IMAGING GENOMICS

LI SHEN

Center for Neuroimaging, Department of Radiology and Imaging Sciences
Center for Computational Biology and Bioinformatics
Indiana University School of Medicine
355 West 16th Street Suite 4100, Indianapolis, IN 46202
E-mail: shenli@iu.edu

LEE A.D. COOPER

Department of Biomedical Informatics, Emory University School of Medicine
Department of Biomedical Engineering, Georgia Institute of Technology
PAIS Building, 36 Eagle Row, 5th Floor South, Atlanta, GA 30322
E-mail: lee.cooper@emory.edu

Imaging genomics is an emerging research field, where integrative analysis of imaging and omics data is performed to provide new insights into the phenotypic characteristics and genetic mechanisms of normal and/or disordered biological structures and functions, and to impact the development of new diagnostic, therapeutic and preventive approaches. The Imaging Genomics Session at PSB 2017 aims to encourage discussion on fundamental concepts, new methods and innovative applications in this young and rapidly evolving field.

1. Introduction

Imaging genomics[1–9] is an emerging research field that arises with the recent advances in acquiring high throughput omics data and multimodal imaging data. Its major task is to perform integrative analysis of genomics data and structural, functional and molecular imaging data. Bridging imaging and genomic factors and exploring their connections have the potential to provide important new insights into the phenotypic characteristics and genetic mechanisms of normal and/or disordered biological structures and functions, which in turn will impact the development of new diagnostic, therapeutic and preventive approaches.

Binformatics strategies for imaging genomics, which is a relatively young field,[1–4] have been rapidly evolving. Early studies started with the simplest strategy to examine pairwise univariate associations[10,11] between genetic markers and imaging phenotypes. To identify more flexible associations involving multiple genetic markers and multiple imaging phenotypes, recent studies employed multiple regression and multivariate models,[12] sometimes coupled with powerful machine learning approaches[13] and valuable prior knowledge[14] to discover relevant imaging and genomic features. To increase statistical power and reduce false positives, meta-analysis studies[15,16] were performed to quantitatively synthesize imaging genomic findings from multiple independent analyses. To hunt for "missing heritability", epistatic studies[17] were performed to examine genetic interaction effects on imaging phenotypes. To identify biologically meaningful findings with increased statistical power, imaging genetic enrichment analysis[18] was proposed to mine set level associations in both imaging and genomic domains.

The topic of imaging genomics has recently been addressed in several medical imaging and bioinformatics conferences. The most focused one is the International Imaging Genetics

Conference (IIGC, http://www.imaginggenetics.uci.edu/), which is an annual meeting organized at the UC Irvine since 2005. The MICCAI Workshop on Imaging Genetics (MICGen, http://micgen.csail.mit.edu/) has been held twice in conjunction with the major medical image computing conference MICCAI in 2014 and 2015. An educational course on "Introduction to Imaging Genetics" has been offered at the annual meeting of the Organization for Human Brain Mapping (OHBM) since 2009. The topic of imaging genomics has also been covered in the following two events in the bioinformatics field: (1) ACM BCB 2015 Workshop on The Computational Pathology: Linking Tissue Phenotypes with Genomics and Clinical Outcomes, and (2) ICIBM 2015 Tutorial in Bioimage Informatics and Integrative Genomics.

As the field of imaging genomics contains a significant genomics (or omics in general) component in addition to biomedical imaging, we feel that it is timely for a major bioinformatics conference such as PSB to address this important, relevant and emerging topic. We believe that PSB offers an ideal and timely opportunity to bring together people with different expertise and shared interests in this rapidly evolving field. Specifically, the computational biology and bioinformatics expertise of the PSB and ISCB communities can provide important new perspective, complementary to the expertise of the IIGC, MICCAI, OHBM, ACM BCB and ICIBM communities, and thus can help contribute new concepts, methods, and applications to the analysis of emerging imaging and genomic data.

The scale and complexity of multidimensional imaging and omics data provide us unprecedented opportunities in enhancing mechanistic understanding of complex disorders such as neurological diseases[19–21] and cancers,[22,23] which can benefit public health outcomes by facilitating diagnostic and therapeutic progress. However, due to the extremely high dimensionality and complex structure of these data sets, this field is facing major computational and bioinformatics challenges. The technological advance in this field is urgently needed and has the potential to significantly contribute to multiple national health priority areas including *the Precision Medicine Initiative*,[24] *the Brain Initiative*,[25] and *the Big Data to Knowledge Initiative*.[26]

The objective of this Imaging Genomics Session at PSB 2017 is to encourage discussion on fundamental concepts, novel methods and innovative applications. We hope that this session will become a forum for researchers to exchange ideas, data, and software, in order to speed up the development of innovative technologies for hypothesis testing and data-driven discovery in Imaging Genomics.

2. Session Summary

This session includes an invited lecture and five accepted presentations with peer-reviewed papers. Three presentations will be delivered as platform talks and the other two as posters.

2.1. *Invited Talk*

Our invited lecture will be given by Dr. Paul Thompson, a world renowned pioneer in imaging genomics. Dr. Thompson is from the University of Southern California (USC). At USC, he is a Professor of Neurology, Psychiatry, Radiology, Pediatrics, Engineering, and Ophthalmology, the director of the USC Imaging Genetics Center, and the director of the ENIGMA

Center for Worldwide Medicine, Imaging & Genomics – an $11M NIH Center of Excellence in Big Data Computing. Dr. Thompson's major contributions to the field of imaging genomics and to the science in general can be summarized by the following text quoted from http://keck.usc.edu/faculty/paul-m-thompson/:

> Paul Thompson directs the ENIGMA Consortium, a global alliance of 307 scientists in 33 countries who conduct the largest studies of 10 major brain diseases – ranging from schizophrenia, depression, ADHD, bipolar illness and OCD, to HIV and addictions on the brain. ENIGMA's genomic screens of over 31,000 people's brain scans and genome-wide data (published in Nature Genetics, 2012; Nature, 2015) have brought together experts from 185 institutions to unearth genetic variants that affect brain structure, disease risk, and brain connectivity. Collaborating with imaging labs around the world, Dr. Thompson and his students have published over 1,300 publications (h-index: 116) describing novel mathematical and computational strategies for analyzing brain image databases, for detecting pathology in individual patients and groups, and for creating disease-specific atlases of the human brain.

2.2. *Papers*

In *Integrative analysis for lung adenocarcinoma predicts morphological features associated with genetic variations*, **Wang et al.** analyzed an imaging genomic data set downloaded from the TCGA portal, containing 201 patients with lung adenocarcinoma (LUAD). The data includes clinical information, mRNA expression profiles, and histopathologic whole slide images of the patients. On the imaging end, the authors calculated 283 morphological features from histopathologic images, and identified features strongly correlated with patient survival outcome. On the genomic end, the authors constructed the gene co-expression network and extracted gene co-expression clusters. To relate imaging with genomics, the authors regressed the outcome-relevant morphological feature on multiple co-expressed gene clusters using Lasso. The study identified gene clusters highly associated with DNA copy number variations. These observations may lead to new insight on lung cancer development, suggesting biological pathways from genetic variations, gene transcription, cancer morphology to survival outcome.

In *Identification of discriminative imaging proteomics associations in Alzheimer's disease via a novel sparse canonical correlation model*, **Yan et al.** analyzed an imaging proteomic data set downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Participants include 42 healthy controls, 67 patients with mild cognitive impairment (MCI), and 67 patients with Alzheimer's disease (AD). The data includes clinical information, magnetic resonance imaging (MRI) scans, and expression data of 229 proteomic analytes (83 from cerebrospinal fluid and 146 from plasma). The authors developed a novel machine learning model, called discriminative sparse canonical correlation analysis (DSCCA), and applied it to the joint analysis of imaging, proteomic and diagnostic data. This analysis yielded a strong imaging proteomic association so that the identified imaging and proteomic components had also high discriminative power. Such an outcome-relevant imaging proteomic pattern has the potential to improve mechanistic understanding of the disease.

In *Enforcing co-expression in multimodal regression framework*, **Zille et al.** analyzed an imaging genomic data set collected by Mind Clinical Imaging Consortium (MCIC). Participants include 116 controls and 92 schizophrenia patients. The data includes clinical information, functional MRI (fMRI) scans, and genotyping data. The authors developed a new machine learning model, called MT-CoReg, by combining sparse regression with canonical correlation analysis; and applied it to the analysis of the MCIC data. The analysis identified imaging and genomic markers that not only induce a strong imaging genomic association but also can jointly predict the outcome.

In *Adaptive testing of SNP-brain functional connectivity association via a modular network analysis*, **Gao et al.** analyzed an imaging genomic data set downloaded from the ADNI database. Participants include 162 ADNI subjects: 73 with no *APOE* E4 allele, 67 with one copy of the *APOE* E4 allele, and 22 with two copies of the *APOE* E4 allele. The authors analyzed the resting-state fMRI data to identify modular structures in brain functional networks, using a weighted gene co-expression network analysis (WGCNA) framework, coupled with topological overlap matrix (TOM) elements in hierarchical clustering. After that, they employed an adaptive association test based on the proportional odds model to identify distinct modular structures in brain functional networks in relation to different *APOE* E4 groups.

In *Exploring brain transcriptomic patterns: a topological analysis using spatial expression networks*, **Kuncheva et al.** analyzed whole genome whole brain gene expression data downloaded from the Allen Human Brain Atlas (AHBA). Participants include six AHBA donors. The authors focused on 16,906 genes selected based on a previous study, and 105 brain regions where at least one measurement in all 6 brains were available. A Spatial Expression Network (SEN) was extracted for each gene to quantify co-expression patterns amongst several anatomical locations. After that, network similarity measures were computed and used to quantify the topological resemblance between pairs of SENs and identify naturally occurring clusters. The analysis identified three stable clusters, including one with genes specifically involved in the nervous system, and the other two representing immunity, transcription and translation.

2.3. *Discussion*

Most of these studies were facilitated by and conducted using the Big Data resources available in the open science domain, including TCGA analyzed in (**Wang et al.**), ADNI analyzed in (**Yan et al.** & **Gao et al.**), and AHBA analyzed in (**Kuncheva et al.**). The imaging data investigated by these studies ranged from histological whole slide images of cancer specimens in (**Wang et al.**), structural MRI scans in (**Yan et al.**), functional MRI scans in (**Zille et al.** & **Gao et al.**), to images of mRNA expression levels across the brain in (**Kuncheva et al.**). The omics data examined in these studies were also diverse, including DNA genotyping data in (**Zille et al.** & **Gao et al.**), mRNA expression profiles in (**Wang et al.** & **Kuncheva et al.**), and proteomic expression profiles in (**Yan et al.**).

These studies were performed to better understand the brain transcriptomic patterns in healthy controls (**Kuncheva et al.**), the brain imaging genomic or imaging proteomic patterns in Alzheimer's disease (**Yan et al.** & **Gao et al.**) or schizophrenia (**Zille et al.**), and biological pathways from gene transcription, tissue morphology to survival outcome in lung cancer

(**Wang et al.**). As to the bioinformatics strategies, a variety of machine learning methods were employed or newly developed in these studies, including network analysis and clustering models used in (**Wang et al.**, **Gao et al.** & **Kuncheva et al.**), regression models used in (**Wang et al.**), an adaptive association test used in (**Gao et al.**), an integrative regression and canonical correlation analysis model used in (**Zille et al.**), and an outcome-regularized sparse canonical correlation analysis model used in (**Yan et al.**).

While **Gao et al.** studied functional brain network as an innovative imaging phenotype, **Kuncheva et al.** aimed to identify gene clusters using whole brain spatial expression networks. The remaining three studies (**Wang et al.**, **Yan et al.** & **Zille et al.**) shared a common theme to examine the relationship among three levels (i.e., omics features, imaging phenotypes, and clinical outcomes). This suggests a promising future direction to integrate imaging genomics with systems biology, which attempts to model complex and interactive multilevel biological systems using multimodal imaging and multidimensional omics data sets.

3. Acknowledgements

We would like to thank all the authors for their high-quality submissions and excellent presentations. We would like to thank all the reviewers for taking their time and effort to evaluate the papers and provide valuable feedbacks. We would like to thank Dr. Paul Thompson of University of Southern California for giving an outstanding invited lecture. We would like to thank Dr. Kun Huang of Ohio State University for sharing his valuable experience on how to organize a successful PSB session. We would like to thank the PSB 2017 chairs and Tiffany Murray of Stanford University for their great help and support.

References

1. A. R. Hariri and D. R. Weinberger. Imaging genomics. *Br Med Bull*, 65:259–70, 2003.
2. V. S. Mattay and T. E. Goldberg. Imaging genetic influences in human brain function. *Curr Opin Neurobiol*, 14(2):239–47, 2004.
3. A. R. Hariri, E. M. Drabant, and D. R. Weinberger. Imaging genetics: perspectives from studies of genetically driven variation in serotonin function and corticolimbic affective processing. *Biol Psychiatry*, 59(10):888–97, 2006.
4. D. C. Glahn, T. Paus, and P. M. Thompson. Imaging genomics: mapping the influence of genetics on brain structure and function. *Hum Brain Mapp*, 28(6):461–3, 2007.
5. P. M. Thompson, N. G. Martin, and M. J. Wright. Imaging genomics. *Curr Opin Neurol*, 23(4):368–73, 2010.
6. L. Shen, P. M. Thompson, S. G. Potkin, L. Bertram, L. A. Farrer, T. M. Foroud, R. C. Green, X. Hu, M. J. Huentelman, S. Kim, J. S. Kauwe, Q. Li, E. Liu, F. Macciardi, J. H. Moore, L. Munsie, K. Nho, V. K. Ramanan, S. L. Risacher, D. J. Stone, S. Swaminathan, A. W. Toga, M. W. Weiner, A. J. Saykin, and Alzheimer's Disease Neuroimaging Initiative. Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. *Brain Imaging Behav*, 8(2):183–207, 2014.
7. M. G. ElBanan, A. M. Amer, P. O. Zinn, and R. R. Colen. Imaging genomics of Glioblastoma: state of the art bridge between genomics and neuroradiology. *Neuroimaging Clin N Am*, 25(1):141–53, 2015.
8. W. B. Pope. Genomics of brain tumor imaging. *Neuroimaging Clin N Am*, 25(1):105–19, 2015.

9. A. J. Saykin, L. Shen, X. Yao, S. Kim, K. Nho, S. L. Risacher, V. K. Ramanan, T. M. Foroud, K. M. Faber, N. Sarwar, L. M. Munsie, X. Hu, H. D. Soares, S. G. Potkin, P. M. Thompson, J. S. Kauwe, R. Kaddurah-Daouk, R. C. Green, A. W. Toga, M. W. Weiner, and Alzheimer's Disease Neuroimaging Initiative. Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans. *Alzheimers Dement*, 11(7):792–814, 2015.
10. L. Shen, S. Kim, S. L. Risacher, K. Nho, S. Swaminathan, J. D. West, T. Foroud, N. Pankratz, J. H. Moore, C. D. Sloan, M. J. Huentelman, D. W. Craig, B. M. Dechairo, S. G. Potkin, Jr. Jack, C. R., M. W. Weiner, A. J. Saykin, and Alzheimer's Disease Neuroimaging Initiative. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *Neuroimage*, 53(3):1051–63, 2010.
11. J. L. Stein, X. Hua, S. Lee, A. J. Ho, A. D. Leow, A. W. Toga, A. J. Saykin, L. Shen, T. Foroud, N. Pankratz, M. J. Huentelman, D. W. Craig, J. D. Gerber, A. N. Allen, J. J. Corneveaux, B. M. Dechairo, S. G. Potkin, M. W. Weiner, P. Thompson, and Alzheimer's Disease Neuroimaging Initiative. Voxelwise genome-wide association study (vGWAS). *Neuroimage*, 53(3):1160–74, 2010.
12. D. P. Hibar, J. L. Stein, O. Kohannim, N. Jahanshad, A. J. Saykin, L. Shen, S. Kim, N. Pankratz, T. Foroud, M. J. Huentelman, S. G. Potkin, Jr. Jack, C. R., M. W. Weiner, A. W. Toga, P. M. Thompson, and Alzheimer's Disease Neuroimaging Initiative. Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. *Neuroimage*, 56(4):1875–91, 2011.
13. M. Vounou, E. Janousova, R. Wolz, J. L. Stein, P. M. Thompson, D. Rueckert, G. Montana, and Alzheimer's Disease Neuroimaging Initiative. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *Neuroimage*, 60(1):700–16, 2012.
14. J. Yan, L. Du, S. Kim, S. L. Risacher, H. Huang, J. H. Moore, A. J. Saykin, L. Shen, and Alzheimer's Disease Neuroimaging Initiative. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. *Bioinformatics*, 30(17):i564–71, 2014.
15. J. L. Stein, S. E. Medland, A. A. Vasquez, D. P. Hibar, R. E. Senstad, A. M. Winkler, R. Toro, K. Appel, R. Bartecek, O. Bergmann, M. Bernard, A. A. Brown, D. M. Cannon, M. M. Chakravarty, A. Christoforou, M. Domin, O. Grimm, M. Hollinshead, A. J. Holmes, G. Homuth, J. J. Hottenga, C. Langan, L. M. Lopez, N. K. Hansell, K. S. Hwang, S. Kim, G. Laje, P. H. Lee, X. Liu, E. Loth, A. Lourdusamy, M. Mattingsdal, S. Mohnke, S. M. Maniega, K. Nho, A. C. Nugent, C. O'Brien, M. Papmeyer, B. Putz, A. Ramasamy, J. Rasmussen, M. Rijpkema, S. L. Risacher, J. C. Roddey, E. J. Rose, M. Ryten, L. Shen, E. Sprooten, E. Strengman, A. Teumer, D. Trabzuni, J. Turner, K. van Eijk, T. G. van Erp, M. J. van Tol, K. Wittfeld, C. Wolf, S. Woudstra, A. Aleman, S. Alhusaini, L. Almasy, E. B. Binder, D. G. Brohawn, R. M. Cantor, M. A. Carless, A. Corvin, M. Czisch, J. E. Curran, G. Davies, M. A. de Almeida, N. Delanty, C. Depondt, R. Duggirala, T. D. Dyer, S. Erk, J. Fagerness, P. T. Fox, N. B. Freimer, M. Gill, H. H. Goring, D. J. Hagler, D. Hoehn, F. Holsboer, M. Hoogman, N. Hosten, N. Jahanshad, M. P. Johnson, D. Kasperaviciute, Jr. Kent, J. W., P. Kochunov, J. L. Lancaster, S. M. Lawrie, D. C. Liewald, R. Mandl, M. Matarin, M. Mattheisen, E. Meisenzahl, I. Melle, E. K. Moses, T. W. Muhleisen, et al. Identification of common variants associated with human hippocampal and intracranial volumes. *Nat Genet*, 44(5):552–61, 2012.
16. D. P. Hibar, J. L. Stein, M. E. Renteria, A. Arias-Vasquez, S. Desrivieres, N. Jahanshad, R. Toro, K. Wittfeld, L. Abramovic, M. Andersson, B. S. Aribisala, N. J. Armstrong, M. Bernard, M. M. Bohlken, M. P. Boks, J. Bralten, A. A. Brown, M. M. Chakravarty, Q. Chen, C. R. Ching, G. Cuellar-Partida, A. den Braber, S. Giddaluru, A. L. Goldman, O. Grimm, T. Guadalupe, J. Hass, G. Woldehawariat, A. J. Holmes, M. Hoogman, D. Janowitz, T. Jia, S. Kim, M. Klein, B. Kraemer, P. H. Lee, L. M. Olde Loohuis, M. Luciano, C. Macare, K. A. Mather, M. Mattheisen, Y. Milaneschi, K. Nho, M. Papmeyer, A. Ramasamy, S. L. Risacher, R. Roiz-Santianez, E. J.

Rose, A. Salami, P. G. Samann, L. Schmaal, A. J. Schork, J. Shin, L. T. Strike, A. Teumer, M. M. van Donkelaar, K. R. van Eijk, R. K. Walters, L. T. Westlye, C. D. Whelan, A. M. Winkler, M. P. Zwiers, S. Alhusaini, L. Athanasiu, S. Ehrlich, M. M. Hakobjan, C. B. Hartberg, U. K. Haukvik, A. J. Heister, D. Hoehn, D. Kasperaviciute, D. C. Liewald, L. M. Lopez, R. R. Makkinje, M. Matarin, M. A. Naber, D. R. McKay, M. Needham, A. C. Nugent, B. Putz, N. A. Royle, L. Shen, E. Sprooten, D. Trabzuni, S. S. van der Marel, K. J. van Hulzen, E. Walton, C. Wolf, L. Almasy, D. Ames, S. Arepalli, A. A. Assareh, M. E. Bastin, H. Brodaty, K. B. Bulayeva, M. A. Carless, S. Cichon, A. Corvin, J. E. Curran, M. Czisch, et al. Common genetic variants influence human subcortical brain structures. *Nature*, 520(7546):224–9, 2015.

17. A. L. Zieselman, J. M. Fisher, T. Hu, P. C. Andrews, C. S. Greene, L. Shen, A. J. Saykin, and J. H. Moore. Computational genetics analysis of grey matter density in Alzheimer's disease. *BioData Min*, 7:17, 2014.
18. Xiaohui Yao, Jingwen Yan, Sungeun Kim, Kwangsik Nho, Shannon L. Risacher, Mark Inlow, Jason H. Moore, Andrew J. Saykin, and Li Shen. Two-dimensional enrichment analysis for mining high-level imaging genetic associations. *Brain Informatics*, pages 1–11, 2016.
19. M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, J. Cedarbaum, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, J. Luthman, J. C. Morris, R. C. Petersen, A. J. Saykin, L. Shaw, L. Shen, A. Schwarz, A. W. Toga, J. Q. Trojanowski, and Alzheimer's Disease Neuroimaging Initiative. 2014 update of the Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimers Dement*, 11(6):e1–120, 2015.
20. R. J. Hodes and N. Buckholtz. Accelerating Medicines Partnership: Alzheimer's Disease (AMP-AD) knowledge portal aids alzheimer's drug discovery through open data sharing. *Expert Opin Ther Targets*, 20(4):389–91, 2016.
21. Parkinson Progression Marker Initiative. The Parkinson Progression Marker Initiative (PPMI). *Prog Neurobiol*, 95(4):629–35, 2011.
22. J. McCain. The cancer genome atlas: new weapon in old war? *Biotechnol Healthc*, 3(2):46–51B, 2006.
23. K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*, 26(6):1045–57, 2013.
24. F. S. Collins and H. Varmus. A new initiative on precision medicine. *N Engl J Med*, 372(9):793–5, 2015.
25. M. McCarthy. US to launch major brain research initiative. *BMJ*, 346:f2156, 2013.
26. L. Ohno-Machado. NIH's Big Data to Knowledge initiative and the advancement of biomedical informatics. *J Am Med Inform Assoc*, 21(2):193, 2014.

