinically significant improvements (≥0.5-1mg/day) in dose prediction accuracy.

2. Methods

2.1. *Study Population*

Using BioVU, the Vanderbilt University biobank linked to electronic medical records (EMR), we selected all adult patients (≥18 years old) with DNA available who also had warfarin mentioned in the active prescription section of their problem list or a note from one of the hospital's anticoagulation clinics as of July of 2015. We used two approaches to extract stable warfarin dose based on whether the patient's warfarin was managed by an anticoagulation clinic or an individual physician.

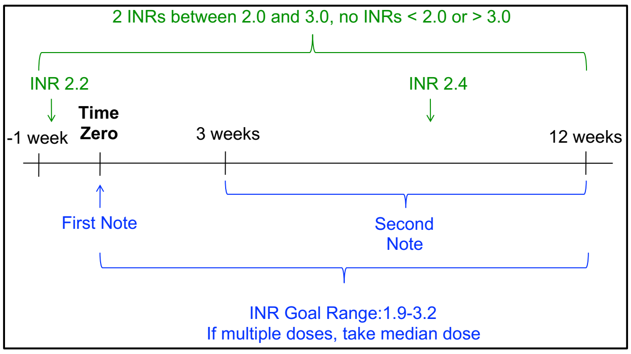


Figure 1. Stable Warfarin Dose Window Algorithm

We used a previously published and validated algorithm[15] to extract stable warfarin dose from patients with their dose managed by a Vanderbilt anticoagulation clinic or, for a subset of African Americans, where the dose was managed by their primary care provider. This approach identifies stable warfarin dose windows, as summarized in **Figure 1**. A stable dose window is defined as the presence of two or more notes from the anticoagulation clinic (or problem list entries for those managed by a primary-care provider) at least three, but not more than 12, weeks apart. During this time (from 7 days before the first note through the second note) the patient must also have two or more International Normalized Ratio (INR) measurements at least one day apart and all INR measurements in the window must be between 2 and 3. For anticoagulation clinic patients the INR goal range at the time of the stable dose window was required to be between 1.9-3.2. Primary-care managed patients were assumed to have an INR goal range of 2-3 unless otherwise specified (where ranges outside of 1.9-3.2 resulted in exclusion from the study). Warfarin dose was extracted from every anticoagulation clinic note in the window using regular expressions. The first window with identical prescribed warfarin doses throughout the window was selected as the stable warfarin dose. Patients lacking a window with identical warfarin doses throughout the window were manually reviewed to confirm accurate dose extraction. If multiple doses were prescribed during the window, the median dose was used. All primary care managed patient records were manually reviewed to extract warfarin dose and verify INR goal range because problem lists are susceptible to copy/paste redundancies and computational extraction may be invalid.

Clinical covariates influencing stable warfarin dose were extracted with a variety of methods. Concomitant therapies (amiodarone, carbamazepine, phenytoin, and rifampin) listed in the problem list before or during the dose window were manually reviewed to confirm the prescriptions were active during the window. Smoking status was identified combination of natural language processing (NLP) and tobacco use International Classification of Disease version 9 (ICD9) codes,[16,17] followed by manual review to confirm active smoking at the time of the stable dose window. Body surface area,[18] was calculated using the median height and weight across the stable dose window or the closest height and weight measurement available within 3-6 months before or after the window (extracted via manual review). Age was defined as the age at the first anticoagulation clinic note or problem list warfarin entry in the stable dose window. "EMR recorded race" is defined by the care provider, but has shown concordance with genetic ancestry.[19] Indication for warfarin treatment, blood clots (i.e., deep venous thrombosis [DVT] or pulmonary embolism[PE]) or atrial fibrillation, was determined through ICD9 codes.[20,21]

2.2. *Genotyping*

This study genotyped twenty-one single nucleotide polymorphisms (SNPs) that had ever been associated with warfarin dose in European or African-descent populations and recorded in the

Table 1. Overview of Dose Prediction Models Tested

	Genetic Model				
	Clinical Only	**Limited Genetic**	**Expanded Genetic**	**Combined SNP**	**Haplotype**
Clinical Vars. (All)	Age (in decades); Body surface area; Smoking status; Amiodarone; Enzyme inducer				
Race Adj. (one of:)	1) Unadjusted 2) EMR Race 3) %African Ancestry 4) White Only 5) Black Only				
Genetic Variables	-	*VKORC1*-1639 *CYP2C9**2 *CYP2C9**3	*VKORC1*-1639 *CYP2C9**2 *CYP2C9**3 *CYP2C9**5 *CYP2C9**6 *CYP2C9**8 *CYP2C9**11 rs2359612 rs2884737 rs7200749 rs8050894 rs9934438 rs17886199 rs10871454 rs2108622 rs11676382 rs12714145 rs339097 rs12777823	*VKORC1*-1639 *CYP2C9**2 *CYP2C9**3 rs7200749 rs9934438 rs17886199 rs10871454 rs2108622 rs11676382 rs12714145 rs339097 rs12777823 *VKORC1* Other[1] *CYP2C9* Other[2]	*VKORC1*-1639 rs2359612 rs2884737 rs7200749 rs8050894 rs9934438 rs17886199 rs10871454 rs2108622 rs11676382 rs12714145 rs339097 rs12777823 *CYP2C9**1/*2 *CYP2C9**1/*3 *CYP2C9**2/*2 *CYP2C9**2/*3 *CYP2C9**3/*3 *CYP2C9* Other Het.[3] *CYP2C9* Other Hom.[4]

[1] If individual carries one or more minor allele at rs2359612 or rs2884737 or rs61162043 or rs8050894 then called 1, else 0; [2] If individual carries one or more minor allele at CYP2C9 *5/*6/*8 or *11 then call 1, else 0; [3] CYP2C9 *1/*11, *1/*5, *1/*6, *1/*8); [4] CYP2C9 *3/*8, *5/*8, *5/*11, *8/*8, *8/*11

Pharmacogenomics Knowledge Base (PharmGKB, www.pharmgkb.org).[22] Three variants (rs9923231, rs1799853, rs1057910) were genotyped using a Taqman assay by the Vanderbilt Technologies for Advanced Genomics (VANTAGE) core. A subset of white subjects had previous genotyping for these variants on the Illumina ADME assay and were not included in the Taqman assay. The remaining 17 variants were genotyped across the entire study population with a Sequenom assay performed by the VANTAGE core. Genotyping data were checked for marker efficiency and samples removed if they were missing one or more genotype calls for the tested variants. Duplicates and HapMap controls were validated.

We used existing genotyping data to calculate percent African ancestry across a subset of the population. Individuals were genotyped on one or more of the following platforms: Illumina Exome Beadchip, Human Omni Express Exome v2, Metabochip and/or OmniQuad. For each platform independently, samples with discrepant genders or sample efficiency <99% were removed. Markers with genotyping efficiencies < 99% or minor allele frequencies<5% were dropped. For the Exome chip, thresholds were set to 97% and 98% for genotyping and sample efficiency respectively as has been done previously to account for low frequency variants.[23] Within each platform, percent African ancestry was calculated using the ADMIXTURE supervised learning method with HapMap Phase III CEU and YRI reference populations.[24] The median estimate was used for individuals genotyped on multiple platforms.

2.3. *Analysis*

We fit and tested 25 different dosing models, combing 5 genetic modeling strategies (including exclusion of genetics altogether) with 5 different methods of race/ancestry adjustment. A summary of the 25 models tested are presented in **Table 1**. For race-stratified models, variants that were monomorphic or non-varying clinical factors were not included. To validate model summaries and prevent overfitting, we bootstrapped 1000 samples with replacement, trained a generalized linear model on each bootstrap, and tested the original dataset against each model. We calculated the mean absolute error (in mg/week) and R^2 for each bootstrap model, then calculated the median and an empiric confidence interval using the 2.5th and 97.5th percentiles of the bootstrap summaries. For all combined race models, we calculated these evaluation criteria across the entire test population and then within each EMR recorded race group separately. Because there are different risks for over- and under-dosing, we also calculated these summary evaluation criteria stratified by low (<21mg/week), medium (21-49mg/week), and high (>49mg/week) stable dose across the entire test

Table 2. Summary of Previous Algorithms Tested for Warfarin Dosing

Algorithm	Clinical Predictors	Genetic Predictors	Notes
Fixed 35mg Weekly Dose	-	-	-
FDA Dosing Table[1]	-	VKORC1-1639 CYP2C9*2 CYP2C9*3	Used mean of dosing range given.
IWPC (International Warfarin Pharmacogenetics Consortium)[2]	Age (in decades) Height Weight Asian African American Amiodarone Enzyme Inducers	VKORC1-1639 CYP2C9*2 CYP2C9*3	-
Ramirez et. al.[3]	Age (in years) Race Sex Body Surface Area Smoking Status DVT/PE Atrial Fibrillation	VKORC1-1639 CYP2C9*2 CYP2C9*3 CYP2C9*6 CYP2C9*8 rs2108622 rs339097	-
Hernandez et. al.[4]	Age (in years) Weight DVT/PE	VKORC1-1639 VKORC1, rs61162043 CYP2C9*2 CYP2C9*3 CYP2C9*5 CYP2C9*8 CYP2C9*11 rs7089580 rs12777823	Performed on subset of population with genotyping for rs61162043. Missing CYP2C9 rs7089580 due to probe failure. Set all patients to reference allele

[1] www.accessdata.fda.gov/drugsatfda_docs/ label/2010/009218s108lbl.pdf; [2] Klein *et. al.* NEJM. 2009; [3] Ramirez *et. al.* Future Medicine. 2010; [4] Hernandez *et. al.* The Pharmacogenomics Journal. 2014.