ADAPTIVE TESTING OF SNP-BRAIN FUNCTIONAL CONNECTIVITY ASSOCIATION VIA A MODULAR NETWORK ANALYSIS

CHEN GAO, JUNGHI KIM and WEI PAN*, for the Alzheimer's Disease Neuroimaging Initiative*

Division of Biostatistics, School of Public Health, University of Minnesota

**E-mail: weip@biostat.umn.edu*

Due to its high dimensionality and high noise levels, analysis of a large brain functional network may not be powerful and easy to interpret; instead, decomposition of a large network into smaller subcomponents called modules may be more promising as suggested by some empirical evidence. For example, alteration of brain modularity is observed in patients suffering from various types of brain malfunctions. Although several methods exist for estimating brain functional networks, such as the sample correlation matrix or graphical lasso for a sparse precision matrix, it is still difficult to extract modules from such network estimates. Motivated by these considerations, we adapt a weighted gene co-expression network analysis (WGCNA) framework to resting-state fMRI (rs-fMRI) data to identify modular structures in brain functional networks. Modular structures are identified by using topological overlap matrix (TOM) elements in hierarchical clustering. We propose applying a new adaptive test built on the proportional odds model (POM) that can be applied to a high-dimensional setting, where the number of variables (p) can exceed the sample size (n) in addition to the usual $p < n$ setting. We applied our proposed methods to the ADNI data to test for associations between a genetic variant and either the whole brain functional network or its various subcomponents using various connectivity measures. We uncovered several modules based on the control cohort, and some of them were marginally associated with the APOE4 variant and several other SNPs; however, due to the small sample size of the ADNI data, larger studies are needed.

Keywords: aSPU test; brain functional connectivity; functional MRI; proportional odds model; single nucleotide polymorphism; weighted gene co-expression network analysis; WGCNA.

1. Introduction

Resting-state functional magnetic resonance imaging (rs-fMRI) is gaining popularity in studies of brain functional connectivity with applications to detection of subtle network reorganizations in Alzheimer's disease.[1] Disruption of connectivity in the brain functional network is related to many pathological conditions in the brain, such as Alzheimer's disease,[2] schizophrenia,[3] or autism.[4] This necessitates the development of methods for modelling the brain functional network its statistical inference.

A network is comprised of nodes and edges connecting the nodes. Based on functional MRI data, a popular choice of nodes are brain regions of interest (ROIs) while the edges are connectivities reflecting statistical dependencies between ROIs. An important network model, the scale-free network,[5] assumes that most nodes in a network are sparsely connected with the exception of a few "hub" nodes that are densely connected with other nodes. In the scale-free network model, new connections are more likely to occur for those hub nodes with already-high

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/howtoapply/ADNIAcknowledgementList.pdf`.

connectivity. There has been empirical evidence supporting this model for brain functional networks,[6] though it is still debatable. In addition, the scale-free network model also admits a modular topological structure, which can be extracted for more efficient analyses for human brains.

Methods for drawing statistical inference to distinguish brain connectivity for different groups of subjects are still under development. The first question encountered is how to define brain functional connectivity. Ref. 7 discussed the choice between Pearson's marginal correlation coefficient and partial correlation coefficient as a network connectivity measure, though other measures are possible and it is yet unclear which one is best. To reduce dimensionality and to reach sparseness, graphical lasso is often used for estimating networks for different groups. Since an estimated network with the imposed sparsity penalty may not demonstrate modular structures, a better approach is to directly discover the modules in a network. A general framework for estimating scale-free networks and detecting modules is proposed in Ref. 8 for gene network analysis, which has gained tremendous popularity in genomics.[9] It starts by defining a similarity measure between two nodes in a network, called adjacency, using the marginal correlation coefficient. Soft-thresholding is then applied, leading to a weighted network. The soft-thresholded adjacency is further transformed to a topological overlap matrix (TOM) element, which is converted to a dissimilarity measure for hierarchical clustering, grouping closely connected nodes together as modules in the network. The above framework not only provides multiple network connectivity measures, but also carries out modular structure identification. The connectivity measures and identified modules in the brain functional network may help statistical inference and offer biological insights.[9]

In this paper, for the first time, we adapt the use of WGCNA for gene expression data to rs-fMRI data, constructing weighted brain functional networks and identifying their subnetworks or modules using the Alzheimer's Disease Neuroimaging Initiative (ADNI) data. We explored using the adjacency matrix element and TOM element, in addition to the marginal correlation or covariance, to characterize connectivity in brain functional networks. Taking advantages of detected network modules, we conduct association analysis of genetic variants with not only the whole brain functional network, but also its various subcomponents, including its modules, which aims to not only improve statistical power, but also offer better biological interpretation. We propose applying a new adaptive association test based on a proportional odds model (POM) accounting for the ordinal nature of the SNP genotype. We found evidence of associations between several network modules and the APOE4 variant, which is by far the most significant genetic risk factor for Alzheimer's disease.

This paper is organized as follows. We first review the method of WGCNA, including its module identification, then introduce the adaptive test based on a POM. We demonstrate the application of our methods to the ADNI data before summarizing our findings and future research directions in the discussion section.

2. Methods

2.1. *Module detection via weighted gene co-expression network analysis*

In this section, we briefly review the work in Ref. 8 on the weighted gene-coexpression network analysis (WGCNA) framework for network construction and module identification.

2.1.1. *Adjacency matrix*

The first step of the WGCNA framework is to define a similarity measure between gene expression profiles; in the current context, we use the BOLD signals in each of multiple ROIs from one or more subjects to calculate a similarity between any two ROIs. The similarity measure is required to take values between 0 and 1. A typical choice of this similarity measure is the absolute value of the Pearson correlation coefficient $s_{uv} = |cor(u, v)|$, for nodes u and v. Another choice, which preserves the sign of correlation, is defined as $s_{uv} = [1 + cor(u, v)]/2$. We refer the first one as unsigned similarity measure, and the second one as the signed similarity measure. From our experience of applications to the ADNI data, the identified modules have negligible differences using either unsigned or signed similarity measure. We used the unsigned similarity measure throughout this paper.

Once the similarity measure is computed, the next step is to transform the similarity matrix $S = [s_{uv}]$ into an adjacency matrix using an adjacency function. Hard thresholding is often used to yield a binary or unweighted network with a 0/1 adjacency indicating no-connection/connection and thus possible loss of information, though a more efficient multi-scale approach with multiple thresholds yielding a set of binary networks has been proposed.[10] Soft thresholding is a simple and popular alternative with more flexibilities. One choice is the power adjacency function

$$a_{uv} = power(s_{uv}, \delta) \equiv |s_{uv}|^{\delta} \tag{1}$$

with parameter δ, which is chosen as the smallest integer such that the scale-free network model fitting is above a certain threshold.

2.1.2. *Topological overlap matrix*

Instead of using only the adjacency matrix, Ref. 11 advocated a topological overlap matrix $\Omega = [\omega_{uv}]$ with its element as a potentially more useful measure that reflects the relative interconnectedness of two nodes u and v after accounting for their shared neighbors. The topological overlap matrix element is defined as

$$\omega_{uv} = \frac{l_{uv} + a_{uv}}{\min\{k_u, k_v\} + 1 - a_{uv}} \tag{2}$$

with $k_u = \sum_v a_{uv}$ and $l_{uv} = \sum_q a_{uq} a_{qv}$. For a binary network with $a_{uv} = 0$ or 1, k_u is the connectivity of node u representing the number of its direct neighbors, while l_{uv} equals the number of nodes that connect both nodes u and v; $\omega_{uv} = 0$ if the nodes u and v are not connected and they are not connected to the same neighbors; in contrast, $\omega_{uv} = 1$ if the nodes u and v are connected and the neighbors of the node with fewer edges are also connected to the one with more edges. For any network, $0 \le a_{uv} \le 1$ implies $0 \le \omega_{uv} \le 1$.

2.1.3. *Module identification*

To identify modules in a network, we need to have a dissimilarity or distance measure. An intuitive way is to convert a similarity measure. Based on the topological overlap matrix element ω_{uv}, we can simply define the dissimilarity measure as $d^{\omega}_{uv} = 1 - \omega_{uv}$. The TOM-based dissimilarity d^{ω}_{uv} is used as the input for average linkage hierarchical clustering. The output from hierarchical clustering is a dendrogram composed of branches and leaves. In a brain functional network, each leaf corresponds to a ROI. The hierarchical clustering algorithm groups the closest ROIs and forms the branches. By cutting the branches of the dendrogram, closely related ROIs are identified as a module. Among the several methods for cutting the branches of the dendrogram, the default used in the WGCNA framework is Dynamic Tree Cut from the R package `dynamicTreeCut`.

Once modules are identified, one can calculate an intramodular connectivity

$$\omega.in_u = \sum_{v \in M} \omega_{uv} \tag{3}$$

for each node u in its module M. Ref. 8 pointed out that intramodular connectivities $\omega.in$ may represent important features of the nodes (i.e. ROIs).

2.2. *An adaptive association test based on the proportional odds model*

Let $Y_i = 0, 1, 2$ denote the count of the minor allele for subject i for a given SNP of interest, then Y_i has $J = 3$ ordered categories. The logistic regression model cannot be applied in this situation, because it only allows the response variable to be binary. A popular choice for ordinal data is the proportional odds model (POM),[12] which we will briefly describe here.

Suppose subject i has p network connectivities denoted by $X_i = (x_{i1}, \ldots, x_{ip})$ and l covariates denoted by $Z_i = (z_{i1}, \ldots, z_{il})$. For the proportional odds model, we define the regression coefficients $\beta = (\beta_1, \ldots, \beta_p)'$ for the network connectivities and $\delta = (\delta_1, \ldots, \delta_l)'$, and a vector of intercepts $\alpha = (\alpha_0, \ldots, \alpha_{J-2})'$. The proportional odds model is

$$\text{logit}[Pr(Y_i \leq j)] = \alpha_j + Z_i\delta + X_i\beta, \qquad j = 0, 1. \tag{4}$$

The likelihood for equation Eq. 4 can be derived based on the multinomial distribution for the categorical variable Y_i, from which maximum likelihood estimates and statistical inference can be obtained as implemented in R package `MASS` or `VGAM`. However, numerical issues such as non-convergence arise when p, the dimension of β, is relatively large as compared to the sample size n.

Here we propose applying a class of tests that are applicable to the high-dimensional setting with $p > n$, from which an adaptive test is constructed to summarize information across the tests. No that most existing tests cannot be applied to the case $p > n$. To test the null hypothesis $H_0 : \beta = (\beta_1, \beta_2, \ldots, \beta_p)' = 0$, we can use the score vector derived in Ref. 13,

$$U_\beta = \sum_{i=1}^{n} \sum_{j=0}^{J-2} (1 - \hat{r}_{i(j-1)} - \hat{r}_{ij}) \cdot I(Y_i = j) \cdot X_i \tag{5}$$

where $\hat{r}_{ij} = \exp(\hat{\alpha} + Z_i\hat{\delta})/[1 + \exp(\hat{\alpha} + Z_i\hat{\delta})]$ comes from the fitted null model of Eq. 4 (i.e. with

$\beta = 0$); $\hat{\alpha}$ and $\hat{\delta}$ are estimated by the `polr` function in the R package `MASS`. Let U_k denote the kth component of the score vector $U_\beta = (U_1, \ldots, U_p)'$. The SPU($\gamma$) test statistic is defined as

$$T_{SPU(\gamma)} = \sum_{k=1}^{p} U_k^\gamma, \tag{6}$$

where $\gamma \geq 1$ is an integer. As the parameter γ increases, a connectivity with a larger absolute value of the score gains a higher weight. In the extreme situation, when $\gamma \to \infty$ as an even integer, $SPU(\infty)$ takes only the maximum component of the score vector, i.e., $T_{SPU(\infty)} = \max_{k=1}^{p} |U_k|$.

The p-values of the SPU tests are computed by permuting the residuals from the null model B times, and the p-value can be calculated as

$$P_{SPU(\gamma)} = \frac{(\sum_{b=1}^{B} I[|T_{SPU(\gamma)}^{(b)}| \geq |T_{SPU(\gamma)}|] + 1)}{(B+1)}, \tag{7}$$

where $T_{SPU(\gamma)}^{(b)}$ is the SPU(γ) statistic based on the bth set of permuted residuals. Since the value of γ that yields highest power cannot be determined a priori, an adaptive SPU (aSPU) test is introduced to combine the evidence across multiple SPU tests,

$$T_{aSPU} = \min_{\gamma \in \Gamma} P_{SPU(\gamma)}, \tag{8}$$

where $P_{SPU(\gamma)}$ is the p-value of $SPU(\gamma)$ test statistics and Γ is a set of integers for the power of aSPU test. In the numerical examples throughout this paper, we chose γ from the set $\Gamma = \{1, 2, \ldots, 8, \infty\}$. To calculate the p-value of T_{aSPU}, we can use the same permutation scheme as used for calculating the p-values of T_{SPU}'s. For each permuted residual set b, after calculating $T_{SPU(\gamma)}^{(b)}$ and its p-value $p_\gamma^{(b)} = (\sum_{b_1 \neq b} I[T_{SPU(\gamma)}^{(b_1)} \geq T_{SPU(\gamma)}^{(b)}] + 1)/B$. Then we can obtain $T_{aSPU}^{(b)} = \min_{\gamma \in \Gamma} p_\gamma^{(b)}$, and the p-value of T_{aSPU} is

$$P_{aSPU} = \frac{(\sum_{b=1}^{B} I[T_{aSPU}^{(b)} \leq T_{aSPU}] + 1)}{(B+1)}. \tag{9}$$

A step-wise procedure is used to gradually increase B if needed. We can start with $B = 10^3$ initially, then increase to $B = 10^5$ (or bigger) if a p-value is smaller than 5×10^{-3} (or smaller). The test is implemented in R package `POMaSPU` to be available on CRAN.

3. Results

3.1. *ADNI Data*

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). We included all subjects from the normal and Alzheimer's disease (AD) groups in the ADNI data. We applied motion correction and global signal regression to reduce noises.

Here we used the power adjacency function $a_{uv} = power(s_{uv}, \beta) = |s_{uv}|^\beta$ (equation (1)). β was selected as the smallest β such that the scale-free model fitting R^2 was above a pre-set threshold 0.85.

3.2. *Distinct modular structures in brain functional networks based on APOE4 SNP genotype scores*

For the ADNI data, we grouped the subjects based on the APOE4 SNP (rs429358) minor allele counts $(0, 1, 2)$. APOE4 plays a major role in the pathogenesis of Alzheimer's disease.[14,15] The APOE4 variant is a major risk factor for both early- and late-onset Alzheimer's disease.[14,15] We removed those subjects with a missing rs429358 value, resulting in a total of 162 subjects. Among them, 73 subjects have no minor allele at rs429358, whereas 67 subjects have one minor allele and 22 subjects have two. In order to establish possible modular structures in brain functional networks in the normal condition, we first applied the WGCNA framework to the rs-fMRI data of the control subjects only. Specifically, for each ROI, we concatenated the BOLD time series of all the control subjects, which were used to calculate the similarity between any two ROIs (i.e. the absolute value of Pearson's correlation between any two BOLD time series), then conducting the subsequent analyses in the WGCNA framework. At the end, we identified four modules based on the data from the control cohort (Figure 1).

Based on the modules identified, we continued to explore them for each APOE4 SNP genotype group. To measure the network connectivities, we used the correlation matrix, covariance matrix, and the topological overlap matrix (TOM). The rows and columns are ordered in the same way as in Figure 1. Distinct modular structures seem to be present in the correlation, covariance and TOM plots across the APOE4 genotype groups (Figure 2).

3.3. *Adaptive testing for SNP-module associations*

Using the APOE4 SNP (rs429358) minor allele counts as the response in a POM, we tested the association between the APOE4 SNP and the network connectivities. Covariates including age, gender and years of education were adjusted. Using the aSPU test, we found that the covariance matrix elements were marginally associated with the APOE4 SNP ($P = 0.033$, Table 1). We further decomposed the whole network connectivities into two exclusive subsets: connectivities within the four modules and those between the modules. Both the between-modular covariance and TOM were associated with the APOE4 SNP with $P < 0.05$.

Next we focused on the network connectivities in each individual module, and tested their association with the APOE4 SNP (Table 2). The network connectivities defined by the correlations in the yellow module showed evidence of association with the APOE4 SNP ($P = 0.017$). In addition, the network connectivities defined by covariance matrix elements in the blue and yellow modules were also associated with the APOE4 SNP ($P = 0.034, P = 0.011$).

Finally we tested for association between each module-specific intramodular connectivity $\omega.in$ and the APOE4 SNP. Only the yellow module showed a significant association with $P = 0.007$.

There are 30 and 19 ROIs in the blue and yellow modules, respectively. The ROIs identified in the yellow modules includes left/right sides of posterior cingulate cortex, angular gyrus, superior frontal cortex, middle frontal cortex, and inferior frontal cortex. For comparison, Ref. 13 identified 18 nodes related to the default mode network (DMN), including left/right sides of superior frontal cortex, medial prefrontal cortex, ventral anterior cingulate cortex, posterior cingulate cortex, parahipppocampal cortex, inferior parietal cortex, angular, middle

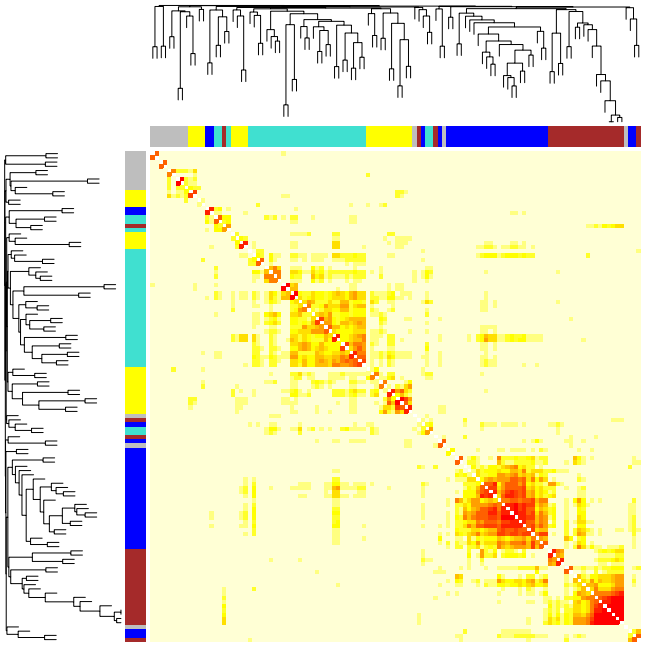

Fig. 1. TOM plot of the whole brain functional network and its modules for normal subjects. The rows and columns are the ROIs, ordered by their distance in the tree.

temporal gyrus, and inferior temporal cortex.[16–18] We found that 15 ROIs in the yellow module are also related to the 18 nodes in the DMN. For example, the posterior cingulate cortex plays a pivotal role in the default mode network of the brain.[19,20] The posterior cingulate cortex is linked to cognitive functions such spatial memory, configural learning, and maintenance of discriminative avoidance learning and.[21,22] It is shown in the DMN that Alzheimer's disease affects the posterior cingulate cortex.[20] Angular gyrus is another region found in both DMN and the yellow module. Loss of grey matter volume in angular gyrus has been associated with dementia and progression to Alzheimer's disease.[23] The association between the APOE4 variant and the network connectivity measures in the yellow module also uncovers some key brain regions in DMN that were found to be affected in Alzheimer's disease.

The ROIs in the blue module includes the left/right sides of hippocampus, lingual gyrus, cuneus, calcarine fissure and superior occipital gyrus, cerebellum and vermis. Hippocampus is well known for its key role in memory.[24] Hippocampal neuronal loss and structural change have been connected with Alzheimer's disease.[25,26] Alzheimer's disease patients have also demonstrated neuronal and glial loss and structural changes in cerebellum and vermis.[27] Lingual gyrus, cuneus, calcarine fissure and superior occipital gyrus are located in the occipital lobe, which are mainly related to vision processing.[28] In addition, lingual gyrus plays an important role in the identification and recognition of words.[29] The association between the APOE4

SNP and the network connectivity measures may reflect the pathological changes of the brain functional network in Alzheimer's disease.

3.4. *GWAS scan with individual modules*

We tested for associations of the SNPs across the whole genome with the functional connectivity measures in the yellow and blue modules respectively. For genotype data, we included all SNPs with a minor allele frequency (MAF) ≥ 0.05, genotyping rate $\geq 90\%$, and passing the Hardy-Weinberg equilibrium test with a p-value > 0.001. After filtering with the above criteria, we obtained 579,382 SNPs.

The genome-wide scan showed that among the SNPs associated with the network connectivities (measured by Pearson's correlation) in the yellow module, rs17114690 on chromosome 14 was the only SNP that had a p-value smaller than 10^{-3}. Three SNPs were founded to be associated with the network connectivities (correlations) in the blue module, with p-values smaller than 10^{-3}. They are located on chromosome 1 (rs7536105, rs11265187) and chromosome 2 (rs17498117). rs7536105 is located in the chromatin interactive region, while rs11265187 is located in the enhancer region of gene olfactory receptor family 10 subfamily J member 9 pseudogene (OR10J9P).

The genome-wide scan also identified 5 SNPs associated with the intramodular network connectivity $\omega.in$ for the yellow module, with $P < 10^{-5}$. They are located on chromosome 1 (rs6656071, rs12043216), chromosome 7 (rs1178127, rs12674460), and chromosome 13 (rs2819239). SNP rs1178127 is a missense variant in gene histone deacetylase 9 (HDAC9),[30] an important gene with function in transcriptional regulation and cell cycle in the Wnt signalling pathway.

4. Discussion

In this paper we adapted WGCNA for network construction and module detection to rs-fMRI data. Based on the identified modules, we also proposed applying a new adaptive association test for single SNP association with the connectivities of the whole network or its components in a proportional odds model. While the whole network was not associated, some module-based connectivities were significantly associated with the APOE4 SNP rs429358. Given the major role of APOE4 in the pathogenesis of Alzheimer's disease, our finding seems plausible, suggesting its possible use for genome-wide scans to detect SNP variants associated with altered brain networks and AD. Although none of the associations was highly or genome-wide significant, it was perhaps due to a too small sample size; larger studies are needed. Our use of modules, with either various ROI-to-ROI connectivities (e.g. TOM in addition to standard correlations) or some module-based node measures (such as intramodular connectivity), not only may reduce the dimension and thus improve the statistical power, but also can enhance result interpretation, highlighting where is the association if any. In particular, we found that intramodular connectivities showed more significant associations with more SNPs, possibly due to their lower dimensions (i.e. p_1 in a module with p_1 ROIs as compared to $p_1(p_1-1)/2$ of ROI-to-ROI connectivities) and/or higher information contents.