Table 3. Population Demographics

	Combined (n = 2181)	**Whites (n = 1928)**	**Blacks (n = 253)**
Weekly Warfarin Dose, mg/wk (median, sd)	35.0 (±17.6)	35.0 (±17.0)	40.8 (± 19.9)
Age, years (mean, sd)	66 (± 15)	66 (± 15)	60 (± 16)
Female (n, %)	911 (41.8%)	784 (40.7%)	127 (50.2%)
African American (n, %)	253 (11.6%)	-	-
% African Ancestry (median, sd)[1]	0.99 (± 31)	0.65 (± 4.5)	81.6 (± 10.3)
Height, cm (median, sd)	173 (± 13.5)	174 (±13.0)	170 (±16.1)
Weight, kg (median, sd)	89 (± 24.0)	88 (± 23.9)	91 (±24.7)
Body Surface Area, m^2 (median, sd)	2.0 (± 0.29)	2.0 (± 0.29)	2.0 (± 0.30)
Current Smokers (n, %)	209 (9.6%)	168 (8.7%)	41 (16.2%)
Amiodarone (n, %)	229 (10.5%)	202 (10.5%)	27 (10.7%)
Enzyme Inducers (n, %)	20 (0.92%)	15 (0.78%)	5 (1.98%)
Indication			
VTE (n, %)	414 (19.0%)	337 (17.5%)	77 (30.4%)
Atrial Fibrillation (n, %)	1592 (73%)	1443 (75%)	149 (59%)

[1] %-African ancestry available for 987 individuals (808 whites, 179 blacks)

population and then within each EMR recorded race separately. To evaluate the validity of our models and compare to existing algorithms, we also calculated mean absolute error and R^2 for a number of existing algorithms. The algorithms tested are summarized in **Table 2**.

3. Results

A total of 3,498 patients (3188 whites, 310 blacks) had a stable dose window (all features in **Figure 1**, except INR goal range filtering) and were genotyped on the Sequenom platform. Of these, 596 whites had *VKORC1*-1369 and *CYP2C9**2/*3 genotypes from the ADME platform, all other individuals were genotyped via Taqman. 291 individuals were missing one or more genotypes (with exceptions of rs7089580 and rs61162043 due to poor probe performance described below) and were removed from the analysis. Of the remaining 3,207 individuals 2,419 (2,192 whites and 227 blacks) had warfarin managed by the anticoagulation clinic. Filtering this population for INR goal ranges between 1.9-3.2 removed a further 233 individuals (212 whites, 21 blacks). Manual review to confirm stable warfarin dose, height and/or weight was performed for 203 whites and 28 blacks. This review removed 52 whites and 9 blacks for missing warfarin dose, height and/or weight. A total of 56 black individuals had warfarin managed by their primary care provider and were manually reviewed to extract warfarin dose and INR goal range. Combining the anticoagulation clinic and primary care populations yielded a final population of 2,181 individuals (1,928 whites, and 253 blacks).

Population demographics are presented in **Table 3**. Blacks had higher warfarin doses (40.8 vs 35mg/week), were younger (60 vs 66 years), were more likely to be current smokers (16% vs 8%), were more likely to be on anticoagulants due to thromboembolic events (30.4% vs 17.5%), and less likely to be on anticoagulants due to atrial fibrillation (59% vs. 75%) than whites. All other demographics factors were similar between blacks and whites.

One marker, rs7089580, failed genotyping in the Sequenom pool. Genotyping efficiency rates and minor allele frequencies are presented for the remaining 20 variants in **Table 4**. One variant, rs61162043 had lower genotyping efficiency (failed genotyping in 111 whites and 21 blacks) and was excluded from the expanded variants model. However, this variant was included in the *VKORC1* combined variable for the Combined Variant model. A summary of

Table 4. Genotyping Quality Control and Minor Allele Frequencies

Gene	SNP	Minor Allele	Call Rate[a]	Minor Allele Frequency (%)		
				Combined (n=2181)	Whites (n=1928)	Blacks (n=253)
VKORC1	rs9923231	T	99.79[b]	35.1	38.3	10.5
VKORC1	rs2359612	A	100	36.4	38.5	20.8
VKORC1	rs2884737	C	99.96	23.3	25.3	3.8
VKORC1	rs61162043	A	93.82	37.2	35.8	49.6
VKORC1	rs7200749	A	99.96	2.6	0.3	20.2
VKORC1	rs8050894	G	99.92	37.3	38.8	26.5
VKORC1	rs9934438	A	99.96	35.1	38.4	10.5
VKORC1	rs17886199	G	100	0.5	0	4.2
STX4	rs10871454	T	99.96	35.3	38.5	10.9
*CYP2C9*2*	rs1799853	T	99.79[b]	13.5	14.9	2.4
*CYP2C9*3*	rs1057910	C	99.95[b]	6.1	6.7	2.0
*CYP2C9*5*	rs28371686	G	100	0.2	0.05	1.6
*CYP2C9*6*	rs9332131	del	99.79	0.2	0	1.4
*CYP2C9*8*	rs7900194	A	100	0.9	0.03	7.3
*CYP2C9*11*	rs28371685	T	99.96	0.4	0.3	1.4
CYP4F2	rs2108622	T	100	28.2	30.4	11.3
GGCX	rs11676382	G	99.96	8.8	9.7	2.6
GGCX	rs12714145	T	100	42.1	41.6	45.8
CALU	rs339097	G	99.83	1.6	0.2	12.3
CYP2C-cluster	rs12777823	A	99.92	16.1	14.6	28.1

[a]Call rates for completed genotyped population – not the final study population (as valid genotypes required for all but rs61162043); [b]Call rate for Taqman group only. ADME QC according to typical procedures.

the frequency of observed diplotypes for *CYP2C9* is presented **Table 5**. The majority of both racial/ethnic populations had a *1/*1 diplotype. Homozygotes and compound heterozygotes for the *2 and *3 variants (i.e., *2/*2, *3/*3, and *2/*3) were only observed in whites. Homozygotes and compound heterozygotes of the less common *5, *6, *8, and *11 alleles were only observed in blacks.

Within our final study population, 978 individuals (800 whites and 178 blacks) had genome-wide data available. A total of 764 individuals were genotyped on two platforms, 98 had genotyped data from three platforms, and 5 individuals had genotyping on four platforms. Of these individuals, the majority (n=437) had a maximum difference of less than 1% between estimates across platforms. Only 7 individuals had estimates across platforms that differed by more than 5% (maximum range of 9.8%). Three individuals had an EMR-recorded race of white, but had more than 50% African ancestry. The median ancestry estimate was used for all analyses.

A summary of the mean absolute error and percent variation explained (R^2) for all twenty-five fitted models, as well as the performance of existing dosing algorithms are provided in **Table 6**. Comparing all new and existing algorithms, the Expanded Genetic unadjusted, Expanded Genetic

Table 5. CYP2C9 Diplotype Frequencies

CYP2C9 Haplotype	Combined (n = 2181)	Whites (n = 1928)	Blacks (n = 253)
*1/*1	1402 (64.3%)	1222 (63.4%)	180 (71.2%)
*1/*2	357 (16.4%)	345 (18%)	12 (4.8%)
*1/*3	214 (9.8%)	205 (10.6%)	9 (3.6%)
*1/*5	8 (0.4%)	2 (0.1%)	6 (2.4%)
*1/*6	7 (0.3%)	-	7 (2.8%)
*1/*8	28 (1.3%)	1 (0.1%)	27 (10.7%)
*1/*11	15 (0.7%)	11 (0.6%)	4 (1.6%)
*2/*2	100 (4.6%)	100 (5.2%)	-
*2/*3	31 (1.4%)	31 (1.6%)	-
*3/*3	11 (0.5%)	11 (0.6%)	-
*3/*8	1 (<0.1%)	-	1 (0.4%)
*5/*8	1 (<0.1%)	-	1 (0.4%)
*5/*11	1 (<0.1%)	-	1 (0.4%)
*8/*8	3 (0.1%)	-	3 (1.2%)
*8/*11	2 (0.1%)	-	2 (0.8%)

EMR recorded race adjusted, Haplotype unadjusted, and Haplotype EMR recorded race adjusted models had the lowest mean absolute error across the combined population, with the Haplotype models explaining slightly more dose variance (54.4% vs 54.1%). The Expanded Variant model with percent ancestry adjustment had the lowest mean absolute errors in whites, and the Expanded Genetic stratified model had the lowest mean absolute error in blacks.

The algorithm performance with respect to mean error within low, medium, and high weekly dose groups are presented in **Figure 2**. When broken down by dose range 362 individuals (336 white and 26 black) had low warfarin requirements (<21mg/week), 1,313 individuals (1,173 whites and 140 blacks) had moderate warfarin requirements (21-49mg/week), and 486 individuals (402 whites and 84 blacks) had high warfarin requirements (>49mg/week). Within the medium dose requirement group (60% of the study population), dose predictions in whites were less than 5mg/week overestimated, while dose predictions in blacks were ~5mg/week overestimated. For the 17% of individuals with low warfarin dose requirements, mean dosing error was <10mg/week overestimated in whites, and 10-20mg/week overestimated in African Americans. The existing algorithm with the best performance among low-dose requiring African Americans was Ramirez *et. al.* (overestimating warfarin dose by 11.6mg/week). Within the high dose requirement individuals (22%), all races were consistently underestimated by 10-20mg/week.