The multiple traits used in this paper, including various network connectivity measures in

the whole network or its various subcomponents, differ from most of the previous neuroimaging studies,[31] in which the focus was on some direct measures on ROIs, not their connectivities as shown here. These phenotypes are often high dimensional with dimension exceeding the sample size. Many software packages cannot handle such a situation with $p > n$, which limits their use. The adaptive association test used in this paper can be applied to such high-dimensional traits. It can be a useful and powerful method for identifying associations between high-dimensional neuroimaging traits and SNPs. In this paper, we have focused on the study of the association between neuroimaging phenotypes and SNP genotype scores; however, other ordinal outcomes such as a disease status (e.g. normal, MCI and AD in the ADNI data) can be tested for their associations with neuroimaging and other endophenotypes.

Acknowledgment

This research was supported by NIH grants R01GM113250, R01HL105397 and R01HL116720, and by the Minnesota Supercomputing Institute.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research Development, LLC.; Johnson Johnson Pharmaceutical Research Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

References

1. Y. I. Sheline and M. E. Raichle, *Biological Psychiatry* **74**, 340 (2013).
2. K. Supekar, V. Menon, D. Rubin, M. Musen and M. D. Greicius, *PLoS Computational Biology* **4**, e1000100 (2008).
3. A. Zalesky, A. Fornito, M. L. Seal, L. Cocchi, C.-F. Westin, E. T. Bullmore, G. F. Egan and C. Pantelis, *Biological Psychiatry* **69**, 80 (2011).
4. M. K. Belmonte, G. Allen, A. Beckel-Mitchener, L. M. Boulanger, R. A. Carper and S. J. Webb, *The Journal of Neuroscience* **24**, 9228 (2004).
5. A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).

6. C. C. Hilgetag and A. Goulas, *Brain Structure and Function* , 1 (2015).
7. J. Kim, W. Pan and the Alzheimer's Disease Neuroimaging Initiative, *NeuroImage: Clinical* **9**, 625 (2015).
8. B. Zhang and S. Horvath, *Statistical Applications in Genetics and Molecular Biology* **4** (2005).
9. L. Zhu, J. Lei, B. Devlin and K. Roeder, *arXiv preprint arXiv:1606.00252* (2016).
10. H. Lee, H. Kang, M. K. Chung, B.-N. Kim and D. S. Lee, *IEEE transactions on medical imaging* **31**, 2267 (2012).
11. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A.-L. Barabási, *Science* **297**, 1551 (2002).
12. P. McCullagh, *Journal of the Royal Statistical Society. Series B (Methodological)* , 109 (1980).
13. J. Kim, W. Pan and the Alzheimer's Disease Neuroimaging Initiative, *Unpublished* (2016).
14. J. Kim, J. M. Basak and D. M. Holtzman, *Neuron* **63**, 287 (2009).
15. E. Genin, D. Hannequin, D. Wallon, K. Sleegers, M. Hiltunen, O. Combarros, M. J. Bullido, S. Engelborghs, P. De Deyn, C. Berr *et al.*, *Molecular Psychiatry* **16**, 903 (2011).
16. M. D. Greicius, G. Srivastava, A. L. Reiss and V. Menon, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4637 (2004).
17. L. Q. Uddin, A. Clare Kelly, B. B. Biswal, F. Xavier Castellanos and M. P. Milham, *Human Brain Mapping* **30**, 625 (2009).
18. S. Passow, K. Specht, T. C. Adamsen, M. Biermann, N. Brekke, A. R. Craven, L. Ersland, R. Grüner, N. Kleven-Madsen, O.-H. Kvernenes *et al.*, *Human Brain Mapping* **36**, 2027 (2015).
19. P. Fransson and G. Marrelec, *Neuroimage* **42**, 1178 (2008).
20. R. L. Buckner, J. R. Andrews-Hanna and D. L. Schacter, *Annals of the New York Academy of Sciences* **1124**, 1 (2008).
21. R. J. Maddock, A. S. Garrett and M. H. Buonocore, *Neuroscience* **104**, 667 (2001).
22. R. Leech and D. J. Sharp, *Brain* **137**, 12 (2014).
23. G. Karas, J. Sluimer, R. Goekoop, W. Van Der Flier, S. Rombouts, H. Vrenken, P. Scheltens, N. Fox and F. Barkhof, *American Journal of Neuroradiology* **29**, 944 (2008).
24. L. R. Squire, *Psychological Review* **99**, 195 (1992).
25. B. T. Hyman, G. W. Van Hoesen, A. R. Damasio and C. L. Barnes, *Science* **225**, 1168 (1984).
26. M. J. West, P. D. Coleman, D. G. Flood and J. C. Troncoso, *The Lancet* **344**, 769 (1994).
27. M. Sjöbeck and E. Englund, *Dementia and Geriatric Cognitive Disorders* **12**, 211 (2001).
28. R. Malach, J. Reppas, R. Benson, K. Kwong, H. Jiang, W. Kennedy, P. Ledden, T. Brady, B. Rosen and R. Tootell, *Proceedings of the National Academy of Sciences* **92**, 8135 (1995).
29. A. Mechelli, G. W. Humphreys, K. Mayall, A. Olson and C. J. Price, *Proceedings of the Royal Society of London B: Biological Sciences* **267**, 1909 (2000).
30. C. Fernandez-Rozadilla, L. De Castro, J. Clofent, A. Brea-Fernandez, X. Bessa, A. Abuli, M. Andreu, R. Jover, R. Xicola, X. Llor *et al.*, *PLoS One* **5**, e12673 (2010).
31. L. Shen, S. Kim, S. L. Risacher, K. Nho, S. Swaminathan, J. D. West, T. Foroud, N. Pankratz, J. H. Moore, C. D. Sloan, M. J. Huentelman, D. W. Craig, B. M. DeChairo, S. G. Potkin, C. R. Jack Jr, M. W. Weiner, A. J. Saykin and the Alzheimer's Disease Neuroimaging Initiative, *Neuroimage* **53**, 1051 (2010).

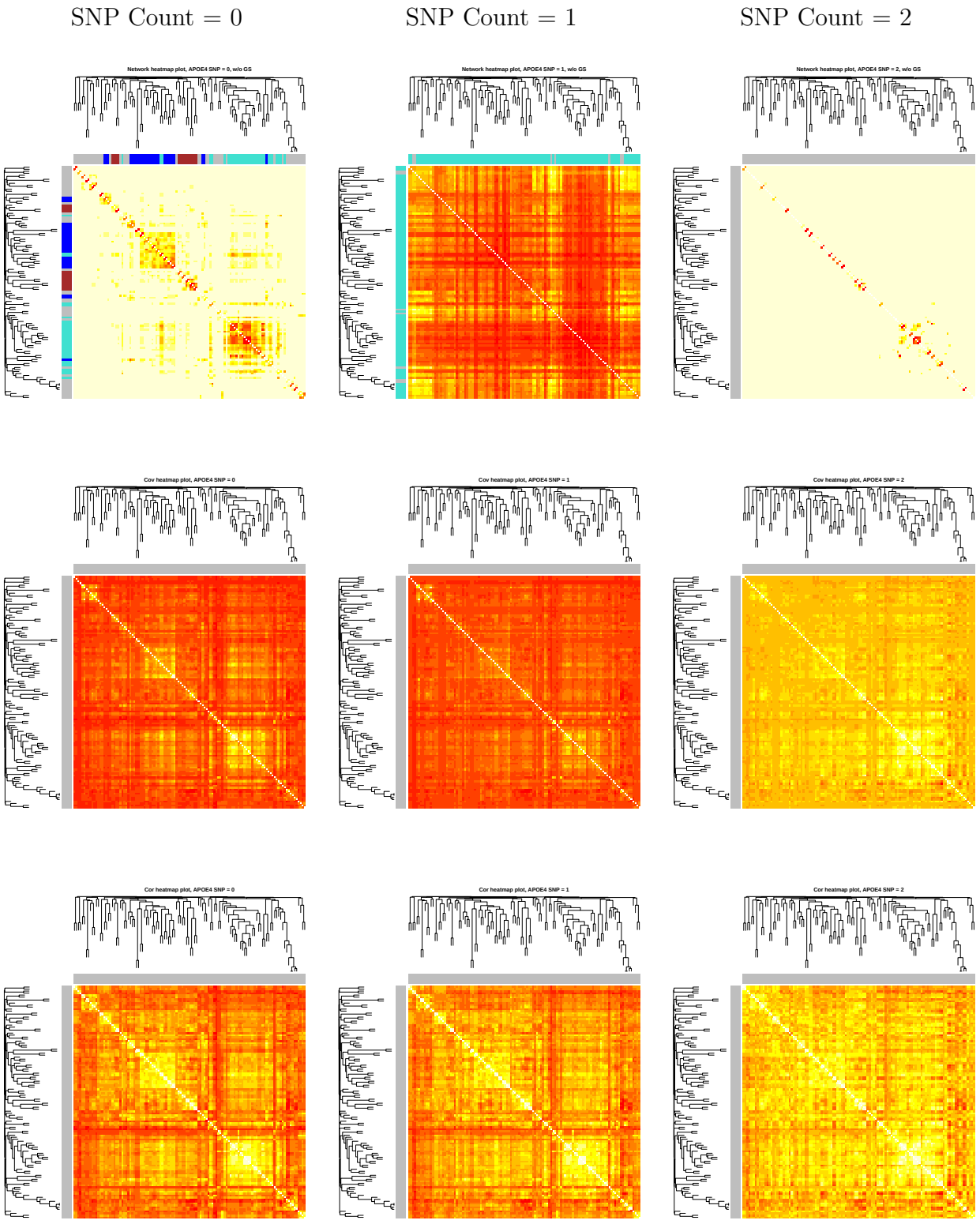

Fig. 2. TOM plot (top), covariance matrix plot (middle) and correlation matrix plot (bottom) of the brain functional networks for the three genotype groups based on APOE4 SNP (rs429358) (with its minor allele counts equal to 0, 1 or 2 from left to right).

Table 1. P-values of the tests for SNP-whole network associations using the correlation, covariance, TOM or adjacency matrix elements as the network connectivity measure respectively. W-mod and Btw-mod stand for within-modular and between-modular, respectively.

	Correlation			Covariance			TOM			Adjacency		
Test	All	W-mod	Btw-mod	All	W-mod	Btw-mod	All	W-mod	Btw-mod	All	W-mod	Btw-mod
SPU(1)	0.477	0.365	0.526	0.052	0.052	0.087	0.530	0.637	0.109	0.477	0.527	0.479
SPU(2)	0.161	0.154	0.207	0.012	0.016	0.008	0.099	0.250	0.014	0.477	0.515	0.487
SPU(3)	0.323	0.202	0.377	0.018	0.025	0.010	0.817	0.902	0.224	0.472	0.482	0.498
SPU(4)	0.150	0.122	0.197	0.019	0.009	0.004	0.325	0.424	0.172	0.463	0.434	0.516
SPU(5)	0.248	0.137	0.299	0.130	0.066	0.004	0.892	0.987	0.317	0.442	0.402	0.528
SPU(6)	0.141	0.101	0.209	0.120	0.008	0.003	0.444	0.533	0.330	0.416	0.365	0.554
SPU(7)	0.216	0.111	0.267	0.429	0.052	0.004	0.657	0.890	0.393	0.381	0.348	0.568
SPU(8)	0.137	0.089	0.225	0.188	0.009	0.004	0.498	0.603	0.410	0.348	0.334	0.583
SPU(∞)	0.122	0.079	0.356	0.210	0.009	0.007	0.463	0.706	0.655	0.263	0.239	0.460
aSPU	0.208	0.146	0.296	0.033	0.025	0.007	0.181	0.384	0.039	0.356	0.328	0.585

Table 2. P-values of the tests for SNP-individual network module associations using the correlation, covariance, TOM or adjacency matrix elements as the network connectivity measure.

		Correlation		Covariance		TOM		Adjacency	
Module	Test	W-mod	Btw-mod	W-mod	Btw-mod	W-mod	Btw-mod	W-mod	Btw-mod
Blue	SPU(1)	0.140	0.272	0.018	0.012	0.946	0.093	0.504	0.488
	SPU(2)	0.090	0.077	0.025	0.001	0.440	0.067	0.515	0.481
	SPU(3)	0.099	0.144	0.028	0.001	0.793	0.396	0.502	0.455
	SPU(4)	0.101	0.071	0.033	0.001	0.528	0.265	0.496	0.423
	SPU(8)	0.139	0.075	0.050	0.011	0.566	0.458	0.464	0.264
	SPU(∞)	0.267	0.070	0.069	0.015	0.571	0.554	0.458	0.180
	aSPU	0.160	0.119	0.034	0.003	0.654	0.141	0.581	0.244
Turquoise	aSPU	0.648	0.172	0.277	0.011	0.139	0.192	0.605	0.309
Brown	aSPU	0.260	0.182	0.016	0.015	0.459	0.219	0.249	0.327
Yellow	SPU(1)	0.040	0.500	0.005	0.084	0.083	0.172	0.357	0.510
	SPU(2)	0.024	0.219	0.005	0.012	0.060	0.155	0.323	0.509
	SPU(3)	0.020	0.350	0.005	0.015	0.147	0.458	0.297	0.514
	SPU(4)	0.016	0.220	0.010	0.006	0.188	0.445	0.262	0.522
	SPU(8)	0.008	0.271	0.100	0.004	0.317	0.690	0.177	0.500
	SPU(∞)	0.006	0.525	0.200	0.008	0.383	0.848	0.129	0.411
	aSPU	0.017	0.354	0.011	0.010	0.106	0.289	0.183	0.524

EXPLORING BRAIN TRANSCRIPTOMIC PATTERNS: A TOPOLOGICAL ANALYSIS USING SPATIAL EXPRESSION NETWORKS

ZHANA KUNCHEVA
Department of Mathematics
Imperial College London, UK
E-mail: z.kuncheva12@imperial.ac.uk

MICHELLE L. KRISHNAN
Perinatal Imaging and Health
King's College London, UK
E-mail: michelle.krishnan@kcl.ac.uk

GIOVANNI MONTANA
Biomedical Engineering Department
King's College London, UK
E-mail: giovanni.montana@kcl.ac.uk

Characterizing the transcriptome architecture of the human brain is fundamental in gaining an understanding of brain function and disease. A number of recent studies have investigated patterns of brain gene expression obtained from an extensive anatomical coverage across the entire human brain using experimental data generated by the Allen Human Brain Atlas (AHBA) project. In this paper, we propose a new representation of a gene's transcription activity that explicitly captures the pattern of spatial co-expression across different anatomical brain regions. For each gene, we define a Spatial Expression Network (SEN), a network quantifying co-expression patterns amongst several anatomical locations. Network similarity measures are then employed to quantify the topological resemblance between pairs of SENs and identify naturally occurring clusters. Using network-theoretical measures, three large clusters have been detected featuring distinct topological properties. We then evaluate whether topological diversity of the SENs reflects significant differences in biological function through a gene ontology analysis. We report on evidence suggesting that one of the three SEN clusters consists of genes specifically involved in the nervous system, including genes related to brain disorders, while the remaining two clusters are representative of immunity, transcription and translation. These findings are consistent with previous studies showing that brain gene clusters are generally associated with one of these three major biological processes.

Keywords: Spatial gene expressions; Biological networks

1. Introduction

The human brain is a complex interconnected structure controlling all elementary and high-level cognitive tasks[1]. This complexity is a result of the cellular diversity distributed across hundreds of distinct brain anatomical structures[2,3]. One of the main tasks of the neuroscience community in the past decade has been to connect the underlying genetic information of the anatomical structures to their underlying biological function[3–5]. A useful data source for such studies is the Allen Human Brain Atlas (AHBA)[3], which provides microarray expression profiles of almost every gene of the human genome with emphasis on an extensive anatomical coverage across the entire human brain.

In this paper, we make use of the experimental data provided by the AHBA project to study the spatial microarray variability at the single gene level. Analyzing the complete transcription architecture of the human brain in this way may be informative of the impact of genetic disorders on different brain regions that would otherwise not be apparent due to the coarse resolution.

To gain new insights into the expression patterns of the human brain and identify potentially important biomarkers, many studies involving the AHBA data explore gene to gene relationships[3,4]. Each gene is represented by its expression levels across anatomical locations. Genes with correlated expression profiles are grouped together based on an appropriate similarity measure. The analysis of the resulting gene co-expression networks provides evidence that transcriptional regulation relates to anatomy and brain function[2–4]. There are also studies that consider the genetic similarity between pairs of regions, and show that transcriptional regulation varies enormously with anatomic location[3,4,6,7]. These findings indicate the necessity to adopt a new representation of a gene's transcription activity that explicitly captures the pattern of spatial co-expression across different anatomical brain regions.

We propose a new and unexplored way to model the spatial variability at the single gene level. For each gene, we create a spatial expression network, or SEN. Each node of the network corresponds to a pre-defined brain region for which we have sufficient transcriptomic data, and each edge weight represents the similarity in gene expression levels, for that gene, between two brain regions. Applying this procedure to genes that have been found to be stably expressed across specimens gives rise to a population of approximately $17,000$ gene networks, each one representing a brain-wide spatial pattern of gene expression. Using this representation, we investigate whether the topological similarity of the SENs reflects the biological similarity of genes through an integrative analysis based on network clustering and gene ontologies. Our hypothesis is that, if clusters of topologically similar SENs can be identified, the corresponding genes within each cluster may also share similar biological properties.

A robust cluster analysis of all SENs has indicated the presence of three large and stable clusters of SENs, each one having significantly different topological features as well as different biological function. In particular, one of the clusters has been found to be uniquely enriched for brain-related terms, neurological diseases and genes with enriched expression in neurons. Overall, our analysis provides evidence supporting the notion that topological proximity of spatial gene networks is indicative of similar biological function.

2. Materials and Methods

2.1. *Spatial Expression Networks (SENs)*

The Allen Human Brain Atlas (AHBA)[3,8] is a publicly available atlas of the human brain with microarray-based genome-wide transcriptional profiling of specific brain regions spanning all major anatomical structures of the adult brain. The data set includes transcriptional profiling data from more than 3500 samples comprising approximately 200 brain regions in 6 clinically unremarkable adult human brains. The Agilent 4×44 Whole Human Genome platform was used for gene expression extraction. Two donors contributed samples representing approximately 1000 structures across the whole brain, while the other four approximately 500 samples

from the left hemisphere. Our analyses is based on 16,906 pre-selected genes from a previous study[5]. We use the normalized expression levels, which were normalized across samples and across different brains as in previous analyses[9].

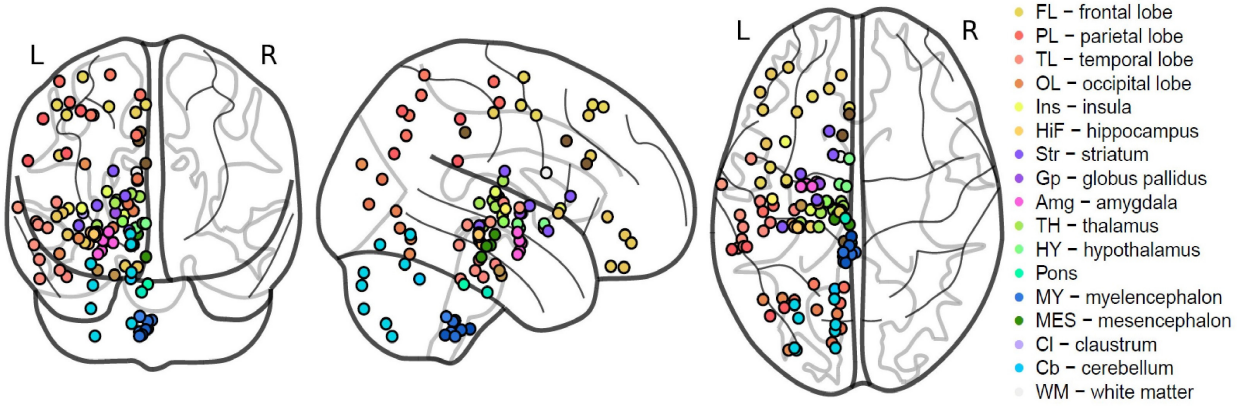

Fig. 1. Anatomical maps of the 105 brain regions used to construct the SENs. The maps show the brain regions as seen from inferior, lateral and superior views, from left to right. All regions are in the left hemisphere and they are located in the Thalamus, Cerebellum, Pons, Midbrain, Medulla and Cerebral cortex. Coloring of the regions is consistent with anatomical tissue and is obtained from AHBA ontology atlas.[8]

For each of the 16,906 genes, we constructed an individual spatial expression network (SEN) representing patterns of expression variability in the brain. Only brain regions with at least one measurement in all 6 brains were included in the analysis resulting in a total of $N = 105$ regions from the left hemisphere, as shown in Fig. 1.

The mean expression level for a gene in brain region i is denoted by g_i. The distribution of the mean and median values for each brain region over all genes were not found statistically different (Kolmogorov-Smirnov test[10]; all $p >> 0.05$). Furthermore, for more than 97% of all region samples across all genes, the standard deviation of the expression values is less than 20% of the mean value, indicating that the mean can be taken as representative of the expression values at a given region for a given gene.

Formally, we define a SEN as a fully connected network $G = (V, E)$ with node set $V = \{i : i = 1, 2, ..., N\}$ indicating the brain regions and weighted edge set $E = \{E_{ij} : i, j = 1, 2, ..., N; i \neq j\}$. Each edge weight $E_{ij} \in [0, 1]$ quantifies the similarity in gene expression between regions i and j. The maximum value is reached when the mean expression levels in the two brain regions are equal. We impose that E_{ij} monotonically decreases with an increasing absolute difference between mean expression levels; accordingly, the edge weights are defined as

$$E_{ij} := \frac{1}{1 + |g_i - g_j|}.$$

This network representation allows us to capture the interconnected variability of gene expression across the brain at the gene level.

2.2. *Clustering SENs*

In order to address our hypothesis that topological similarity may reflect biological similarity, initially we set out to explore whether SENs form naturally occurring clusters. For this we first required an appropriate measure of topological dissimilarity between pairs of SENs. We first mapped each SEN G to a N-dimensional feature vector $\boldsymbol{d} = (d_1, d_2, ..., d_N)$ with each elements representing the node degree, i.e. $d_i = \sum_{j=1}^{N} E_{i,j}$. The degree for each node captures the global transcriptomic similarity of the corresponding brain region to all other brain regions for a given gene. If the node degrees for two SENs are very different, then the corresponding genes have very different global transcriptomic patterns. The dissimilarity between two SENs, G_l and G_k, was taken to be the Euclidean distance between the corresponding feature vectors, $\boldsymbol{d}_l$ and $\boldsymbol{d}_k$.