Table 6. Performance of Predictive Dosing Algorithms

Algorithm	Mean Absolute Error (mg/week) Median (95% Confidence Interval)			Percent Variation Explained (R^2) Median (95% Confidence Interval)		
	Combined	Whites	Blacks	Combined	Whites	Blacks
			Existing Algorithms			
Fixed 35 mg/week	13.5	13.2	16.1	-2.3	-1.1	-23.7
US FDA Table mid-range	12.0	11.7	14.7	17.5	18.3	1.1
IWPC	9.5	9.0	13.4	42.7	45.2	20
Ramirez *et. al.*	9.2	8.8	12.9	47.5	49.5	28.7
Hernandez *et. al.*	10.4	9.9	14.2	37.3	39.4	17.2
			New Models			
Clinical						
Unadjusted	12.0 (11.9-12.0)	11.7 (11.7-11.8)	14.0 (13.8-14.2)	20.3 (19.9-20.5)	19.6 (19.0-19.9)	12.3 (9.9-14.9)
Race Adjusted	11.9 (11.9-11.9)	11.7 (11.7-11.7)	13.6 (13.4-13.8)	21.5 (21.1-21.6)	19.8 (19.3-20.0)	20.8 (18.9-22.1)
% Ancestry Adjusted	11.5 (11.5-11.7)	10.9 (10.8-11.0)	14.5 (14.3-14.9)	23.8 (22.9-24.2)	20.6 (19.3-21.3)	21.7 (17.9-24.2)
Race Stratified	-	11.7 (11.7-11.7)	13.4 (13.3-13.7)	-	19.8 (19.4-20.0)	21.7 (18.0-23.1)
Limited Genetic						
Unadjusted	9.3 (9.2-9.3)	8.8 (8.8-8.8)	12.9 (12.8-13.1)	51.5 (51.2-51.6)	53.8 (53.4-54.1)	27.8 (25.4-29.7)
Race Adjusted	9.3 (9.2-9.3)	8.8 (8.8-8.8)	12.8 (12.7-13.0)	51.8 (51.5-52.0)	54.0 (53.6-54.2)	30.0 (27.8-31.2)
% Ancestry Adjusted	9.5 (9.4-9.5)	8.5 (8.4-8.6)	13.7 (13.5-14.0)	49.8 (49.0-50.2)	52.8 (51.5-53.4)	32.4 (28.9-34.7)
Race Stratified	-	8.8 (8.8-8.8)	12.7 (12.5-13)	-	54.0 (53.7-54.2)	30.9 (27.1-32.5)
Expanded Genetic						
Unadjusted	**9.0 (9.0-9.1)**	8.6 (8.6-8.7)	12.2 (11.9-12.6)	54.1 (53.6-54.4)	55.4 (54.9-55.7)	37.7 (34.0-40.0)
Race Adjusted	**9.0 (9.0-9.1)**	8.6 (8.6-8.7)	12.2 (11.9-12.6)	54.1 (53.5-54.4)	55.4 (54.9-55.8)	37.5 (33.5-40.1)
% Ancestry Adjusted	9.2 (9.1-9.3)	**8.4 (8.3-8.5)**	12.9 (12.5-13.4)	52.5 (50.8-53.3)	54.2 (52.5-55.1)	39.7 (32.9-43.8)
Race Stratified	-	8.6 (8.6-8.7)	**11.9 (11.6-12.4)**	-	55.7 (55.2-55.9)	39.5 (33.8-42.5)
Combined SNP						
Unadjusted	9.9 (9.9-10.0)	9.5 (9.5-9.6)	12.8 (12.6-13.1)	44.0 (43.5-44.3)	44.7 (44.2-45.0)	30.2 (27.2-32.5)
Race Adjusted	9.9 (9.9-10.0)	9.6 (9.5-9.6)	12.8 (12.6-13.1)	44.0 (43.5-44.3)	44.7 (44.2-45.1)	30.3 (27.0-32.6)
% Ancestry Adjusted	10 (9.9-10.1)	9.2 (9.1-9.3)	13.6 (13.3-14.0)	44.4 (43.3-45.0)	44.8 (43.2-45.8)	34.2 (29.5-37.9)
Race Stratified	-	9.6 (9.5-9.6)	12.5 (12.2-12.9)	-	44.9 (44.3-45.1)	33.9 (29.6-36.4)
Haplotype						
Unadjusted	**9.0 (9.0-9.1)**	8.6 (8.6-8.7)	12.1 (11.8-12.5)	**54.4 (54.0-54.7)**	**55.8 (55.3-56.1)**	38.0 (34.5-40.1)
Race Adjusted	**9.0 (9.0-9.1)**	8.6 (8.6-8.7)	12.1 (11.9-12.5)	**54.4 (53.9-54.7)**	**55.8 (55.3-56.1)**	37.8 (34.3-40.2)
% Ancestry Adjusted	9.2 (9.1-9.3)	**8.4 (8.3-8.5)**	12.7 (12.4-13.3)	53.1 (51.6-53.9)	54.6 (52.3-55.5)	**41.5 (36.1-44.5)**
Race Stratified	-	8.6 (8.5-8.6)	12.0 (11.7-12.5)	-	56.2 (55.7-56.4)	39.6 (34.6-42.1)

Bold and shaded cells indicate the best performing algorithm for each population.

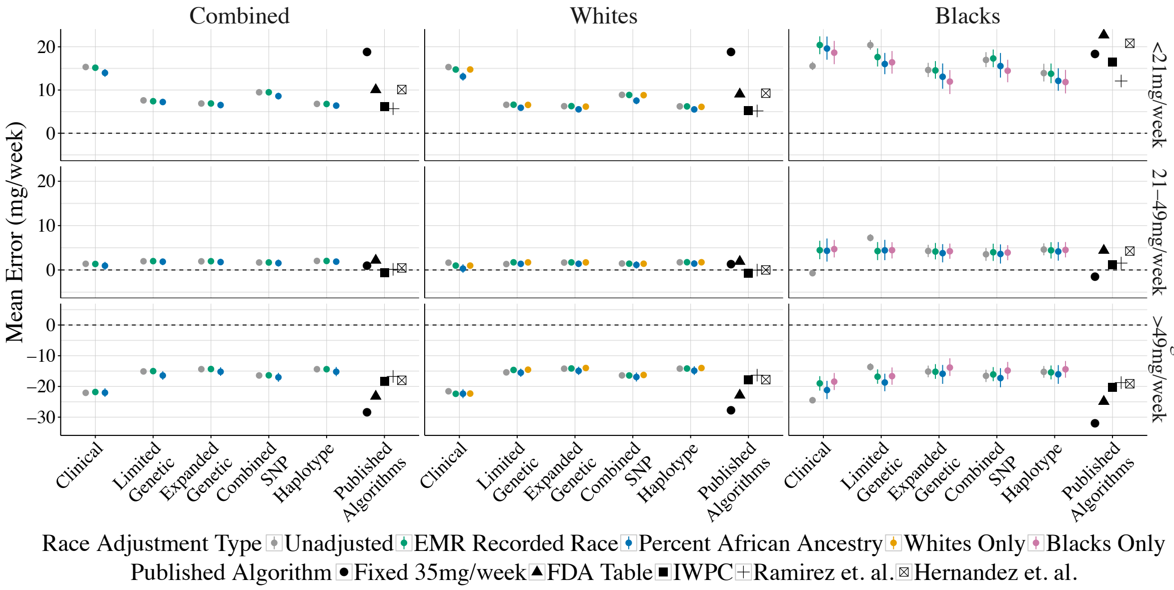


Figure 2. Performance of Dosing Algorithms by Stable Dose Range
This figure shows the algorithm performance (mean error in mg/week) divided by EHR recorded race and the stable dose range, e.g. patient’s stable warfarin dose is a low weekly dose (<21mg/week), medium weekly dose (21-49mg/week), or high weekly dose (>49mg/week). Mean errors greater than 0 indicate over dosing, while mean errors less than 0 indicate underdosing.

4. Discussion

The goal of this study was to: 1) account for variants associated with warfarin dose in African Americans, 2) investigate whether race-stratified dosing leads to clinically significant improved dose predictions, 3) investigate whether race adjustment using percent ancestry offers improved prediction accuracy compared to EMR recorded race. The last goal was predicated on a study of lung function predictions (a continuous trait that, like warfarin dose, differs by race) that found improved model fit when they included percent African ancestry.[14] This hypothesis was bolstered by a study among Caribbean Hispanics that found adjusting for admixture improved warfarin dose prediction.[25]

Although this study required that individuals have DNA available in our biobank, because we took a complete cross-section of all individuals with warfarin exposure and DNA, the relative percentage of African Americans in this study (~10%) is consistent the broader Vanderbilt clinical population. As previously observed in the literature,[13] our black study population had a higher incidence of DVT/PE as an indication for anticoagulation. The genetics of our population were consistent with expected allele frequencies from the HapMap populations, with African Americans having allele frequencies lying between the Yoruba in Ibadan, Nigeria (YRI) and African Americans in the Southwest USA (ASW). Ancestry estimates for the black population were as expected with African Americans having approximately 80% African ancestry,[26] and allele frequencies for *CYP2C9**2/*3 and *VKORC1*-1639 were consistent with other studies within the Vanderbilt clinical population (that are not necessarily part of the biobank population).[27] Importantly, *CYP2C9* *2 and *3 homozygotes and compound heterozygotes were only observed in our white population, lending support to the notion that use of only *CYP2C9**2/*3 for warfarin dosing algorithms may be insufficient for African Americans.