Three different clustering algorithms were used – partitioning around medoids (PAM)[11], k-means[12] and fuzzy C-means[13] – all providing a partition of all the SENs into k different clusters. To determine an appropriate number of clusters k using each one of these algorithms we performed a stability analysis[12]. The k clusters are deemed "stable" if random changes in the SEN configurations generate almost identical k clusters. To introduce random changes in the networks, we use a randomization strategy by which the observed networks in network space Γ are perturbed slightly. For this analysis we used two different randomization procedures: (a) vertex permutations, i.e. we permuted the node labels of a random subset of networks so as to preserve the node degrees but not their order, (b) edge perturbation, i.e. we perturbed the edge weights of a random subset of networks so as to make the cluster robust against white noise.

To obtain a measure of cluster instability, we use the following steps: First, we generate perturbed versions $\Gamma_b\,(b = 1, 2, ..., b_{\max})$ of Γ, and cluster the networks in Γ_b into k clusters thus obtaining $\mathcal{C}_b(k)$. In addition, we randomize the cluster assignments[14] in $\mathcal{C}_b\,(k)$ to obtain random clustering $\mathcal{C}_{b,\mathrm{rand}}\,(k)$. Second, for $b, b' = 1, 2, ..., b_{\max}$, we compute the pairwise distances $[1 - NMI(\mathcal{C}_b(k), \mathcal{C}_{b'}(k))]$ between the clusterings $\mathcal{C}_b(k)$ and $\mathcal{C}_{b'}(k)$, and between the randomized clusterings $\mathcal{C}_{b,\mathrm{rand}}(k)$ and $\mathcal{C}_{b',\mathrm{rand}}(k)$. The normalized mutual information (NMI) is used as a similarity measure between partitions[15]. The cluster instability index is defined as the mean distance between clusterings $\mathcal{C}_b(k)$, i.e.

$$I(k) = \frac{1}{b_{\max}^2} \sum_{b,b'=1}^{b_{\max}} \left[1 - NMI(\mathcal{C}_b(k), \mathcal{C}_{b'}(k))\right]. \tag{1}$$

We use the normalized instability index, $I_{\mathrm{norm}}(k) := I(k)/I_{\mathrm{rand}}(k)$, which corrects for a scaling[14] of $I(k)$ with an increasing number of clusters k. We choose number of clusters k that gives the lowest $I_{\mathrm{norm}}(k)$.

2.3. *Topological characterization of SEN clusters*

To characterize the topological properties of SENs in each cluster, we use global topological measures that capture different aspect of the network such as its density, the tendency of its nodes to cluster and form communities, the presence of central and hub nodes. Overall, we use eight such different measures: average node degree[16], average closeness centrality[16], weighted

diameter[17], global clustering coefficient for weighted networks[17], number of non-overlapping communities, average authority score[18], the number of nodes with authority score > 0.95, and the number of nodes with authority score < 0.05. All measures were computed for all SENs within each cluster. To test for statistically significant differences in network topology across clusters, we performed a multivariate ANOVA test[19].

Furthermore, for each SEN we derived a measure of community structure[20]. In our context, the presence of a community in a given SEN indicates that there is a set of highly interconnected brain regions whose gene expression similarity is higher compared to the rest of the network. For this analysis we used the Fast Greedy algorithm[21], which is based on the optimization of the modularity function that sums the edge weights within a community and corrects for the expected edge weights by chance. The algorithm is discriminative of small edge weight differences and can yield sensitive separation of brain regions into communities. Genes with similar community structures indicate the presence of similar local coherent transcriptomic patterns for groups of brain regions.

For each cluster, we quantify the similarity of a pair of brain regions using the communities detected in all the SENs by counting the number of times the two regions fall within the same community. This count is then divided by the total number of SENs in the cluster in order to obtain an index lying in the $[0, 1]$ range, which we call the "coherence index". Values close to 1 indicate high coherency between the two brain regions, i.e. the average tendency to fall within communities of highly interconnected brain regions.

2.4. *Biological characterization of SEN clusters*

In order to investigate whether naturally occurring clusters formed by SENs can be related to distinct biological function, we require a procedure which assigns representative biological terms to each cluster. For this purpose we use a Gene Ontology (GO) enrichment analysis pipeline which first collects broad GO information for the biological context of genes in each of the main clusters, and then reduces this information to representative GO terms for final interpretation of the clusters.

Each SEN cluster was first annotated for significantly enriched Biological Process (BP) terms using a standard hypergeometric test for over-represented terms ($p < 0.001$) implemented in the GOstats R package[22]. Using a clustering methodology implemented in the tool REVIGO[23], we group semantically similar GO terms based on the established *SimRel* measure. The algorithm finds a representative term for each group based on the enrichment p-values, with a bias away from very general parent GO terms. The size of the resulting summary list is controlled by setting the threshold for the *SimRel* similarity measure at 0.5. Results are summarized by retaining the cluster representatives for each GO term that can reveal underlying function of these clusters.

Genes in each of the clusters were also annotated for disease enrichment using the WebGestalt tool[24], which interfaces with the GLAD4U platform[25] to retrieve and prioritize disease-gene links from publications, using a hypergeometric test with multiple testing correction and the genome as background.

3. Experimental results

3.1. *Topologically different SEN clusters*

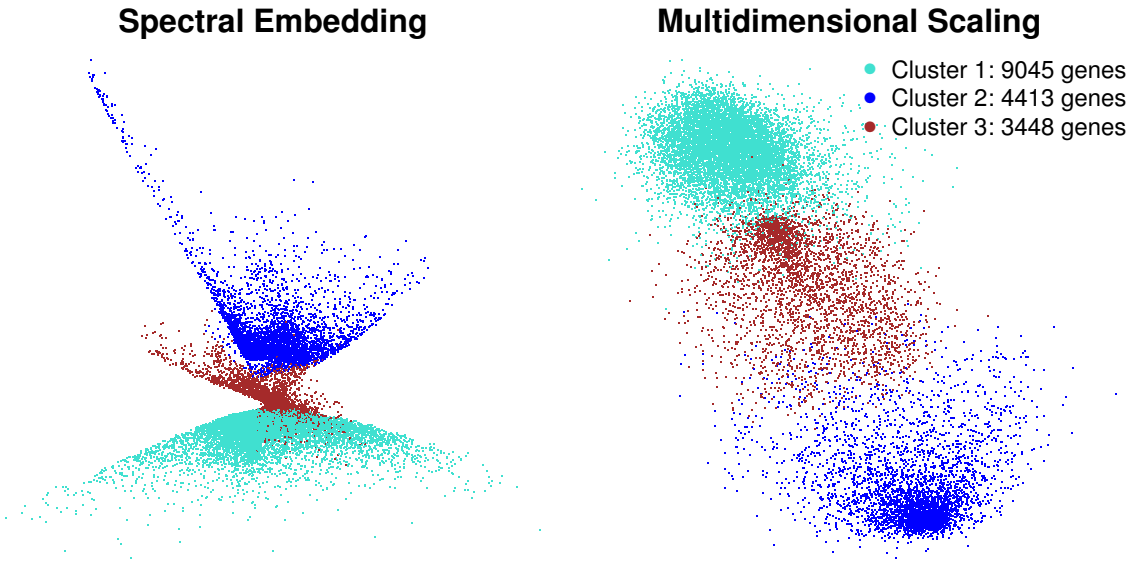

Fig. 2. Two-dimensional visualization of all SENs using two different dimensionality reduction algorithms: spectral embedding[26] (left) and multidimensional scaling[27] (right). The color scheme indicates the cluster membership as determined by the PAM algorithm. Both visualizations indicate three main clusters.

All SENs were clustered into up to six clusters using the procedures outlined in Sec. 2.2. The two instability analyses were each performed using $b_{\max} = 500$. Using the first randomization scheme, 5% of networks were randomly sampled for node permutation, while in the second procedure 20% of networks were randomly sampled and white noise was introduced by adding ±20% to each edge weight. The results for all three clustering procedures, Tab. 1, show that PAM clustering has the lowest instability followed by fuzzy C-means. Furthermore, for all three clustering methods grouping data into two and three clusters leads to the lowest instabilities.

Table 1. Different stability analyses for three different clustering algorithms using two randomization strategies (vertex and edge permutation).

	Vertex permutation			Edge perturbation		
$I_{\text{norm}}(k)$	PAM	Cmeans	k-means	PAM	Cmeans	k-means
$I_{\text{norm}}(2)$	0.016	0.020	0.065	0.009	0.015	0.065
$I_{\text{norm}}(3)$	**0.018**	**0.023**	**0.076**	**0.010**	**0.016**	**0.071**
$I_{\text{norm}}(4)$	0.023	0.031	0.092	0.019	0.033	0.180
$I_{\text{norm}}(5)$	0.026	0.038	0.171	0.025	0.030	0.191
$I_{\text{norm}}(6)$	0.031	0.080	0.187	0.027	0.086	0.208

The PAM algorithm was chosen to generate the final partitions as it yields the lowest instability index. As an additional validation to support the choice of three PAM clusters, we used three internal validation measures: the Sillhouette width[28], the Dunn index[13] and the within-cluster variance[29]. The Dunn index and Silhouette width support the presence of two to three clusters, see Tab. 2. However, the intra-cluster variance, which is known to be more sensitive to the existence of sub-clusters[30], shows that grouping data into two clusters leads to high within-cluster variability compared to a higher number of clusters. By taking all these criteria into account, we have chosen to consider $k = 3$ since this leads to the lowest instability and within-cluster variability whilst having as high as possible Dunn and Sillhouette scores.

Table 2. Cluster validation measures for clustering SENs into k clusters using PAM.

k	Dunn	Silhouette	Within-cluster Variance
2	2.20	0.66	0.276
3	**1.20**	**0.44**	**0.225**
4	0.61	0.30	0.215
5	0.63	0.23	0.223
6	0.48	0.18	0.211

In an attempt to visually assess whether this choice seems appropriate, we used a distance-preserving projection of all 16906 SENs into a 2D-dimensional space using two different dimensionality reduction procedures: spectral embedding[26] and multidimensional clustering[27]. The resulting projections can be found in Fig. 2. All three clusters – 1 (turquoise), 2 (blue) and 3 (brown) – appear well-separated.

3.2. *Topological differences amongst SEN clusters*

To validate that the three SEN clusters have distinct topological structure, we used the eight global network measures outlined in Sec. 2.3. The frequency distribution of the topological measures for each cluster is summarized in Fig. 3 where a clear mean difference can be observed for each individual measure across clusters. Using a MANOVA test, we reject the null hypothesis of equality of topological features across clusters ($p < 2.2e-16$; Wilk's $\Lambda = 0.3589$).

We have found that Cluster 1 mostly consists of SENs with the highest node degree, centrality measures, diameter, authority score and number of nodes with high authority score, while there are only a few number of communities and few nodes with low authority score. These properties imply coherent expression levels across all brain regions. On the other hand, Cluster 2 comprises of SENs with the lowest node degree, centrality measures, diameter, authority score and number of nodes with high authority scores, and the highest number of communities and nodes with low authority score. This indicates that most SENs within this cluster are sparse, and that there is high variability between expression levels across brain regions. Finally, Cluster 3 consists of SENs with medium ranged values for all network

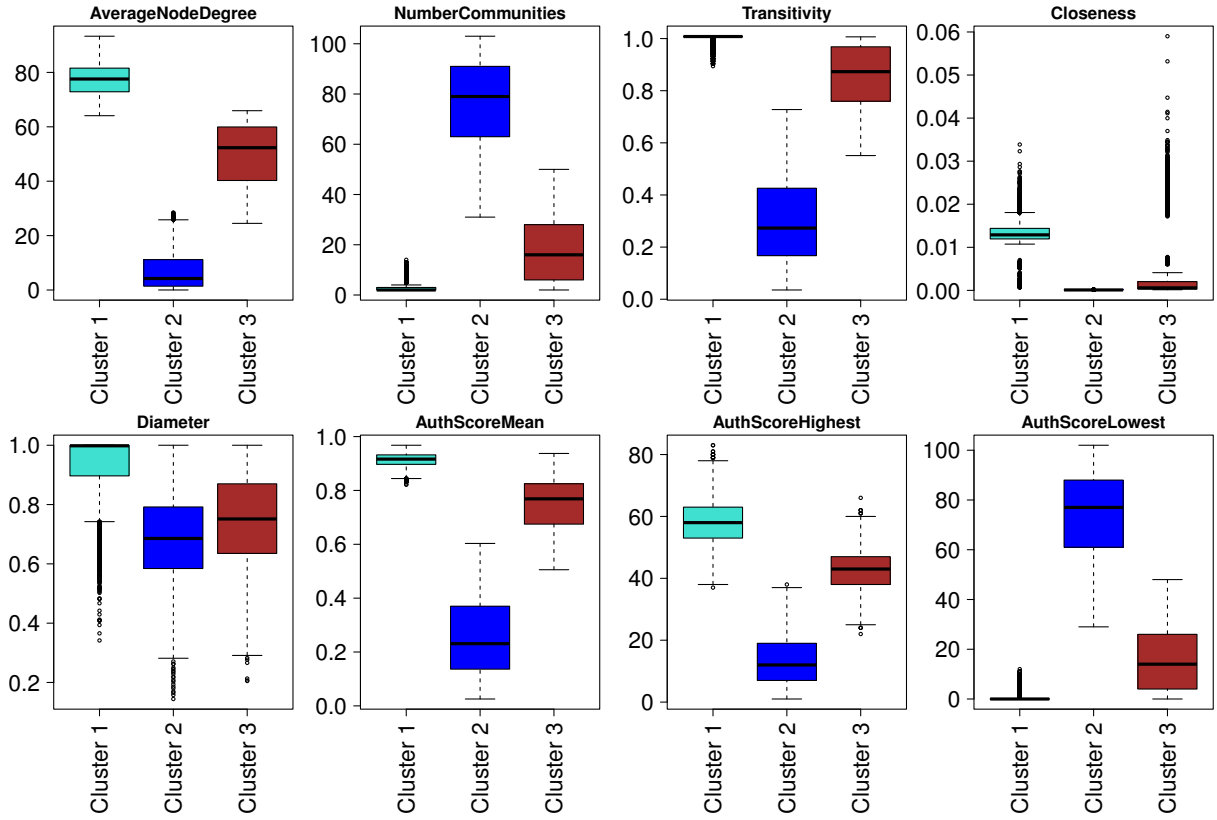

Fig. 3. Distribution of topological network measures in the three clusters obtained using the PAM algorithm. High node degrees imply high edge weights with fewer low-weighted shortest paths and fewer discrepancies in edge values. This leads to high transitivity and closeness values, simultaneously reducing the number of communities SENs are partitioned into. Higher node degrees lead to more nodes having high authority scores thus increasing both the average authority scores and the number of nodes with high authority. Low node degrees signify sparseness of the SENs and more low-weighted shortest paths. This results in more nodes being grouped in their own communities, in addition to low closeness and transitivity. Sparse networks and low node degrees result in lower authority scores and fewer nodes with high authority score.

measures, implying moderate variability between expression levels across brain regions.

3.3. *Biological differences amongst SEN clusters*

We investigated the local transcriptomic patterns within each of the three clusters using the "coherence index" defined in Sec. 2.3. The three clusters have different transcriptomic patterns, Fig. 4, and comparing heatmaps of the three clusters to one for all 16906 genes shows that Cluster 1 is closest to the genome-wide global patterning, while Cluster 2 and Cluster 3 are carriers of imposed heterogeneity. The patterns of the 16906 genes are also consistent with existing work, and largely replicate previous findings[3,4,6]. In particular, homogeneity within the Neocortex and Cerebellum, and increased heterogeneity in the Basal Ganglia, have been previously reported. Cluster 2 has few coherency patterns in the Basal Ganglia regions and Cerebellum. Cluster 1 exhibits high homogeneity within the Cerebellum and the Neocortex, and between subdivisions of the subcortical structure and the Hippocampus. Cluster 3 appears to have coherent patterns in the Cerebellum and the Neocortex but increased variability in

the Basal Ganglia.

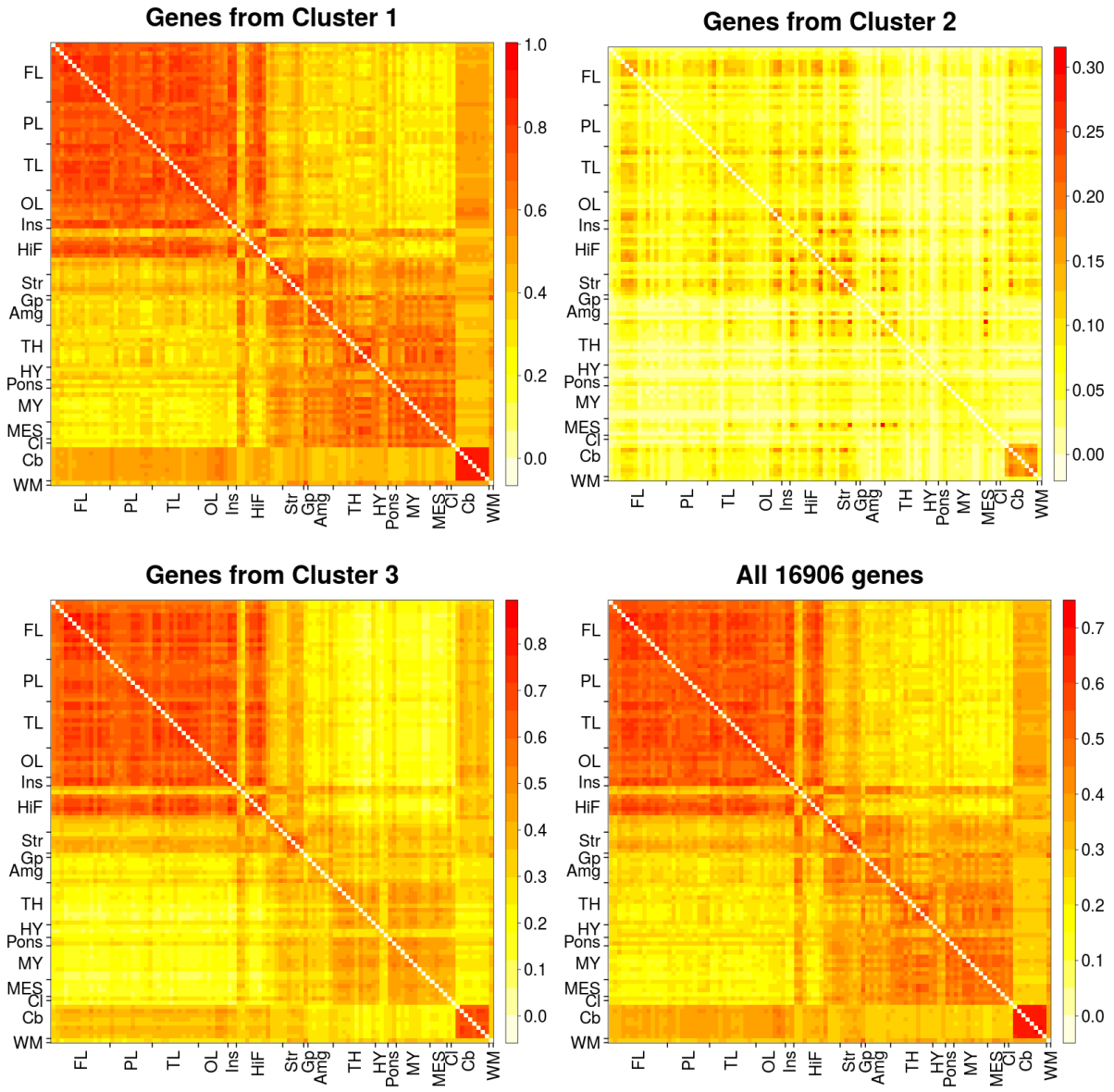

Fig. 4. Heatmaps representing the "coherence index" between pairs of brain regions in each of the three SEN clusters and across all 16906 genes. Each pixel on the heatmap is the "coherence index" between the two corresponding brain regions. Each heatmap is accompanied by a color key, where higher values indicate high homogeneity of expression levels and lower values indicate heterogeneous expression levels. The 105 brain regions are mapped to 17 major brain structures using the AHBA ontology atlas[8] and abbreviated as indicated in Fig. 1.

Obtaining detailed annotation as described in Sec. 2.4 revealed that all three clusters are significantly enriched ($p < 0.001$) for a variety of GO BP terms. We reduced these large sets of GO terms to smaller non-redundant sets by applying REVIGO[23].

The BP representative terms selected on the basis of enrichment p-values and semantic

similarity indicate that Cluster 1 genes can be described primarily by "RNA processing" and "ribonucleoprotein complex biogenesis". Cluster 2 genes are predominantly involved in immunity including "immune system process", "leukocyte proliferation" and "G-protein coupled receptor signaling pathway" terms primarily associated to the immune system, whereas Cluster 3 genes are uniquely involved in "behavior", "metal ion transport" and "nervous system development". On closer inspection of Cluster 3, these representative terms comprise several linked biological processes specific to the Nervous System, and which are not found on either Cluster 1 or 2, such as "synaptic transmission" and "dendrite extension".

The significant disease enrichment (adjusted $p < 0.001$) also supported the functional distinctiveness of the three clusters, with Cluster 1 being enriched for Mitochondrial disease, Cluster 2 being significantly enriched for genes involved with Immune System and Inflammatory disease, and Cluster 3 being principally involved in Nervous system disorders. Given the observed functional differentiation between clusters, we investigated whether this might correspond to cell-type specialization. We obtained lists of neuron- and microglia-enriched genes in a repository of detailed RNA-sequencing and splicing data from purified cell cultures[31] , and computed significant intersections using the SuperExactTest[32]. This showed that genes in Cluster 3 have significant overlap with neuron- and microglia-specific genes ($p < 0.05$). Cluster 2, on the other hand, has a unique association to microglia-specific genes only ($p < 0.05$).

4. Discussion

Analyzing the transcriptome architecture of the human brain is a challenging task due to the high-dimensionality and biological complexity of the data. This is compounded by technical factors related to sample acquisition and measurement error that can influence the results. We addressed the issue of anatomical variability in gene expression by proposing to model each gene's spatial co-expression pattern across anatomical regions as an individual spatial network, or SEN. To explore whether topological similarity of gene expression as captured by SENs is related to biological similarity, we used network dissimilarity to obtain clusters of genes with similar patterns of spatial co-expression. We aimed to gain additional insights into the biological interpretation of regional anatomical specialization of the brain.

We demonstrated that there is evidence to support the presence of three topologically distinct clusters of SENs, with each cluster being characterised by particular network properties. Furthermore, investigating the community structure of the SENs, we identified possible anatomical basis for the difference in the topological properties in the three clusters. The differences between clusters are mainly due to the heterogeneity of expression levels in the Basal Ganglia, and between the Neocortex and Cerebellum.

We also found these three topologically distinct clusters to have biologically distinct properties. On closer inspection we find Cluster 3 to be specific to the nervous system, while Cluster 2 appears to be involved with immunity and Cluster 1 with transcription and translation. These associations are in line with previous results on the AHBA data set[3,4], where the majority of clusters obtained using WGCNA[33], a well-known gene clustering procedure, were also associated to immunity, nervous system or transcription and translation.