Examining algorithm performance over the entire study population, the inclusion of additional variants associated with warfarin dose did increase dosing accuracy (mean absolute error) and percentage of dose variation explained for the combined, white and black populations. In all three populations one of the novel algorithms using SNPs independently (Expanded Genetic) or combined by *CYP2C9* diplotype (Haplotype) outperformed existing algorithms, the Clinical, and the Limited Genetic models. When considering confidence intervals, the Expanded Genetic and Haplotype models performed at similar levels across all populations. This is important for future clinical implementation as algorithms such as MyDrugGenome use *CYP2C9* diplotype. These diplotypes do not always have unambiguous assignments and are subject to change as the number of known genetic variants in a gene rise.[28] Our results suggest that algorithms utilizing unique SNPs can perform at similar levels to those using diplotypes and may be preferable due their more stable identification.

When considering only mean absolute error, stratified dosing models outperformed combined models only in African Americans. Interesting, stratified dosing did not result in improved performance over combined models in whites. This may be due to race misclassification of the three individuals recorded as white in the EMR, but who nevertheless had greater than 50% African ancestry. We chose not to manually change these individuals' race, as this misclassification is a real, generalizable[14] problem in the clinic, and would have an effect on algorithms' accuracy if clinically deployed. Although stratified dosing did improve algorithm performance among African Americans, it did not increase percent of warfarin dose explained by the model as has been seen in other studies.[13]

Correcting for race with percent ancestry yielded interesting results. Within the clinical model, percent ancestry improved model fit (lower mean absolute error, higher R^2) in the combined population, but not when pharmacogenomic markers were added into the model. Interestingly, percent ancestry improved dosing among whites across all models including those with pharmacogenomic markers. It is possible that the race misclassification also affected the algorithms using percent African ancestry. While this misclassification would be an important limitation in clinical implementation, at the current time this is less important because genetic ancestry is typically unavailable in current clinical systems. However, should this information have increased clinical utility in the future, panel testing of ancestry informative markers could enable implementation of these data.

While the algorithms developed in this study outperformed existing algorithms when considering the mean absolute error of prediction, we advocate using **Figure 2** to evaluate algorithm performance for desired implementation. We also caution that to determine the overall "best" algorithm, one must think within the context of clinical implementation of these algorithms. "Best" needs to be defined not just by performance, but also the generalizability and feasibility of implementation. For example, the Ramirez *et. al.* algorithm outperforms all algorithms for blacks with low warfarin doses and performs similarly to the best algorithms across most other race/dose requirement groups. However, the Ramirez *et. al.* algorithm requires the reason for anticoagulation (DVT/PE or atrial fibrillation), information typically computationally unavailable at the time of warfarin initiation. Many settings implementing prospective pharmacogenomic testing rely on automated clinical decision support and active intervention at the time of ordering to tailor the prescription. Although our overall best performing algorithm/s are not clinically significantly improved over the

Ramirez *et. al.* algorithm, they can all be computed with information readily available in a patient's medical record, allowing for immediate calculation of starting warfarin dose at the time of prescription.

In addition to the question of implementation one must also consider that the clinical impact of dose misclassification is not consistent across all dosing groups. Overdosing individuals with low warfarin requirements (warfarin dose <21mg/week) can lead to serious bleeding events, while under-dosing those with high warfarin requirements (doses >49mg/week) can lead to clotting events.[29,30] Although the IWPC algorithm performs similarly to the highest performance algorithms, it is particularly poor at predicting doses of low dose African Americans (~4.5 mg worse than the best performing algorithms). Depending on the frequency of low dose African Americans in the health system (determined with retrospective data), the IWPC algorithm may not be the best option. However, if the health system had a significant Asian population, use of the IWPC algorithm may be preferred because it takes these variables into account even if performance among low dose African Americans is reduced.

An important limitation of this study is that one of the previously tested algorithms, Ramirez *et. al.* was derived on a subset of patients included in this study. Thus it is possible that the high performance of the Ramirez algorithm in our population is inflated and may not be generalizable. The novel algorithms were also likely positively biased given the lack of an external validation set. Further, the results of the Hernandez *et. al.* algorithm were likely negatively biased as two SNPs predicting higher dose in African Americans were not included in this study due to poor genotyping quality. This study was also limited by the small number of African Americans studied. Additionally, since these data are from a single institution the results may not generalize to other populations. Warfarin dose is highly affected by vitamin K intake and the eating habits/cultural norms in the South may not reflect other parts of the US and world. Similarly, since this study only included whites and blacks, it is not clear how well the derived algorithms will perform among other ancestry groups.