To gain an insight into possible cellular contributions to these differences, we included

cell-type specific data and observe that the overlap of neuron- and microglia- specific genes in Cluster 3 is in keeping with current hypotheses regarding the significant interactions between these two cell-types, including the possible modulatory activity of microglia in synaptic pruning and cell communication beyond purely immune functions[34].

We found significant disease associations for all three clusters, implying the high biological impact of the genes involved and the utility of our modular clustering approach for the identification of therapeutic targets. There is a preponderance of neurological and neuropsychiatric conditions linked to Cluster 3 genes, and immune disorders linked to Cluster 2, reflecting their biological functions as described above and supporting those annotations.

One important concern was whether the above results were specific to using node degrees or they could be reproduced using other feature vectors. Thus we constructed two different sets of feature vectors based on node centrality as captured by the authority score and based on the raw edges of the SEN. Based on each new set of feature vectors, results not included in this paper demonstrated evidence to support the presence of three topologically distinct clusters of SENs. For both feature vectors, the three clusters were again marked by different topological properties although there were shifts in the distributions of those properties. Even so, in both cases the three clusters were uniquely associated to the immune system, nervous system or transcription and translation.

For comparison purposes, we used WGCNA on the gene expression values of the 16906 genes for the 105 brain regions. Results not included in this paper showed that WGCNA did not assign a cluster membership to the majority of genes in Cluster 2 due to the sparseness of their expression levels. More and smaller clusters were discovered with higher instability. The advantage of our method compared to WGCNA is that the structure of SENs allows us to use a number of clustering procedures to detect stable gene clusters, whose validation could be achieved using both topological and biological measures. We determine the biological function of a cluster using the gene ontology of the entire set of genes in the cluster, which is robust to slight changes in the cluster membership.

A next step in the analysis of SENs should consider additional clusters to detect more specialized biological functions. Furthermore, it is well known that gene expressions in the cerebellum, subcortical and cortical regions differ significantly from each other based on their composition of different cell types[3,4]. Future work in this direction will include an analysis where only neocortex regions are used to construct SENs.

5. Conclusion

An important and challenging task in studying the brain transcriptional architecture is integrating and modelling the high dimensionality of the gene expression across the brain. To the best of our knowledge, our work is the first to perform a region-wise comprehensive profiling of gene-specific co-expression patterns across the human brain. By modelling gene expression as SENs and employing network embeddings, we identified distinct clusters of genes associated to specific biological functions, topological properties and cell-types, with potential implications for neuropsychiatric disease. Modelling genes as SENs across brain regions could be used for future studies in helping to identify genes with particular co-expression patterns across

a set of spatial brain locations of interest, enabling the identification of genes that act in spatially contextualized clusters with high biological impact. As more microarray gene expression data become available at higher spatial resolution and cell-type specificity, modelling gene co-expression across the brain will be increasingly important to understanding the brain transcriptome architecture at a microstructural scale.

References

1. M. C. Oldham, G. Konopka *et al.*, *Nat. Neurosci.* **11**, 1271 (nov 2008).
2. M. Hawrylycz, L. Ng *et al.*, *Neural Netw.* **24**, 933 (nov 2011).
3. M. J. Hawrylycz, E. S. Lein *et al.*, *Nature* **489**, 391 (sep 2012).
4. M. Hawrylycz, J. A. Miller *et al.*, *Nat. Neurosci.* **18**, 1832 (nov 2015).
5. J. Richiardi, A. Altmann *et al.*, *Science* **348**, 1241 (jun 2015).
6. A. Mahfouz, M. van de Giessen *et al.*, *Methods* **73**, 79 (mar 2015).
7. P. Goel, A. Kuceyeski *et al.*, *Hum. Brain Mapp.* **35**, 4204 (aug 2014).
8. Allen Institute for Brain Science., Allen Human Brain Atlas (2014).
9. Allen Human Brain Atlas, *Technical White Paper: Microarray Data Normalization*, tech. rep., Allen Institute (2013).
10. G. Marsaglia, W. W. Tsang *et al.*, *J. Stat. Softw.* **8**, 1 (2003).
11. L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (John Wiley & Sons Ltd, 1990).
12. U. Von Luxburg, *Clustering Stability: An Overview* (Now Publishers Inc., 2010).
13. J. C. Dunn, *J. Cybern.* **3**, 32 (1973).
14. T. Lange, V. Roth *et al.*, *Neural Comput.* **16**, 1299 (2004).
15. M. Meila, *J. Multivar. Anal.* **98**, 873 (may 2007).
16. T. Opsahl, F. Agneessens *et al.*, *Soc. Networks* **32**, 245 (jul 2010).
17. A. Barrat, M. Barthélemy *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **101**, 3747 (mar 2004).
18. J. M. Kleinberg, *J. ACM* **46**, 604 (sep 1999).
19. S. Scheiner, *Des. Anal. Ecol. Exp.* , 94 (2001).
20. S. Fortunato, *Phys. Rep.* **486**, 75 (2010).
21. A. Clauset, M. Newman *et al.*, *Phys. Rev. E* **70**, p. 066111 (dec 2004).
22. S. Falcon and R. Gentleman, *Bioinformatics* **23**, 257 (jan 2007).
23. F. Supek, M. Bošnjak *et al.*, *PLoS One* **6**, p. e21800 (jan 2011).
24. B. Zhang, S. Kirov *et al.*, *Nucleic Acids Res.* **33**, W741 (jul 2005).
25. J. Jourquin, D. Duncan *et al.*, *BMC Genomics* **13 Suppl 8**, p. S20 (jan 2012).
26. U. V. Luxburg, *A Tutorial on Spectral Clustering*, tech. rep., Max Planck Institute for Biological Cybernetics (2007).
27. I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications* (Springer Science & Business Media, 2005).
28. P. J. Rousseeuw, *J. Comput. Appl. Math.* **20**, 53 (nov 1987).
29. M. Halkidi and M. Vazirgiannis, Clustering Validity Asessment: Finding the Optimal Partitioning of a Data Set, in *Proc. 2001 IEEE Int. Conf. Data Min.*, (IEEE Comput. Soc, 2001).
30. Y. Liu, Z. Li *et al.*, Understanding of Internal Clustering Validation Measures, in *2010 IEEE Int. Conf. Data Min.*, (IEEE, dec 2010).
31. Y. Zhang, K. Chen *et al.*, *J. Neurosci.* **34**, 11929 (sep 2014).
32. M. Wang, Y. Zhao *et al.*, *Sci. Rep.* **5**, p. 16923 (jan 2015).
33. S. Horvath, *Weighted Network Analysis: Applications in Genomics and Systems Biology* (Springer, 2011).
34. M.-È. Tremblay, R. L. Lowery *et al.*, *PLoS Biol.* **8**, p. e1000527 (jan 2010).

INTEGRATIVE ANALYSIS FOR LUNG ADENOCARCINOMA PREDICTS MORPHOLOGICAL FEATURES ASSOCIATED WITH GENETIC VARIATIONS*

CHAO WANG
Electrical and Computer Engineering, The Ohio State University
Columbus, Ohio, 43210, USA
Email: wang.2031@osu.edu

HAI SU
Biomedical Engineering, University of Florida
Gainsville, Florida, 32611, USA
Email: hai.su@bme.ufl.edu

LIN YANG
Biomedical Engineering, University of Florida
Gainsville, Florida, 32611, USA
Email: lin.yang@bme.ufl.edu

KUN HUANG
Biomedical Informatics, The Ohio State University
Columbus, Ohio, 43210, US
Email: kun.huang@osumc.edu

Lung cancer is one of the most deadly cancers and lung adenocarcinoma (LUAD) is the most common histological type of lung cancer. However, LUAD is highly heterogeneous due to genetic difference as well as phenotypic differences such as cellular and tissue morphology. In this paper, we systematically examine the relationships between histological features and gene transcription. Specifically, we calculated 283 morphological features from histology images for 201 LUAD patients from TCGA project and identified the morphological feature with strong correlation with patient outcome. We then modeled the morphology feature using multiple co-expressed gene clusters using Lasso-regression. Many of the gene clusters are highly associated with genetic variations, specifically DNA copy number variations, implying that genetic variations play important roles in the development cancer morphology. As far as we know, our finding is the first to directly link the genetic variations and functional genomics to LUAD histology. These observations will lead to new insight on lung cancer development and potential new integrative biomarkers for prediction patient prognosis and response to treatments.

1. Introduction

Lung cancer is one the most deadly cancers in the world. Among lung cancers, lung adenocarcinoma (LUAD) is a subtype of the non-small cell lung cancer (NSCLC) and is the most common histological type of lung cancers (*1*). However, despite the fact that it is a sub-classification of lung cancer, LUAD is a heterogeneous group of tumors with a highly variable prognosis and responses to treatment (*2*).

The high-throughput sequencing technologies are making targeted therapies possible for LUAD (*3*). The advance of these technologies allows molecular diagnostic biomarkers for the detection of lung cancer in addition to computed tomography (CT) screening (*4–7*). For example, the utility of epidermal growth factor receptor (EGFR) mutation testing is strongly recommended (*8*) in clinical practice. However, although EGFR-mutant lung cancers are

* This work is supported by UK-OSU Joint CCTS grant, NCI ITCR 1U01CA188547-01A1 grant, the OSU Pelotonia Fellowship, and the Ohio Supercomputer Center.

sensitive to EGFR tyrosine kinase inhibitors (TKIs), they develop resistance (*9*). Therefore, novel biomarkers for for LUAD are needed for enhanced personalized treatment.

Lung cancer diagnosis and classification have been traditionally based on imaging approaches, such as CT and histopathology (*10*, *11*). For instance, five distinct histologic subtypes and radiologic patterns have been reported recently. Traditionally, histopathology images serve as a golden standard for lung cancer diagnosis. Cellular and inter-cellular level morphology are usually used by the pathologists for making diagnosis decisions. However, the current pathology diagnosis is commonly based on individual pathologists' interpretations of the samples which are subject to large inter-observer variations and low throughput analysis. Unbiased quantitative pathology methods are showing promise by offering more cellular information (*12–14*). Recently, pathology informatics study on lung cancer has attracted more interests. In one study (*15*), the diagnostic significance of nuclear features in differentiating small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) was investigated. Edwards et.al.(*16*) showed that adenocarcinoma diagnosis is more challenging compared to squamous carcinoma. An early automatic pathology analysis system was proposed in (*17*). In the study by Mijović et al. (*18*), diagnostic values of seven Karyometric variables are examined for diagnosis of major histological types of lung carcinoma. In Zhang et al's study (*19*), an image classification system is proposed to differentiate lung adenocarcinoma and squamous carcinoma. The work by Yao et al (*20*) developed topological features for lung cancer diagnosis. Compared to genomic biomarkers, advanced imaging may provide more clinically relevant information.

In order to take advantage of both the richness of histopathological information and molecular profiles, we aim to develop an integrative computational pipeline that exploits diagnostic images and mRNA expression. A related work on lung cancer was recently published on integrating histopathologcal images with genetic data for outcome prediction (*21*). The pipeline allowed us to discover the associations between cellular level and molecular level phenotypes, and thus novel biomarkers can be unveiled. In this paper, we extracted 283 histopathological features from LUAD tissue slides and initially attempted to identify co-expression gene clusters that have high correlation with these image features. Such approach in other cancers has led to new insight on cancer biology and new potential biomarkers (*22*). However, as shown in this paper, the morphology of LUAD is much more complicated and it turned out that the morphological features have low correlations with gene expression profiles. Figure 1 shows a 'highly-correlated' pairs between the imaging features and gene clusters. It is thus plausible that the LUAD morphology is regulated by any particular group of genes; instead a specific morphological characteristic is a manifestation of a combined effect from multiple groups of genes. Based on these quantitative experiments, we assert that a multivariate model is needed.

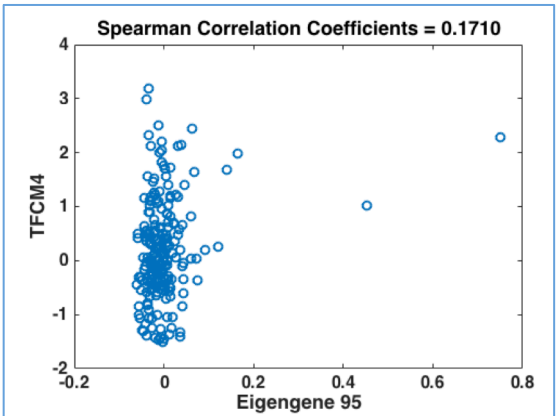

Figure 1: The scatter plot between the eigengene 95 and the *tfcm4*, which have the highest SCC between all eigengenes with tfcm4 (SCC = 0.170).

Therefore in this paper, we demonstrate that the morphological characteristics of LUAD can be explained by a combination of multiple gene clusters identified using sparse modeling based on the Lasso algorithm. In addition, we found that many of the gene clusters are associated with putative copy number variations, implying that genetic variations play important roles in the development cancer morphology. As far as

we know, our finding is the first to directly link the genetic variations and functional genomics to LUAD histology. These observations will lead to new insight on lung cancer development and potential new integrative biomarkers for prediction patient outcome and response to treatments.

2 Methods and Materials

Our analysis involve molecular and histological analysis based on data from The Cancer Genome Atlas (TCGA) LUAD project. The data we use include mRNA profiling, histological images and clinical data including survival information.

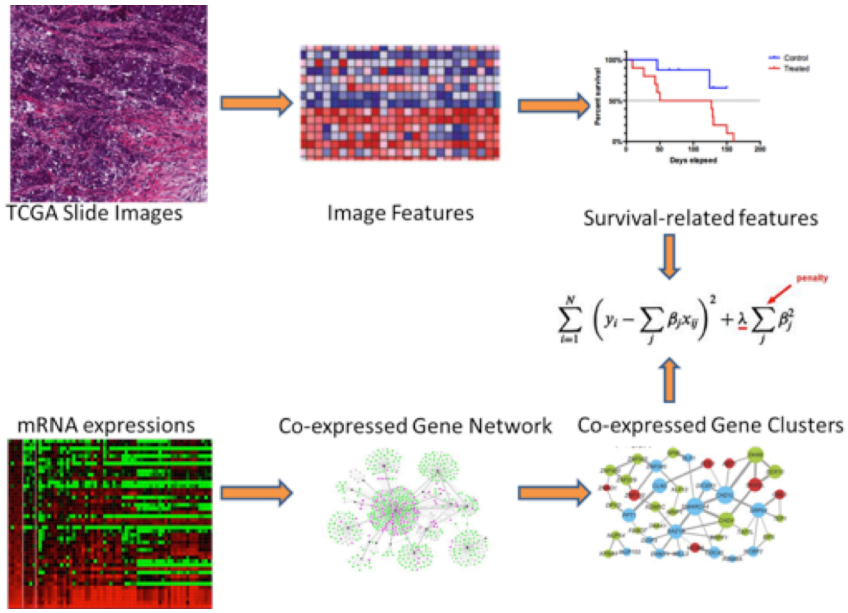

Figure 2: Schematic overview of constructing gene co-expression networks and analyzing the relationships between gene networks and morphological features.

2.1 Integrated Analysis Pipeline

We collected matched diagnostic images and gene expression data for a discovery dataset of 201 LUAD patients from the TCGA. The integrative analysis workflow is shown in Figure 2. Our automatic imaging processing pipeline detected cell nuclei and extracted predefined features evaluating staining variations. To select imaging features with clinical relevance, survival-related imaging features were identified. At molecular level, gene expression profiles (mRNA levels) were filtered and clustered using our co-expression network analysis algorithm. Strongly co-expressed gene clusters were represented by eigengenes. Then, we built a lasso regression model to select gene clusters that regulate the image feature that has the strongest association with survival times. By finding the co-expression patterns that are associated with the selected imaging feature, we can discover biological processes and genetic variations associated with cancer histology.

2.2 Image and Genomic Data Collection

We focus on LUAD patients with clinical information, genomic information, and histopathologic whole slide images. The data were downloaded from TCGA (The Cancer Genome Atlas) Data Portal. Data for 201 LUAD patients with all the three data types are downloaded for the experiments in 2014. For each patient, a representative image patch of size 1712 x 952 without damage or artifact is cropped from the tumor region. Expression profiles of 20,530 unique genes were investigated in the 201 patients (*23*).

2.3 Data Preprocessing and Imaging Feature Extraction

2.3.1 Imaging features

We adopt the cell detection and segmentation methods proposed in (*24*). In the cell

detection stage, a radial voting scheme with Gaussian pyramid is employed (*25*). For each image, a Gaussian pyramid is created. A single–pass voting is applied to each layer. The voting region receives scores weighted by a distance transform. Therefore, such weighted voting encourages the pixels closer to the cell center accumulates higher voting scores. The final voting score is obtained by summing up the voting scores from different layers. In the segmentation stage, a marker based active contour with a repulsive term is applied to the images using the detection results as the markers. An initial contour associated with each detected marker is created first. The contours evolve through an iterative procedure to reach the real boundaries of the cells. The repulsive term serves to prevent the contours from crossing and merging with each other.

Group 1: Geometry Features. Based on the segmentation results, five geometry features are calculated for each lung cancer cell to capture the cell shape information, including cell area, contour perimeter, circularity, major-minor axis ratio, and contour solidity. Contour solidity is defined as the ratio of the area of a cell region over the convex hull defined by the segmentation boundary.

Group 2: Pixel Intensity Statistics. Pixel intensity statistics features are used to capture the color of the segmented cells. This group of features are calculated based on the intensity of the pixels within the segmented cells, including intensity mean, standard deviation, skewness, kurtosis, entropy, and energy. *Lab* color space is used for a better color representation.

Group 3: Texture Features: Texture is an important feature found to be closely related to cancer diagnosis in radiomics. This is rooted in the fact that texture patterns are linked to difference in protein expressions (*26*). This group of features consists of co-occurrence matrix (*27*), center symmetric auto-correlation (CSAC) (*28*), local binary pattern (LBP) (*29*), texture feature coding method (TFCM) (*30*). The co-occurrence matrix (*27*) computes an estimation of the joint probability distribution of the intensity of two neighboring pixels. CSAC is a measure of the local patterns with symmetrical structure. These patterns are characterized by a series of local auto-correlation and covariance introduced in (*28*), including symmetric texture covariance (SCOV), variance (SVR), and within-pair variance (WVAR), and between-pair variance (BVAR). 3×3 pixel unit of each channel is considered. LBP (*29*) feature measures the local textures by assigning a binary code to a pixel with respect to its intensity and those of its neighboring pixels. A histogram of the generated binary codes reveals the distribution of the present repeated local patterns. Similar to LBP, in TFCM (*30*), a texture feature number (TFN) is assigned to each pixel by comparing this pixel with its neighbors in four directions: 0°, 45°, 90°, and 135°. A histogram is calculated based on the TFNs of one image patch.

2.3.2 Gene transcriptome data

The expression profiles of 201 samples with primary lung cancer adenocarcinoma from TCGA LUAD project were downloaded from TCGA data portal in January 2014. Specifically, RNA-seq data for the tumor samples were obtained using Illumina sequencing and processed as described in (*6*). The mapping results were converted to RPKM (read per kilobase per million reads) values for 20,530 genes. Genes with low expression levels (with no data in the top 15 percentile) and low variance (in the lowest 10 percentile) were removed resulting in 9,179 genes.

2.4 Gene co-expression network analysis and summarization

While our goal is to establish the relationships between gene expression levels and the imaging features, we first carry out gene co-expression network analysis (GCNA) to cluster

genes into co-expressed clusters. There are multiple reasons for carrying out GCNA before associating them with the imaging features. First, there is a large number of genes. If the association between every pair of gene and imaging feature is calculated and tested for significance, then more than half a million tests will be carried out which leads to low statistical power. In addition, since we will explore the association beyond univariate relationships using sparse analysis, the large number of genes (which are not always independent), pose serious computing challenges to the sparse modeling algorithms such as Lasso. Thus we first group genes with highly correlated expression profiles into co-expression clusters using GCNA then summarize the expression profiles within each cluster as an "eigengene" using the protocol described in (*31*). Essentially the expression profiles of each gene are first centralized (by subtracting the mean for each gene) and then standardized to have norm one. After the processing steps, singular value decomposition is applied to obtain the *eigengene* as the principal vector in the direction with the largest variance among the samples. Another advantage of the GCNA approach is that the highly co-expressed gene clusters are usually highly enriched in specific biological processes, regulatory factors or structural variations (e.g., copy number variations) (*32*), making the interpretation of the results straightforward.

While there are many algorithms for performing GCNA including the well known WGCNA package, we use a weighted network mining algorithm called local maximum quasi-clique merging (lmQCM) algorithm we recently developed (*32*). Unlike WGCNA which uses hierarchical clustering and does not allow overlaps between clusters, our algorithm is a greedy approach allowing genes to be shared among multiple clusters, in consistent with the fact the genes often participate in multiple biological processes. In addition, we have shown that lmQCM can find smaller co-expressed gene clusters which are often associated structual mutations such as copy number variations in cancers. The lmQCM algorithm has four parameters γ, α, t, and β. Among these parameters, γ is the most influencial, it decides if a new cluster can be initiated by setting the weight threshold for the first edge of the cluster as a subnetwork. In our GCN analysis, we directly use the absolute values of the Spearman correlation coefficients between expression profiles of genes as weights for which we have shown to be effective in previous studies.

2.4 Associations between Morphology and Transcriptomes

2.4.1 Correlation analysis

We first examined the correlation between the imaging features and the eigengenes for the gene clusters identified using lmQCM by calculating the Spearman correlation coefficients between them. However, as shown in the Results, the correlations between imaging features and eigengenes are not strong (none of them is significant if Bonferroni correction is applied for multiple test compensation). While this is different from the case in breast cancer, it suggests that the tissue morphology development is a complicated process involving in multiple processes and genetic factors. Thus in order to explain the morphology development, we need to resort to multi-variate modeling methods such as lasso regression.

2.4.2 Sparse modeling using Lasso regression

We model imaging features as manifestations of gene expression. Given the data availability, we focus on transcriptome data. Lasso regression model minimizes the residual sum of squares while at the same time enforcing sparsity of the model by adding a penalty term of the L_1-norm of the model coefficients.

Consider the linear regression model: we have (x_i, y_i), $i = 1,2, \dots, N$, where $x_i = (x_{i1}, \dots, x_{ip})^T$ and y_i are eigen-gene expression and image feature value for the i th

observation(patient sample), respectively. With regular regression model, the least square estimates are obtained by minimizing the residual squared error. However, in feature selection models to predict biomarkers, only imperative transcriptomes contribute to biological functions and processes, requiring more stringent and interpretable features. With large number of features, we would like to determine a small subset of them that can predict strong correlations. Let $\beta = (\beta_1, \dots, \beta_p)^T$ and β_0 to be a scalar. The lasso model estimate (β, β_0) by solving the following problem

$$\min_{\beta, \beta_0} \left(\frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i{}^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right), \quad (1)$$

where λ is nonnegative regularization parameter giving the weight for the model complexity term. As λ increases, the number of nonzero components of β decreases, leading to smaller numbers of predictors.

2.5 Identification of Survival-Related Image Features

Univariate Cox Proportional Hazard models are used to identify morphological features and genes that have expression related significantly to survival. Morphological features that have p-values less than 0.05 are recorded.

Table 1: Prognostic values of various image features in discover dataset. The features are listed by their significance in the survival model.

Feature Names	p-value		Feature Names	p-value
tfcm4	0.00456904		contrast1	0.01210092
tfcm9	0.00532429		tfcm12	0.01247155
tfcm3	0.00563955		tfcm11	0.01361604
tfcm1	0.0064998		csac23	0.01754474
tfcm2	0.00657692		tfcm7	0.0178572
tfcm10	0.00685436		fourier15	0.0178766
contrast2	0.0082282		csac5	0.01896244
tfcm8	0.0093341		entry4	0.01995154

2.6 Validating the Identified Genes in Other Data Set

For genes associated with the imaging features with the highest potential for predicting patient survival, we also test them on another publicly available dataset obtained from the NCBI Gene Expression Omnibus. The dataset GSExxxx contains transcriptome data of 149 non-small cell lung cancer patients, among which 90 are unique lung adenocarcinoma patients with clinical outcome (survival time and status). We use the genes to be tested as features to separate the 90 patients into two groups using K-means algorithms (K=2, Euclidean distance, average linkage, and 10 replicates). The survival times of the two groups are then visualized using Kaplan-Meier curves and compared using Cox Proportional Hazard regression.

2.7 Enrichment analysis of gene clusters

To interpret the biological meaning of the identified gene clusters, enrichment analysis tools such as TOPPGene (https://toppgene.cchmc.org/enrichment.jsp) are used. In addition, information about the genes are extracted from cBioPortal (http://www.cbioportal.org/).

3 Results

3.1 Image Feature Calculation

As shown in Figure 3 Left, the images reveal clear heterogeneity of the tumors among the patients. We calculated 283 image features from the images. As described in Section 2.3.1, there are multiple types of features and many features are strongly correlated (Figure 3 Right) such as part of the TCFM family (the block of 211 to 222). In this paper, we analyze

each feature individually, but some of the highly features can be combined in future analysis.

3.2 Survival-related Image Features and Gene Cluster

Using a univariate Cox proportional hazards regression model, we assessed the image features related risk score in the prediction of the LUAD patient survival. Significant

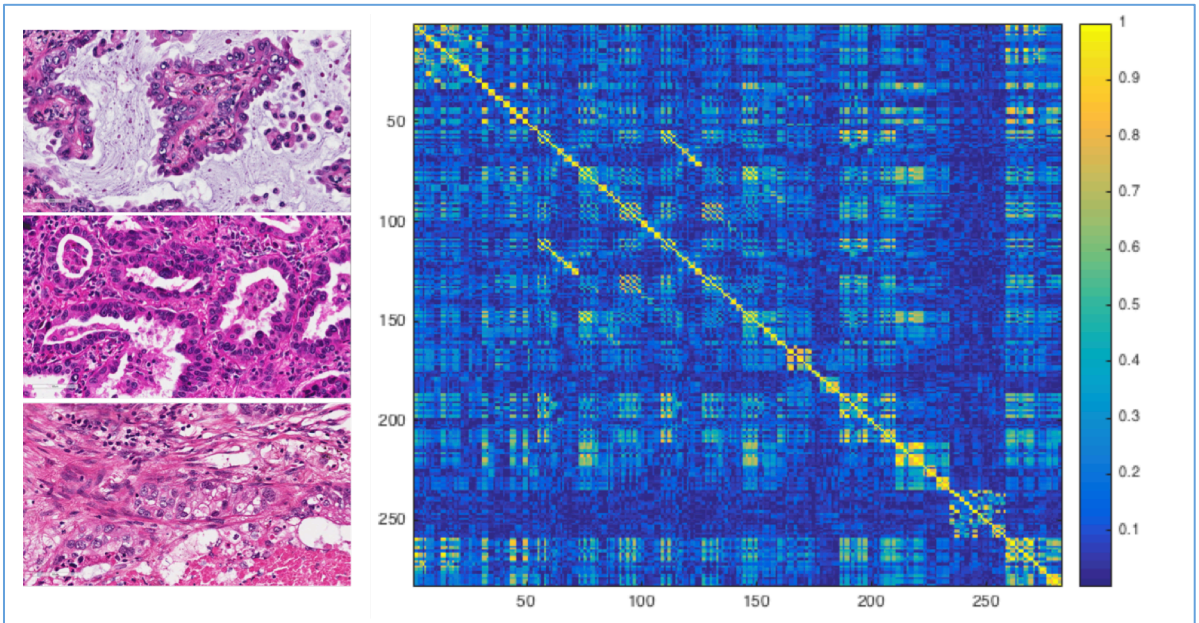

Figure 3. **Left**: Examples of the image patches from different patients. **Right**: Heatmap of the correlation (Spearman) matrix for the 283 image features.

morphological features are listed in Table 1. Among the six categories of imaging features, the *tfcm* category shows the most significant prognostic power, indicating texture features in lung adenocarcinoma have a strong potential for predicting patients' outcomes. In fact, all of the top six survival-related imaging features are in the *tfcm* category. Other features that capture prognosis are *contrast2*, *contrast*, *csac23*, *fourier15*, *csac5* and *entry4*.

3.2 Gene Co-Expression Network Analysis

As mentioned in Section 2.3.2, *9,179* genes were kept for analysis. The absolute value of the Spearman rank correlation coefficients were used for cluster detection using lmQCM algorithm. We allow the smallest gene clusters to have five genes. Then we found with γ = 0.75, t = 1, α = 1, and β = 0.4 the algoirthm yielded co-expressed gene clusters with balanced sizes. Specifically, it led to 95 clusters ranging from 5 to 120 genes. Many of the gene clusters are consistent to the ones frequently found in cancers. Most of these clusters involved in hallmark cancer biological processes such as cell cycle/genome stability (cluster 1), immune respones (cluster 2), translation / protein synthesis (cluster 3), and extracellular matrix development (cluster 7). However, some of them are more associated with specific cytobands (e.g., chr19p13), implying potential CNV sites.

3.3 Correlations between Image Features and Gene Clusters

The image analysis pipeline allowed us to quantify tumor characteristics on cellular level and associate these tumor characteristics with patient outcomes. In this study, we calculated 283 imaging features for the 201 patients and correlated with the 95 eigengenes. The correlation coefficient with the large absolute value is -0.2990 (p=1.7728e-05). In Table 2, we list the strongest correlation between eigengenes and the top five imaging features (in

Table 1) with the most significant power for predicting patient outcome. It is clear from the table that none of such correlations is statistically significant (after multiple test compensation), suggesting that complex phenomena such as cell and tissue morphology in lung cancers can only be explained by multiple molecular and genetic factors.

Table 2. Imaging features and the eigengenes with the strongest correlations with them.

Imaging feature	Eigengene (cluster)	SCC/p-value	Enrichment
tfcm4	95	0.1710/0.0153	18q12.1 (p=1.175e-9), all five genes on 18q12
tfcm9	59	0.1677/0.0174	16p11.2 (p=1.364e-10), all seven genes on 16p11
tfcm3	59	-0.1658/0.0188	
tfcm1	59	0.1704/0.0157	
tfcm2	59	0.1508/0.0327	

3.4 Lasso Regression Model for Imaging Features Using Eigengenes

Since the imaging features with prognostic power do not have strongly correlated gene clusters, we resort to multivariate models to explain the cell and tissue morphology using molecular data. Specifically, we built a lasso regression model. The lasso model selects a sparse set of eigengenes to explain the selected imaging feature. We rank the importance of image features by their significance in survival analysis. The top 10 image features in Table 1 belong to only two categories – TFCM and Contrast. Features within each category are highly correlated (for the eight TFCM features, the smallest of the absolute value of the SCC is 0.6840, the two SCC between the two Contrast features is 0.9923). Since eight out of 10 top image features are from the TFCM family, we chose one feature from for our modeling, namely *tfcm4*.

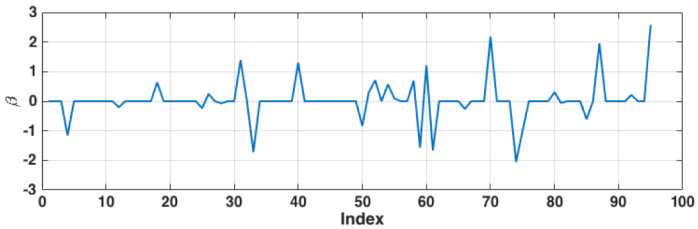

Figure 4. Coefficients (β values) of the lasso regression for *tfcm4*.

For *tfcm4*, it is found that the lowest MSE is found at λ = 0.0371 for the cost function in Eq.(1). Figure 4 shows the values of the coefficients β. Among the 95 eigengenes, 28 have non-zero coefficients among which 18 are larger than 0.5 and 12 are larger than 1. For the analysis of genes, we collected 185 genes from the 18 clusters with absolute value of coefficients larger than 0.5. In addition, Figure 5 shows the correlation between the combined eigengenes using the calculated β values with the *tfcm4* values in contrast to the correlation between the 95th eigengene (as listed in Table 2) and *tfcm4* (Figure 1).

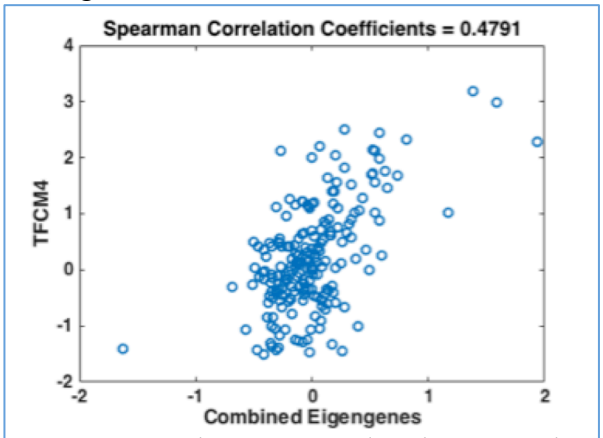

Figure 5. The scatter plot between the combined eigengenes using the lasso coefficients and the *tfcm4* (SCC = 0.4791).

3.5 Functional and Genetic Analysis of Gene Clusters Associated with Imaging Features

In order to understand the functional roles of the gene clusters associated with *tfcm4*, enrichment analysis was carried out

using TOPPGene and the results for the 18 gene clusters are shown in Table 2. Among the gene clusters whose eigengenes are associated with tfcm4, the largest cluster is the cluster #4, consisting of 59 genes and is highly enriched with ribosomal genes and thus protein translation function. Other related biological processes including immune response (response to virus, cluster #18), response to steroid hormone, negative regulation of epithelial cell proliferation, and mitochondrial ATP synthesis.

Interestingly, 14 out of the 18 gene clusters are highly enriched on specific cytobands. It has been previously noticed that many of the co-expressed clusters in cancers are associated with copy number variations (CNVs) in specific cytobands (*32*). CNVs are common genetic variations playing important roles cancer initiation and development. Functional CNVs usually lead to changes in expression levels of genes on that region due to the "dose effect", which also leads to co-expression of the transcribed genes. Figure 6 Left shows an example of the RPRD1A gene in cluster #95, whose mRNA level has a strong correlation with its copy number measurement and it shows a strong co-expression relationship with the ELP2 genes on the same cytoband.

Table 2: Gene clusters showing strong correlation with texture image feature tfcm, and their Gene Ontology terms and enriched cytobands.

Gene Cluster (size)	beta	GO Biological Process/p-values	Cytobands/p-values	Notes:
4 (59)	-1.1558	GO:0006614 SRP-dependent cotranslational protein targeting to membrane / 9.105E-98		
18 (14)	0.6328	GO:0009615response to virus/ 9.965E-15		
31 (10)	1.3894			Genes down-regulated in nsopharyngeal carcinoma relative to the normal tissue (p = 5.074e-19, all 10 genes)
33 (10)	-1.7213		19q13.42/5.525e-6	All 10 genes on 19q13.3-4
40 (8)	1.2977		8q24.13/3.263e-5	Seven genes on 8q21-24, one on 8q13
50 (8)	-0.8343	GO:0048545response to steroid hormone / 2.290E-8		
52 (8)	0.7075		7q33/4.800E-5	All eight genes on 7q21-36
54 (7)	0.5669	GO:0006413 translational initiation /1.096E-5	Yq11/2.305E-6, Xq13.2/2.856E-5	Four genes on Yq11, two on Xq13.2, one on Yp11.3
58 (7)	0.6952		8p21.1/ 6.631E-6	Five genes on 8p21, two on 8p12
59 (7)	-1.5729		16p11.2/1.364e-10	All seven genes on 16p11
60 (7)	1.2103		Xq28/1.982e-13	All seven genes on Xq27-28
61 (7)	-1.6639		6p21.1/4.436e-7	Six genes on 6p21-22, one on 6p12
70 (6)	2.1783		17q21.31/5.532e-	All six genes are on 17q21

			10	
74 (6)	-2.0544		8p11.2/1.048e-9	All six genes are on 8p11.2
75 (6)	-1.0093	GO:0050680 negative regulation of epithelial cell proliferation / 3.290E-6	17q11.2/6.880e-7	All six genes are on 17q11-12
85 (5)	-0.6095	GO:0042776 mitochondrial ATP synthesis coupled proton transport / 6.311E-9	21q22.11/3.344E-5	Four genes on 21q21-22
87 (5)	1.9569		19q13.2/1.131e-6	All five genes on 19q13
95 (5)	2.5783		18q12.1/1.175e-9	All five genes on 18q12

3.4 Prognostic Validation

Validation on heterogeneous external data sets allows for evaluation of the generalizability. To test the importance of cilium-related genes, we further performed survival analysis on a publicly available dataset with 90 LUAD patients. Among the 185 genes correlated with the image feature category tfcm, 118 of the gene symbols can be matched exactly to the external dataset. In the validation dataset, lung adenocarcinoma patients were stratified into two groups using K-means based on their expression levels of the 118 genes. In both datasets, a statistically significant group of patients with worse outcomes were differentiated (n = 44 and n = 46, respectively). The difference between the two groups is significant (Cox hazard proportional model p-value 2.1285e-6). Figure 7 shows the Kaplan-Meier curves of the two patient cohorts.

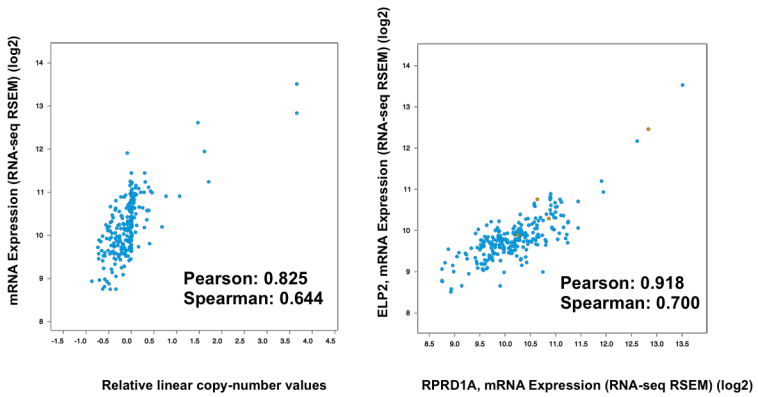

Figure 6. **Left**: Gene RPRD1A shows strong correlation between mRNA level and copy number values. **Right**: RPRD1A and ELP2 (both in cluster #95) are strongly co-expressed.

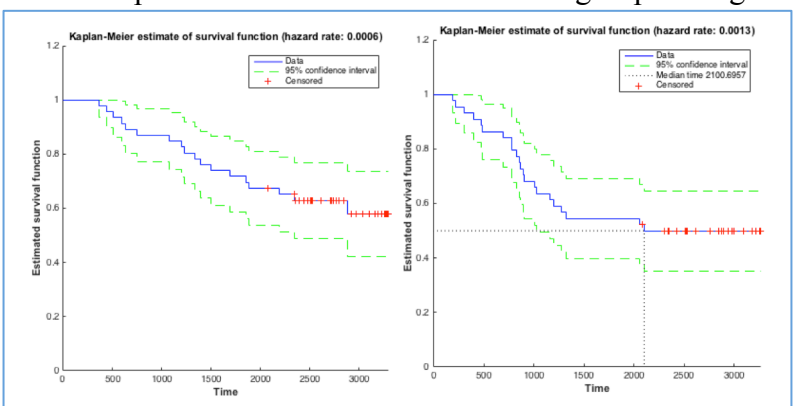

Figure 7: Kaplan–Meier survival curves of prognostic model on two clusters from the lung adenocarcinoma dataset. Survival curves of two groups (group 1 with 44 patients on the left and group 2 with 46 patients on the right). The difference of survival time between the two groups is significant (Cox hazard proportional model p-value 2.1285e-6).

4 Discussion and Conclusion

Our integrative analysis

pipeline allows us to find survival related textural features of lung adenocarcinoma. In addition to the image features, we also demonstrated that modeling of the histology at cellular and tissue levels using "omics" data may involve multiple groups of genes. Interestingly, our results showed that that the histological phenotype may be manifestations of multiple genetic variations, especially copy number variations. Specifically, many of the enriched cytobands we identified have been previously associated with lung cancer development including 19q13 (*33*, *34*) , 8q24 (*33*), 7q21-36 (*35*), 8p21 (*33*), 16p11 (*36*), Xq27-28, 6p21 (*34*), 17q21 (*34*), 21q22 (*35*), and 18q12 (*33*). While there is no report on the association of Xq27-28 with lung cancer, Xq26 has been shown to be associated with lung cancers (*36*), suggesting that the genetic variations should be further explored to identify potential "driver" genes for lung cancer. We also showed that the genes in the clusters can indeed predict patient prognosis, which leads to discovery of potential biomarkers. While our study is focused on patient prognosis, the process can be repeated for patient treatment response prediction with appropriate data. Overall we demonstrated that the morphology is a complex phenomenon and its development may involve multiple groups of genes. In cancers, this process is even more complex as the genetic variations also contribute significantly to this process. Our findings indeed support this notion.

References

1. S. Couraud, G. Zalcman, B. Milleron, F. Morin, P.-J. Souquet, *Eur. J. Cancer*. **48**, 1299–311 (2012).
2. P. A. Russell *et al.*, *J. Thorac. Oncol.* **6**, 1496–504 (2011).
3. E. Conde *et al.*, *Clin. Transl. Oncol.* **15**, 503–8 (2013).
4. D. Hokka *et al.*, *Lung Cancer*. **79**, 77–82 (2013).
5. X. Li *et al.*, *Neoplasma*. **59**, 500–7 (2012).
6. E. A. Collisson *et al.*, *Nature*. **511**, 543–50 (2014).
7. C. Camps, Jantus-Lewintre, Usó, Sanmartin, *Lung Cancer Targets Ther.*, 21 (2012).
8. P. F. Robert T. Adamson, *Am. J. Manag. Care*. **19** (2013).
9. J. Chmielecki *et al.*, *J. Thorac. Oncol.* **7**, 434–42 (2012).
10. J. H. M. Austin *et al.*, *Radiology*. **266**, 62–71 (2013).
11. L. M. Solis *et al.*, *Cancer*. **118**, 2889–99 (2012).
12. Y. Yuan *et al.*, *Sci. Transl. Med.* **4**, 157ra143–157ra143 (2012).
13. a. H. Beck *et al.*, *Sci. Transl. Med.* **3**, 108ra113–108ra113 (2011).
14. H. Wang, F. Xing, H. Su, A. Stromberg, L. Yang, *BMC Bioinformatics*. **15**, 1–12 (2014).
15. F. B. Thunnissen *et al.*, *Pathol. Res. Pract.* **188**, 531–5 (1992).
16. S. L. Edwards *et al.*, *J. Clin. Pathol.* **53**, 537–40 (2000).
17. K. Kayser, D. Radziszowski, P. Bzdyl, R. Sommer, G. Kayser, *Rom. J. Morphol. Embryol.* **47**, 21–8 (2006).
18. M. Mijovic, Zaklina; Mihailovic, Dragan; Kostov, *Med. Biol.* **15**, 28 – 32 (2008).
19. X. Zhang, L. Yang, W. Liu, H. Su, S. Zhang, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014* (2014), pp. 479–486.
20. J. Yao *et al.*, in *Proceedings of the 6th International Workshop on Machine Learning in Medical Imaging - Volume 9352* (Springer-Verlag New York, Inc., 2015; http://link.springer.com/10.1007/978-3-319-24888-2_35), pp. 288–295.
21. X. Zhu *et al.*, in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (IEEE, 2016; http://ieeexplore.ieee.org/document/7493475/), pp. 1173–1176.
22. C. Wang *et al.*, *J. Am. Med. Inform. Assoc.* **20**, 680–7.

23. TCGA, The Cancer Genome Atlas - Data Portal.
24. F. Xing, L. Yang, in *2013 IEEE 10th International Symposium on Biomedical Imaging* (IEEE, 2013), pp. 386–389.
25. X. Qi, F. Xing, D. J. Foran, L. Yang, *IEEE Trans. Biomed. Eng.* **59**, 754–65 (2012).
26. P. Lambin *et al.*, *Eur. J. Cancer*. **48**, 441–6 (2012).
27. R. M. Haralick, K. Shanmugam, I. Dinstein, *IEEE Trans. Syst. Man. Cybern.* **3**, 610–621 (1973).
28. K. Laws, in *24th Annual Technical Symposium*, T. F. Wiener, Ed. (International Society for Optics and Photonics, 1980), pp. 376–381.
29. T. Ojala, M. Pietikainen, T. Maenpaa, *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 971–987 (2002).
30. M.-H. Horng, Y.-N. Sun, X.-Z. Lin, *Comput. Med. Imaging Graph.* **26**, 33–42 (2002).
31. P. Langfelder, S. Horvath, *BMC Bioinformatics*. **9**, 559 (2008).
32. J. Zhang, K. Huang, *Cancer Inform.* **1**, 1 (2016).
33. B. R. Balsara *et al.*, *Cancer Res.* **57**, 2116–20 (1997).
34. P. P. Medina *et al.*, *Hum. Mol. Genet.* **18**, 1343–52 (2009).
35. F. Li, L. Sun, S. Zhang, *Oncol. Rep.* **34**, 1701–7 (2015).
36. N. A. Levin *et al.*, *Cancer Res.* **54**, 5086–91 (1994).