In conclusion, expanding the variants in a warfarin dosing model does increase model accuracy, but not in clinically significant ways over existing algorithms in the literature. Similarly, race stratification resulted in the best model fits for African Americans, but the difference is unlikely to be clinically significant. Finally, percent ancestry surprisingly improves model fit – especially in the context of race misspecification in EMR recorded white race. However, the improvement in model fit among the white population is not clinically significant. When determining which dosing model to use, care must be given to selecting a model that not only matches the racial distribution of the population, but is also technically and financially achievable.

Acknowledgements

Thanks to Cara Sutcliffe and Paxton Baker of the VANTAGE core and Sarah Collier of the BioVU team for their help facilitating the genotyping in this project. This project was funded by Vanderbilt University, NCI (K07CA172294), NHLBI (U01HL122904, U19HL065962), NHGRI (U01HG007253, U01HG008672), and NLM (T15LM007450). The data used for this analysis were obtained from BioVU, supported by institutional funding and funding from NCATS (UL1TR000445).

References

1. B.F. Gage, C. Eby, J.A. Johnson, et al., *Clin Pharmacol Ther* **84**, 326 (2008).
2. J.A. Johnson and L.H. Cavallari, *Trends Cardiovasc Med* **25**, 33 (2015).
3. T.E. Klein, R.B. Altman, N. Eriksson, et al., *N Engl J Med* **360**, 753 (2009).
4. S.E. Kimmel, B. French, S.E. Kasner, et al., *N Engl J Med* **369**, 2283 (2013).
5. M. Pirmohamed, G. Burnside, N. Eriksson, et al., *N Engl J Med* **369**, 2294 (2013).
6. S.A. Scott and S.A. Lubitz, *Pharmacogenomics* **15**, 719 (2014).
7. M.A. Perera, L.H. Cavallari, N.A. Limdi, et al., *Lancet* **382**, 790 (2013).
8. M.A. Perera, E. Gamazon, L.H. Cavallari, et al., *Clin. Pharmacol. Ther.* **89**, 408 (2011).
9. Y. Liu, H. Jeong, H. Takahashi, et al., *Clin Pharmacol Ther* **91**, 660 (2012).
10. L.H. Cavallari, M. Perera, M. Wadelius, et al., *Pharmacogenet. Genomics* **22**, 152 (2012).
11. R. Daneshjou, E.R. Gamazon, B. Burkley, et al., *Blood* **124**, 2298 (2014).
12. K. Drozda, S. Wong, S.R. Patel, et al., *Pharmacogenet Genomics* **25**, 73 (2015).
13. N.A. Limdi, T.M. Brown, Q. Yan, et al., *Blood* **126**, 539 (2015).
14. R. Kumar, M.A. Seibold, M.C. Aldrich, et al., *N Engl J Med* **363**, 321 (2010).
15. A.H. Ramirez, Y. Shi, J.S. Schildcrout, et al., *Pharmacogenomics* **13**, 407 (2012).
16. M. Liu, A. Shah, N.B. Peterson, et al., *AMIA Annu Symp Proc* (2012).
17. L.K. Wiley, A. Shah, H. Xu, et al., *J Am Med Inf. Assoc* (2013).
18. D. Du Bois and E.F. Du Bois, *Nutrition* **5**, 303 (1989).
19. L. Dumitrescu, M.D. Ritchie, K. Brown-Gentry, et al., *Genet. Med.* **12**, 648 (2010).
20. E.R. McPeek Hinz, L. Bastarache, and J.C. Denny, *AMIA Annu Symp Proc* **2013**, 975 (2013).
21. S. Khurshid, J. Keaney, P.T. Ellinor, et al., *Am J Cardiol* **117**, 221 (2016).
22. M. Whirl-Carrillo, E.M. McDonagh, J.M. Hebert, et al., *Clin. Pharmacol. Ther.* **92**, 414 (2012).
23. Y. Guo, J. He, S. Zhao, et al., *Nat Protoc* **9**, 2643 (2014).
24. D.H. Alexander, J. Novembre, and K. Lange, *Genome Res* **19**, 1655 (2009).
25. J. Duconge, A.S. Ramos, K. Claudio-campos, et al., *PLoS One* **11**, (2016).
26. F. Zakharia, A. Basu, D. Absher, et al., *Genome Biol* **10**, R141 (2009).
27. S.L. Van Driest, Y. Shi, E.A. Bowton, et al., *Clin Pharmacol Ther* **95**, 423 (2014).
28. J.D. Robarge, L. Li, Z. Desta, et al., *Clin. Pharmacol. Ther.* **82**, 244 (2007).
29. E.M. Hylek, A.S. Go, Y. Chang, et al., *N Engl J Med* **349**, 1019 (2003).
30. E.M. Hylek, C. Evans-Molina, C. Shea, et al., *Circulation* **115**, 2689 (2007).