IDENTIFICATION OF DISCRIMINATIVE IMAGING PROTEOMICS ASSOCIATIONS IN ALZHEIMER'S DISEASE VIA A NOVEL SPARSE CORRELATION MODEL

JINGWEN YAN*

Department of BioHealth Informatics, Indiana University,
Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University
Indianapolis, 46202, USA
E-mail: jingyan@iupui.edu

SHANNON L. RISACHER, KWANGSIK NHO, ANDREW J. SAYKIN

Department of Radiology and Imaging Sciences, School of Medicine, Indiana University,
Indianapolis, 46202, USA
E-mail:{srisache,knho,asaykin}@iupui.edu

LI SHEN*

Department of Radiology and Imaging Sciences, School of Medicine, Indiana University,
Center for Computational Biology and Bioinformatics, School of Medicine, Indiana University
Indianapolis, 46202, USA
E-mail:shenli@iu.edu

FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE†

Brain imaging and protein expression, from both cerebrospinal fluid and blood plasma, have been found to provide complementary information in predicting the clinical outcomes of Alzheimer's disease (AD). But the underlying associations that contribute to such a complementary relationship have not been previously studied yet. In this work, we will perform an imaging proteomics association analysis to explore how they are related with each other. While traditional association models, such as Sparse Canonical Correlation Analysis (SCCA), can not guarantee the selection of only disease-relevant biomarkers and associations, we propose a novel discriminative SCCA (denoted as DSCCA) model with new penalty terms to account for the disease status information. Given brain imaging, proteomic and diagnostic data, the proposed model can perform a joint association and multi-class discrimination analysis, such that we can not only identify disease-relevant multimodal biomarkers, but also reveal strong associations between them. Based on a real imaging proteomic data set, the empirical results show that DSCCA and traditional SCCA have comparable association performances. But in a further classification analysis, canonical variables of imaging and proteomic data obtained in DSCCA demonstrate much more discrimination power toward multiple pairs of diagnosis groups than those obtained in SCCA.

Keywords: Imaging genomics; Alzheimer's disease; Proteomics; Canonical correlation analysis; Multi-class discrimination.

*To whom correspondence should be addressed

†Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

1. Introduction

Alzheimer's disease (AD) has been well known as one of the most common brain dementia, a major neurodegenerative disorder that has been characterized by gradual memory loss and brain behavior impairment. According to the latest report,[1] more than 5 million Americans are living with Alzheimer's and it has been officially listed as the 6th leading cause of death. Also, due to the significant decline of self-care capabilities during disease, it is not only the patients who suffer, but also the family members, friends, communities and the whole society considering the time-consuming daily care and high health care expenditures needed. In the past decade, deaths attributed to Alzheimer's disease has increased 68 percent, while deaths attributed to the number one cause, heart disease, has decreased 16 percent. And all of these situations will continue to deteriorate as the population ages during the next several decades. To prevent such health care crisis, substantial efforts have been made to help cure, slow or stop the progression of the disease.

In the last few years, many efforts have been dedicated to explore whether the combination of multi-modal measures, e.g. brain atrophy measured by magnetic resonance imaging (MRI), hypometabolism measured by functional imaging and quantification of proteins, can better predict the clinical outcomes of AD, such as disease status and cognitive outcomes.[19] In many of these works, it has been found that brain imaging and protein expression, from both cerebrospinal fluid (CSF) and blood plasma, hold some complementary information.[12,18] But how they are related with each other still remains elusive.

In this work, we will explore the relationships between brain imaging and protein expression using bi-multivariate association models. Sparse Canonical Correlation Analysis (SCCA)[11,16] is a typical example that has been widely used for associative analysis in both real[8,15] and simulated[3] -omics data sets.[2,11,17] But it can not guarantee the selection of disease-relevant biomarkers and therefore the associations generated in SCCA are not necessarily related to a specific disease either, unless the input features are already prefiltered disease-related biomarkers.[5] On the other hand, most existing SCCA algorithms use the soft threshold strategy for solving the Lasso[11,16] regularization terms, which assumes the independence structure of data features. Unfortunately, this independence assumption does not hold in neither imaging nor proteomics data, and will inevitably limit the capability of yielding optimal solutions.

To overcome these limitations, we propose a novel discriminative SCCA (DSCCA) model, coupled with a new algorithm to eliminate the independence assumption, to explore the imaging and proteomic associations. Given imaging, proteomic and diagnostic data, the proposed model can perform a joint association and multi-class discrimination analysis. As such, we can not only identify disease-relevant multimodal biomarkers, but also reveal strong association between them. We perform an empirical comparison between the proposed DSCCA algorithm and a widely used SCCA implementation in the PMA software package (http://cran.r-project.org/web/packages/PMA/).[16] The results show that DSCCA and SCCA have comparable association performances. But in a further classification analysis, canonical variables of imaging and proteomic data obtained in DSCCA demonstrate much more discrimination power toward diagnosis groups than those obtained in SCCA.

2. Discriminative SCCA (DSCCA)

Throughout this section, we denote vectors as boldface lowercase letters and matrices as boldface uppercase ones. For a given matrix $\mathbf{M} = (m_{ij})$, we denote its i-th row and j-th column to $\mathbf{m}^i$ and $\mathbf{m}_j$ respectively. Let $\mathbf{X} = \{\mathrm{x}_1, ..., \mathrm{x}_n\} \subseteq \mathfrak{R}^p$ be the imaging data and $\mathbf{Y} = \{\mathrm{y}_1, ..., \mathrm{y}_n\} \subseteq \mathfrak{R}^q$ be the protein data, where n is the number of participants, p and q are the number of brain regions and proteins respectively.

Canonical correlation analysis (CCA) is a bi-multivariate method that explores the linear transformations of variables $\mathbf{X}$ and $\mathbf{Y}$ to achieve the maximal correlation between $\mathbf{Xu}$ and $\mathbf{Yv}$, which can be formulated as:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} \qquad s.t. \;\; \mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} = 1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v} = 1 \tag{1}$$

where $\mathbf{u}$ and $\mathbf{v}$ are canonical loadings or weights, reflecting the significance of each feature in identified associations.

However, the power of CCA in biomedical applications is quite limited due to 1) its requirement on the relatively large number of observations n which is expected to exceed the combined dimension of $\mathbf{X}$ and $\mathbf{Y}$, and 2) its nonsparse outputs $\mathbf{u}$ and $\mathbf{v}$ which make the ultimate pattern hard to interpret. To address this concerns, sparse CCA (SCCA) method was later proposed, where two penalty terms on both weight vectors $P_1(\mathbf{u}) \leq c_1$ and $P_2(\mathbf{v}) \leq c_2$ were introduced to help generate sparse results.

A widely used SCCA implementation, PMA package,[16] applied L_1 norm penalty for both P_1 and P_2. But without diagnosis information, its capability in identifying disease-relevant biomarkers is quite limited. Thus the ultimate association relationships are not necessarily related to a specific disease either. Another limitation of PMA is that it takes the soft threshold strategy in the solution, which requires the input data to have an linear independence design $\mathbf{X^TX = I}$ and $\mathbf{Y^TY = I}$ (see Section 10 in[14]). Unfortunately, this independence assumption does not hold in both imaging and proteomics data (e.g., correlated voxels in an ROI, correlated protein expressions), and will inevitably limit the capability of identifying meaningful imaging proteomics associations.

To overcome these limitations, we propose a novel discriminative SCCA (denoted as DSCCA) algorithm to not only take into account the diagnosis information but also eliminate the independence assumption. Inspired by the application of locality preserving projection (LPP) in linear discriminative analysis,[10] we add two new constraints as P_1 and P_2 for multi-class discrimination.

$$\begin{aligned} P_1(\mathbf{u}) &= ||\mathbf{u}||_D = \alpha\mathbf{u}^T\mathbf{X}^T\mathbf{L}_w\mathbf{X}\mathbf{u} - (1-\alpha)\mathbf{u}^T\mathbf{X}^T\mathbf{L}_b\mathbf{X}\mathbf{u}, \\ P_2(\mathbf{v}) &= ||\mathbf{v}||_D = \alpha\mathbf{v}^T\mathbf{Y}^T\mathbf{L}_w\mathbf{Y}\mathbf{v} - (1-\alpha)\mathbf{v}^T\mathbf{Y}^T\mathbf{L}_b\mathbf{Y}\mathbf{v}, \end{aligned} \tag{2}$$

Here, we construct two graphs $\mathbf{G}_w$ and $\mathbf{G}_b$ to account for the diagnosis groups, where each vertex indicates one subject (Fig. 1). In $\mathbf{G}_w$, only subjects within the same diagnosis group have connections to each other. In other words, we build a complete graph for all the subjects belonging to the same diagnosis group. In $\mathbf{G}_b$, only subjects from different diagnosis

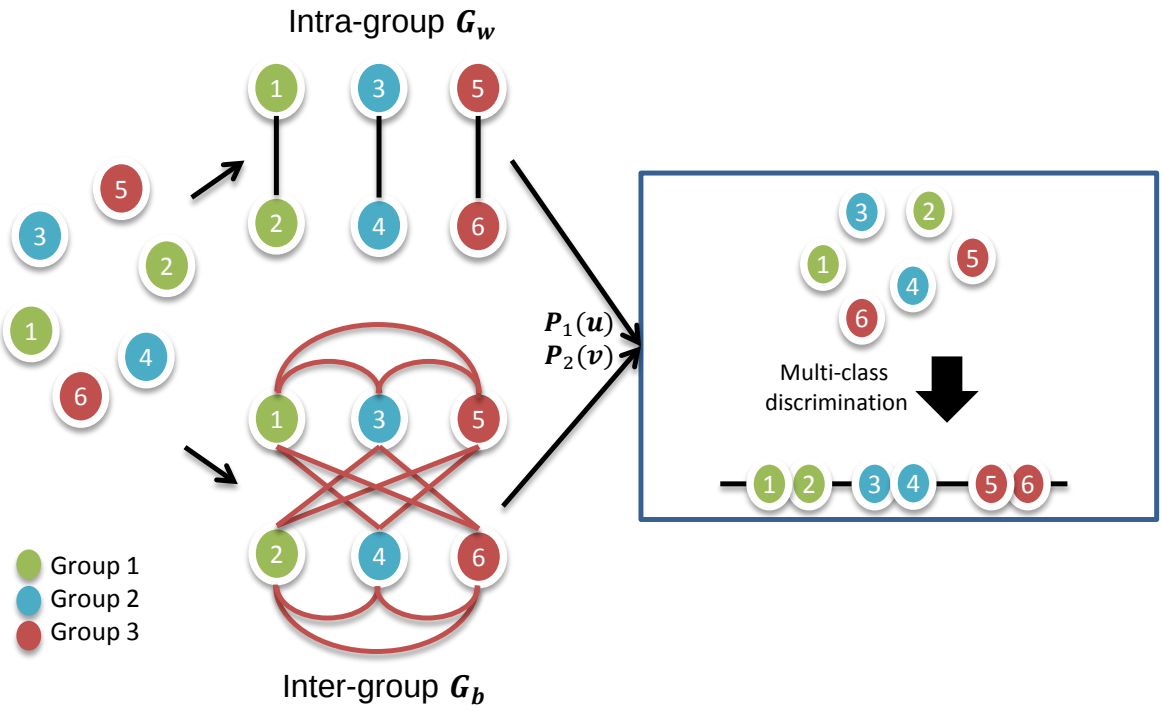

Fig. 1. Illustration of within- and between-group graphs $\mathbf{G}_w$ and $\mathbf{G}_b$. Each circle indicates one subject and subjects from the same diagnosis group are colored the same.

groups have connections. $\mathbf{L}_w$ and $\mathbf{L}_b$ are the Laplacian graphs of $\mathbf{G}_w$ and $\mathbf{G}_b$ respectively. While the traiditonal L_1 norm helps ascertain the sparsity of selected imaging and protein biomarkers, the new penalty term $||\cdot||_D$ encourages the closeness between subjects within the same diagnosis groups and distance between subjects from different diagnosis groups after projection. α is a trade off parameter that help balance the within- and between-group constraints. Since canonical variables $\mathbf{Xu}$ and $\mathbf{Yv}$ have the exact same length, we use the same α for both penalties P_1 and P_2.

The final objective function of DSCCA can be written as follows:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \frac{\beta_1}{2}P_1(\mathbf{u}) - \frac{\beta_2}{2}P_2(\mathbf{v}) \tag{3}$$

$$s.t.\ \ \mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} = 1, \mathbf{v}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v} = 1, ||\mathbf{u}||_1 \leq c_1, ||\mathbf{v}||_1 \leq c_2$$

Using Lagrange multipliers, Eq. (3) can be reformulated as follows:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} - \frac{\gamma_1}{2}||\mathbf{X}\mathbf{u}||_2^2 - \frac{\gamma_2}{2}||\mathbf{Y}\mathbf{v}||_2^2 - \frac{\beta_1}{2}P_1(\mathbf{u}) - \frac{\beta_2}{2}P_2(\mathbf{v}) - \lambda_1||\mathbf{u}||_1 - \lambda_2||\mathbf{v}||_1 \tag{4}$$

Eq. (4) is known as a bi-convex problem, which can be easily solved using an alternating algorithm as discussed in.[16] By fixing $\mathbf{u}$ and $\mathbf{v}$ respectively, we will have the following two minimization problems shown in Eq. (5) and (6).

$$\min_{\mathbf{u}} -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} + \frac{\gamma_1}{2}\mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} + \frac{\beta_1}{2}P_1(\mathbf{u}) + \lambda_1||\mathbf{u}||_1, \quad (5)$$

$$\min_{\mathbf{v}} -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} + \frac{\gamma_2}{2}\mathbf{v}^T\mathbf{Y}^T\mathbf{Y}\mathbf{v} + \frac{\beta_2}{2}P_2(\mathbf{v}) + \lambda_2||\mathbf{v}||_1, \quad (6)$$

Both objective functions can be efficiently solved using the Nesterovs accelerated proximal gradient optimization algorithm.[9] Algorithm 2.1 summarizes the optimization procedure. The convergence is based on the value changes of the objective function and we use 10^{-6} as stop criteria. Five-fold nested cross-validation was applied to automatically tune the parameters β_1, β_2, λ_1 and λ_2. According to,[2] the learned pattern and performance are insensitive to γ_1 and γ_2 settings. Therefore in this paper we set both of them to 1 for simplicity. The optimization method used in steps 3 and 4 is similar to that proposed in.[9]

Algorithm 2.1 Discriminative SCCA (DSCCA)

Require:

$\mathbf{X} = \{\mathrm{x}_1, ..., \mathrm{x}_n\}$, $\mathbf{Y} = \{\mathrm{y}_1, ..., \mathrm{y}_n\}$, $\mathbf{L}_w \subseteq \Re^{n\times n}$, $\mathbf{L}_b \subseteq \Re^{n\times n}$

Ensure:

Canonical vectors $\mathbf{u}$ and $\mathbf{v}$.

1: $t = 1$, Initialize $\mathbf{u}_t \in \Re^{p\times 1}$, $\mathbf{v}_t \in \Re^{q\times 1}$;
2: **while** not converge **do**
3: Solve Eq. (5) using Nesterov's method and obtain $\mathbf{u}$;
4: Solve Eq. (6) using Nesterov's method and obtain $\mathbf{v}$;
5: Scale $\mathbf{u}$ so that $\mathbf{u}^T\mathbf{u} = 1$
6: Scale $\mathbf{v}$ so that $\mathbf{v}^T\mathbf{v} = 1$
7: $t = t + 1$.
8: **end while**

3. Results

3.1. *Data and Experimental Setting*

The MRI data, quantification of proteins in CSF and blood plasma were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see adni.loni.usc.edu.

We totally extracted 246 subjects with all MRI, CSF and plasma proteomic data available. To balance the diagnostic groups, we randomly removed some mild cognitive impairment (MCI) participants. Finally, 176 subjects (67 AD, 67 MCI and 42 healthy control (HC)), were included in this study (Table 1). For each baseline MRI scan, FreeSurfer (FS) V4 was employed to extract 73 cortical thickness measures and 26 volume measures, as well as to extract the intracranial volume (ICV). CSF and blood plasma samples were evaluated by Rules Based Medicine, Inc. (RBM) proteomic panel and 229 proteomic analytes survived the

quality control process, with 83 from CSF and 146 from plasma. Using the regression weights from HC participants, all the MRI, CSF and blood plasma proteomic measures were pre-adjusted for the baseline age, gender, education, and handedness, with ICV as an additional covariate for MRI only.

Table 1. Participant characteristics

	HC	MCI	AD
Number	67	67	42
Gender(M/F)	38/29	45/22	22/20
Handedness(R/L)	64/3	64/3	38/4
Age(mean±std)	75.15±7.68	74.28±7.25	75.93±5.82
Education(mean±std)	15.12±3.01	15.96±2.92	15.88±2.77

3.2. *Experimental Results*

Both DSCCA and PMA were performed on the normalized FS and proteomic measures. To avoid the over-fitting problem, 5-fold nested cross-validation was applied, which also helped to optimally tune the parameters. Table 2 shows 5-fold cross-validation canonical correlation results. It is observed that proposed DSCCA and PMA have comparable performances in identifying imaging proteomic associations, whereas DSCCA is slightly better in performance stability.

Next, we examined the discriminative power of canonical variables **Xu** and **Yv** generated by DSCCA and PMA. Area under ROC curve (AUC) was calculated for each single canonical variable of five folds. Both imaging and proteomic canonical variables of PMA and imaging canonical variable of DSCCA were found to have little discrimination power in all HC vs MCI, HC vs AD and MCI vs AD cases. Proteomic canonical variable **Yv** of DSCCA has the best performance, with an averaged AUC around 0.7 for all three cases. Shown in Fig. 2 is an example plot of **Xu** against **Yv** in one fold. Dot colors represent different diagnostic groups. Compared to one single canonical variable, we observe that combination of two canonical variables generated in DSCCA demonstrated much more discrimination power than PMA. In Fig. 2(a) three diagnosis groups are all very well separated, whereas in Fig. 2(b) subjects are mixing together.

To further validate our results, a follow up classification analysis was performed using both imaging and proteomic canonical variables as predictors. Canonical loadings learned in the training data set are applied to both training and test data to calculate the training and test canonical variables respectively. The LIBSVM toolbox was employed to implement the SVM using a linear kernel under default settings. Three pair-wise binary classification analyses were performed between HC vs MCI, HC vs AD, and MCI vs AD respectively. Shown in Table. 3 are the classification performance comparison between DSCCA and PMA. The results are very encouraging. Canonical variables of DSCCA significantly outperformed those of PMA in terms of the overall accuracy in almost all the cases. The resulting best prediction rates for HC vs AD (92.1%), HC vs MCI (75.3%) and MCI vs AD (70.3%) were competitive with prior

multi-modal studies,[6,19] especially considering that it is under default parameter settings.

All five-fold experiments generated similar sparse results in terms of selection of imaging and proteomic markers. Fig. 3 shows the imaging and proteomic markers commonly identified across all folds using DSCCA, where the color represents the weights of corresponding brain regions. Top brain regions identified include entorhinal cortex, amygdala volume, hippocampal volume, etc. (Fig. 3(a)), which are all aligned with previous AD findings.[12,19] In terms of proteomic markers, expression levels of 12 proteins from CSF and 19 proteins from blood plasma were found to be strongly associated with those brain regions. According to the STRING database (http://string-db.org/), these proteins are highly interconnected with each other, as shown in Fig. 3(b). Edges are colored based on the evidence of the connection, such as experimental interaction, co-expression or co-occurrence in the literature. The more edges two proteins have, the more confident their connection will be.

In particular, four proteins, apolipoprotein E (*APOE*), AXL receptor tyrosine kinase(*AXL*), interleukin 6 receptor (*IL6R*) and vascular endothelial growth factor (*VEGF*), were identified in both CSF and blood plasma. *APOE* is the top risk gene of AD. *AXL* is a member of the Tyro3-Axl-Mer (TAM) receptor tyrosine kinase subfamily, which has been previously reported to be involved in Amyloidogenic APP Processing and β-Amyloid Deposition in AD.[20] For growth factor VEGF, both its variants and expression changes are found to be associated with AD.[4,13] *IL6R* is less explored in terms of its relationship with dementia. But in a recent study it was reported to have significant associations with proteins involved in amyloid processing and inflammation.[7] These findings suggest the existence of certain connections between brain and blood biomarkers. Thus, more accessible fluid biomarkers from blood should have potential to provide extra insights of AD and guidance for future therapeutic intervention activities.

Table 2. Five-fold cross validation canonical correlation results

		f1	f2	f3	f4	f5	mean
DSCCA	Train	0.796	0.670	0.820	0.680	0.636	0.720
	Test	0.424	0.476	0.281	0.392	0.312	0.377
PMA	Train	0.529	0.629	0.505	0.524	0.504	0.538
	Test	0.410	0.095	0.324	0.201	0.460	0.298

4. Discussion

We performed an integrative analysis of brain imaging and protein expression data to jointly identify AD related biomarkers and their associations using a new sparse learning model DSCCA. The overall association performance of DSCCA is better than SCCA. the combination of its two canonical variables are much more powerful in discriminating multiple diagnostic groups simultaneously. Using both imaging and proteomic canonical variables in DSCCA as predictors, we obtained very promising prediction performances: HC vs AD (92.1%), HC vs MCI (75.3%) and MCI vs AD (70.3%), which were competitive with prior multi-modal studies. Since the classification was done under default parameter settings and the sample size is

Table 3. Five-fold cross validation classification performances (%) using canonical variables **Xu** and **Yv**. HC vs MCI, MCI vs AD, and HC vs AD are performed as three tasks separately.

		Train			Test		
		HC vs MCI	HC vs AD	MCI vs AD	HC vs MCI	HC vs AD	MCI vs AD
DSCCA	f1	97.17	100.00	94.19	75.00	91.30	60.87
	f2	86.79	96.51	84.88	85.71	95.65	60.87
	f3	96.23	100.00	94.19	85.71	91.30	86.96
	f4	93.40	95.35	75.58	57.14	100.00	78.26
	f5	72.32	82.61	69.57	72.73	82.35	64.71
	mean	89.18	94.89	83.68	75.26	92.12	70.33
PMA	f1	60.38	77.91	65.12	71.43	86.96	73.91
	f2	66.98	84.88	74.42	71.43	95.65	60.87
	f3	66.04	80.23	63.95	50.00	86.96	60.87
	f4	68.87	80.23	59.30	42.86	82.61	78.26
	f5	65.18	77.17	60.87	31.82	64.71	64.71
	mean	65.49	80.09	64.73	53.51	83.38	67.72

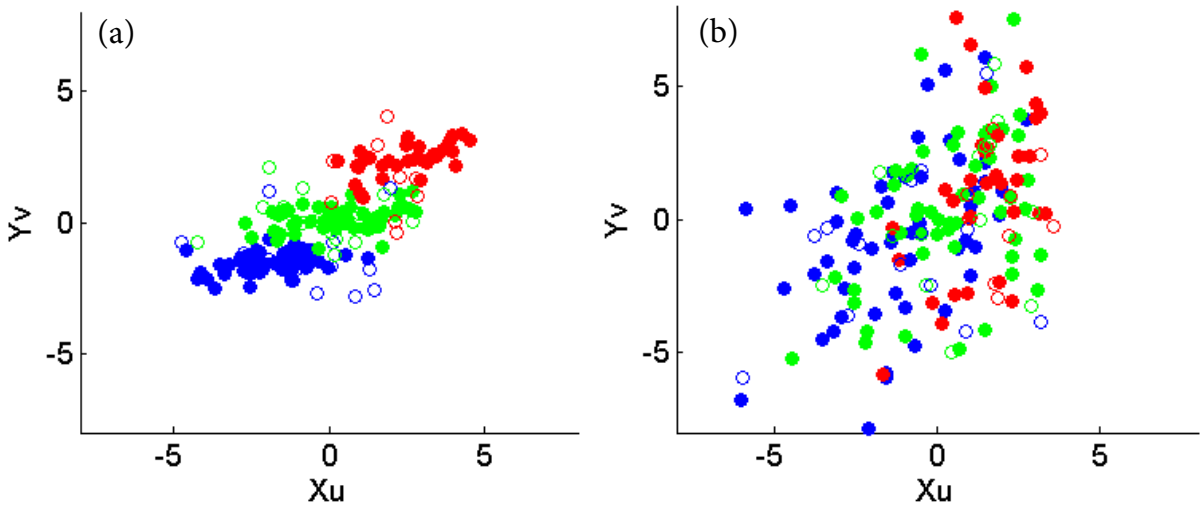

Fig. 2. Plot of canonical variables **Xu** and **Yv**. Left: DSCCA; Right: PMA; Red: AD; Green: MCI; Blue: HC; Solid: Training; Circle: Test.

very limited, we expect improved performances with more advanced parameter optimization strategies and/or larger sample sizes.

In real applications, many identified proteomic markers are found to be interconnected, but the underlying mechanisms still warrant further investigation. Replication in independent large samples will be important to confirm these findings. Further pathway enrichment analysis could be performed as a future direction to identify underlying biological pathways of relevant genes and proteins. Considering the ever increasing data volume and diversity in many complex diseases, another potential future topic is to investigate whether DSCCA can help identify valuable complementary information between new -omics features and further improve the classification performance.

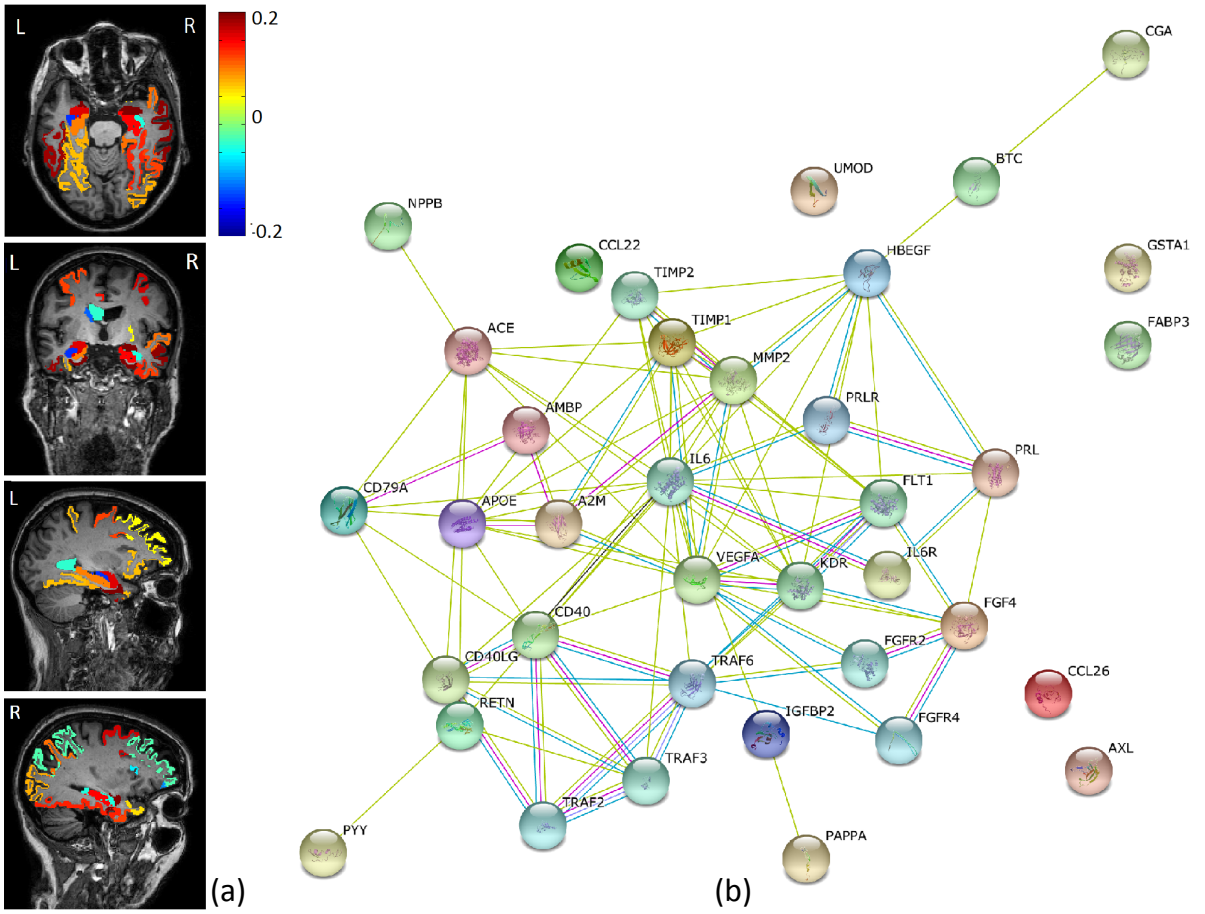

Fig. 3. Common imaging and proteomic markers across 5-fold cross-validation. (a): Mapping of imaging canonical loadings onto the brain; (b): Known interactions between identified protein biomarkers from STRING database.

Acknowledgement

This work was supported by NIH R01 EB022574, R01 LM011360, U01 AG024904, R01 AG19771, P30 AG10133, UL1 TR001108, K01 AG049050 and R00 LM011384; DOD W81XWH-14-2-0151, W81XWH-13-1-0259, and W81XWH-12-2-0012; and NCAA 14132004 at Indiana University.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceuti-

cal Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

1. Alzheimers-Association: Alzheimers disease facts and figures. Alzheimers and Dementia 12, 4 (2016)
2. Chen, X., Liu, H., Carbonell, J.G.: Structured sparse canonical correlation analysis. In: International Conference on Artificial Intelligence and Statistics (2012)
3. Chi, E., Allen, G., et al.: Imaging genetics via sparse canonical correlation analysis. In: Biomedical Imaging (ISBI), 2013 IEEE 10th Int Sym on. pp. 740–743 (2013)
4. Del Bo, R., Ghezzi, S., Scarpini, E., Bresolin, N., Comi, G.: Vegf genetic variability is associated with increased risk of developing alzheimer's disease. Journal of the neurological sciences 283(1), 66–68 (2009)
5. Du, L., Yan, J.W., Kim, S., Risacher, S.L., Huang, H., Inlow, M., Moore, J.H., Saykin, A.J., Shen, L., Initia, A.D.N.: A novel structure-aware sparse learning algorithm for brain imaging genetics. Medical Image Computing and Computer-Assisted Intervention - Miccai 2014, Pt Iii 8675, 329–336 (2014)
6. Hinrichs, C., Singh, V., Xu, G., Johnson, S.C.: Predictive markers for ad in a multi-modality framework: an analysis of mci progression in the adni population. Neuroimage 55(2), 574–89 (2011)
7. Kauwe, J., Bailey, M., Ridge, P., Perry, R., Wadsworth, M., Hoyt, K., Ainscough, B.: Genome-wide association study of csf levels of 59 alzheimer's disease candidate proteins: significant associations with proteins involved in amyloid processing and inflammation. Plos Genetics 10(10), e1004758 (2014)
8. Lin, D., Calhoun, V.D., Wang, Y.P.: Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. Med Image Anal (2013)
9. Liu, J., Ji, S., Ye, J.: Multi-task feature learning via efficient l2,1-norm minimization. In: In Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. pp. 339–348. AUAI Press (2009)
10. Lu, K., Ding, Z.M., Ge, S.: Sparse-representation-based graph embedding for traffic sign recognition. Ieee Transactions on Intelligent Transportation Systems 13(4), 1515–1524 (2012)
11. Parkhomenko, E., Tritchler, D., Beyene, J.: Sparse canonical correlation analysis with application to genomic data integration. Statistical Applications in Genetics and Molecular Biology 8, 1–34 (2009)
12. Shen, L., Kim, S., Qi, Y., Inlow, M., Swaminathan, S., Nho, K., Wan, J., Risacher, S.L., Shaw, L.M., Trojanowski, J.Q., Weiner, M.W., Saykin, A.J., Adni: Identifying neuroimaging and proteomic biomarkers for mci and ad via the elastic net. Multimodal Brain Image Analysis 7012, 27–34 (2011)
13. Tarkowski, E., Issa, R., Sjgren, M., Wallin, A., Blennow, K., Tarkowski, A., Kumar, P.: Increased intrathecal levels of the angiogenic factors vegf and tgf- in alzheimers disease and vascular de-

mentia. Neurobiology of aging 23(2), 237–243 (2002)
14. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58(1), 267–288 (1996)
15. Wan, J., Kim, S., et al.: Hippocampal surface mapping of genetic risk factors in AD via sparse learning models. MICCAI 14(Pt 2), 376–83 (2011)
16. Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10(3), 515–34 (2009)
17. Yan, J., Du, L., Kim, S., Risacher, S.L., Huang, H., Moore, J.H., Saykin, A.J., Shen, L.: Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. Bioinformatics 30(17), i564–71 (2014)
18. Yan, J., H, H., Kim, S., Moore, J., Saykin, A., Shen, L., Initia, A.D.N.: Joint identification of imaging and proteomics biomarkers of alzheimer's disease using network-guided sparse learning. In: In Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. pp. 665–668. IEEE (2014)
19. Zhang, D.Q., Wang, Y.P., Zhou, L.P., Yuan, H., Shen, D.G., Initia, A.D.N.: Multimodal classification of alzheimer's disease and mild cognitive impairment. Neuroimage 55(3), 856–867 (2011)
20. Zheng, Y., Wang, Q., Xiao, B., Lu, Q., Wang, Y., Wang, X.: Involvement of receptor tyrosine kinase tyro3 in amyloidogenic app processing and -amyloid deposition in alzheimer's disease models. Plos One 7(6), e39035 (2012)

ENFORCING CO-EXPRESSION IN MULTIMODAL REGRESSION FRAMEWORK

PASCAL ZILLE[1], VINCE D. CALHOUN[2] and YU-PING WANG[1,*]

[1] *Biomedical Engineering Department, Tulane University.*
[2] *The Mind Research Network, University of New Mexico.*
[*] *E-mail: wyp@tulane.edu*

We consider the problem of multimodal data integration for the study of complex neurological diseases (e.g. schizophrenia). Among the challenges arising in such situation, estimating the link between genetic and neurological variability within a population sample has been a promising direction. A wide variety of statistical models arose from such applications. For example, Lasso regression and its multitask extension are often used to fit a multivariate linear relationship between given phenotype(s) and associated observations. Other approaches, such as canonical correlation analysis (CCA), are widely used to extract relationships between sets of variables from different modalities. In this paper, we propose an exploratory multivariate method combining these two methods. More Specifically, we rely on a 'CCA-type' formulation in order to regularize the classical multimodal Lasso regression problem. The underlying motivation is to extract discriminative variables that display are also co-expressed across modalities. We first evaluate the method on a simulated dataset, and further validate it using Single Nucleotide Polymorphisms (SNP) and functional Magnetic Resonance Imaging (fMRI) data for the study of schizophrenia.

Keywords: Multimodal Analysis, Collaborative Regression, CCA, Sparse Models, Schizophrenia.

1. Introduction

An increasing amount of high-dimensional biomedical data such as micro arrays (mRNA, SNP) or brain imaging sequences (MRI, PET) is collected every day. Classical unimodal analysis often ignore the potential joint effects that may exist, for example, between genes and specific brain regions for diseases such as Schizophrenia, Alzheimer, etc. By harnessing these joint effects across modalities, we might be able to identify new mechanisms that uni-modal methods may fail to capture. Imaging genomics is an emerging field whose aim is precisely to leverage the wealth of biomedical knowledge provided by genomic and brain imaging data. Integrating such multimodal data sets is critical to extract meaningful bio-markers, improve clinical outcome prediction or identify potential associations across modalities. Unfortunately, as mentioned by Lin[1], such studies using genomic and brain imaging data often run into two limitations: The first one is an average small sample size, which may result in over fitting issues. In order to address such constraint, many authors relied on the use of sparse models. One classical method introduced by Tibshirani[2] is the Lasso regression. The second limitation is poor biomarker reproducibility across studies. Although this issue remains an open problem, one may hope that using appropriate priors over the solution will lead to an improved consistency of the result across different studies.

1.1. *Motivation: the study of Schizophrenia*

Schizophrenia is a serious neurological disorder that affects around 1% of the general population. It is regarded as the result of various factors including genetic variants, brain development abnormalities and environmental effects. Identifying critical genes or SNPs related to schizophrenia[3,4] has been a challenging issue. Many studies relied as well on brain imaging techniques[5,6] to pinpoint functional abnormalities in brain regions for schizophrenia patients. Multimodal analysis (e.g. using both genomic and brain imaging) often improve generalization in situations in which many irrelevant features are present. In their recent paper, Cao et al.[7] proposed a sparse representation based variable selection (SRVS) algorithm relying on sparse regression model to integrate both SNP and fMRI in order to perform biomarker selection for the study of schizophrenia. Lin[8] proposed a group sparse canonical correlation analysis (CCA) method based on SNP and fMRI data to extract correlation between genes and brain regions. Le Floch et al.[9] combined univariate filtering and Partial Least Squares (PLS) to identify SNPs covarying with various neuroimaging phenotypes. It appears that both regression and CCA methods display promising behaviors when combining SNP and fMRI data for the study of schizophrenia. In this work, we will try to merge these two methods in order to make the most out of both formulations.

The rest of this paper is organized as follows: we introduce in Section 2 some of the relevant methods as well as the motivation for this work. A novel approach to multivariate regression problems is then proposed in Section 3. Such method is then evaluated on both synthetic and real datasets in Section 4, followed by some discussions and concluding remarks in Section 5.

2. Methods

2.1. *Learning with L_1 penalty*

We consider $M \in \mathbb{N}^+$ distinct (i.e. from different modalities) datasets with n samples and $p_m \in \mathbb{N}^+$ $(m = 1, .., M)$ variables each. The m-th dataset is represented by a matrix $\mathbf{X}_m \in \mathbb{R}^{n \times p_m}$. Additionally, each sample is assigned a class label (e.g. case/controls) $y_i \in \{-1, 1\}$, $i = 1, .., n$. Our goal is to look for a linear link between those class labels and the M data matrices. Let us consider the following regression model:

$$\min_{\beta} \sum_{m=1}^{M} \|\mathbf{y} - \mathbf{X}_m \beta_m\|_2^2 + \lambda \|\beta\|_1 \tag{1}$$

The model described by Eq. 1 performs both variable selection and regularization. It often improves the prediction accuracy and interpretability of the results compared to the use of classical ℓ_2 norm regularization terms, especially when the number of variables is far greater than the number of observations. In some situations, we have several output vectors $\mathbf{y}_m, \forall m = 1, .., M$ and the m datasets are from the same modality: multi-task Lasso[10] was proposed to capture shared structures among the various regression vectors. We consider the following model:

$$\min_{\beta} \sum_{m=1}^{M} \|\mathbf{y}_m - \mathbf{X}_m \beta_m\|_2^2 + \lambda \sum_{p=1}^{P} \|\beta^p\|_2 \tag{2}$$

where P is the dimension of the problem and β^p is the p-th row of the matrix such that $\beta = [\beta_1, .., \beta_m]$ (i.e. the β_m are stacked horizontally). Such norm is also referred to as the ℓ_1/ℓ_2 norm, and is used to both enforce joint sparsity across the multiple β_m and estimate only a few non-zero coefficients. Enforcing regularity within a modality[11,12] (and across tasks) has been an active aspect of regression models, and has proven to increase reliability and results. However, since often pair-wise closeness is looked for in the common subspace, such methods will often fail to capture relationships across modalities.

2.2. *Collaborative learning*

Collaborative (or Co-regularized) methods[13] are based on the optimization of measures of agreement and smoothness across multi-modal datasets. Smoothness across modalities is enforced through a joint regularization term. Their general model can be expressed as follows:

$$J(\beta) = \sum_{m=1}^{M} \|\mathbf{y} - \mathbf{X}_m\beta_m\|_2^2 + \gamma \sum_{m,q=1}^{M} \|\mathbf{U}_m\beta_m - \mathbf{U}_q\beta_q\|_2^2 + \lambda\|\beta\|_1 \tag{3}$$

where the $\mathbf{U}_m,\ m = 1, .., M$ are arbitrary matrices whose roles are to control the cross-view joint regularization between each pair of vectors $(\beta_m, \beta_q),\ m, q = 1, .., M$. Scalar parameter $\gamma \geq 0$ controls the influence of such cross-regularization term. Notice that if $\gamma = 0$, we fall back on the original Lasso formulation. Collaborative learning is an interesting extension of Eq.1 allowing the user to explicitly enforce regularization across modalities. In this work, we rely on a special case of collaborative methods (introduced later in section 3) to address the following aspects: (i) Extend the regularization idea across modalities; (ii) Assume that relationships between variable are not available as a prior knowledge (as opposed, e.g., to Xin[11]); (iii) Define links between components using correlation measure. To do so, we first briefly introduce in the next section some of the classical methods to extract meaningful relationships between variables across modalities.

2.3. *Extracting relationship between datasets*

A wide variety of problems amount to the joint analysis of multimodal datasets describing the same set of observations. Often, a mean to perform such analysis is to learn projection subspaces using paired samples such that structures of interest appear more clearly. Some of these methods are for example: Canonical correlation analysis[14] (CCA), Partial least squares[9] (PLS) or cross-modal factor analysis (CFA). Among them, CCA is probably the most widely used. Its goal is to extract linear combinations of variables with maximal correlation between two (or more) datasets. Using similar notations as in the previous section,and assuming $M = 2$, one formulation of CCA is expressed as follow:

$$\underset{\beta_1,\beta_2}{\operatorname{argmin}}\ J_{cca}(\beta_1, \beta_2) \ = \ \|\mathbf{X}_1\beta_1 - \mathbf{X}_2\beta_2\|^2 \tag{4}$$

to which a constraint on the norm of canonical vectors β_1, β_2 is added to avoid the trivial null solution. In recent years, CCA has been widely applied to genomic data analysis. As a consequence, many studies on sparse versions of CCA (sCCA) have been proposed[8,15–18] to

cope with the high dimension but low sample size problem. In the next section, we will rely on a CCA term to measure co-expression between variables from different modalities.

3. Enforcing cross-correlation in regression problems

3.1. *MT-CoReg formulation*

As discussed in Section 1, several methods have been proposed to: (*i*) Associate a phenotype and datasets while enforcing prior over solution; (*ii*) Extract relationships between coupled or co-expressed datasets. In the present study, we propose to associate both the regression and CCA frameworks in the case of $M = 2$ datasets. Our motivation is to extract informative features that also display a significant amount of correlation across modalities. A simple way to combine Lasso and sparse CCA would be a weighted combination of Eq.(1) and Eq.(4):

$$\min_{\beta} J(\beta) = (1-\gamma) \sum_{m=1}^{2} \|\mathbf{y} - \mathbf{X}_m \beta_m\|_2^2 + \gamma \|\mathbf{X}_1 \beta_1 - \mathbf{X}_2 \beta_2\|^2 + \lambda \|\beta\|_1 \tag{5}$$

where $\gamma \in [0, 1]$ is a weight parameter. Notice that Eq.(5) can be expressed within the collaborative framework introduced in Section 2.2. If we take a look at Eq.(3) with $M = 2$, $\mathbf{U}_1 = \mathbf{X}_1$ and $\mathbf{U}_2 = \mathbf{X}_2$, we fall back on Eq.(5). Let us call this model CoReg for *Collaborative Regression*. Interestingly, a similar model has been considered before by Gross[19] to perform prediction using breast cancer data. However, to our opinion, such formulation might prove to be too constraining. It essentially amounts to force each component of the β_m's to fit both the regression term and the CCA one. We illustrate such behaviour using a toy dataset later in Section 3.4. Since our goal is to perform feature selection, we may allow the model to be slightly more flexible. We thus propose an alternative formulation by first duplicating each β_m into two components such that:

$$\beta_m = [\alpha_m, \theta_m] \ , \ \forall m = 1, 2 \tag{6}$$

where α_m, θ_m are vectors from $\mathbb{R}^{p_m}$. As a consequence, the β_m's are now matrices such that $\beta_m \in \mathbb{R}^{p_m \times 2}$ $\forall m = 1, 2$. We then propose the following MT-CoReg formulation:

$$\min_{\beta} J(\beta) = (1-\gamma) \sum_{m=1}^{2} \|\mathbf{y} - \mathbf{X}_m \alpha_m\|_2^2 + \gamma \|\mathbf{X}_1 \theta_1 - \mathbf{X}_2 \theta_2\|^2 + \lambda \sum_{m=1}^{2} \sum_{i=1}^{p_m} \|\beta_m^i\|_2 \tag{7}$$

where β_m^i is the i-th row of β_m, i.e. $\beta_m^i = [\alpha_m(i), \theta_m(i)] \in \mathbb{R}^2$. The third term of Eq.(3.3) is simply the ℓ_1/ℓ_2 norm of each of the β_m. As we can observe from looking at Eq.(3.3), each 'component' (i.e. column of β_m) will be involved in separate parts of the functional J: (*i*) components α_m are the fit to the regression term of Eq.(3.3); (*ii*) components θ_m are the fit of the CCA term of Eq.(3.3). Each pair (α_m, θ_m) and $m = 1, 2$ is coupled through the use of the ℓ_1/ℓ_2 norm from the third term in Eq.(3.3). Although their values are different, shared sparsity patterns are encouraged within each pair (α_m, θ_m). As a consequence, we allow the method to be significantly more flexible in terms of solutions: different values can be taken to simultaneously fit the Regression and CCA parts. We hope that such framework will encourage the selection of features that are discriminative (via the regression part) but also co-expressed across modalities (via the CCA part). Note that when $\gamma = 0$, criterion (3.3)

essentially reduces to the initial regression problem of Eq.(1), while setting $\gamma = 1$ amounts to solving a conventional sparse CCA problem. A schematic view of the MT-CoReg pipeline can be seen in Fig.(1). In the next section, we briefly explain how to solve the problem described in Eq.(3.3).

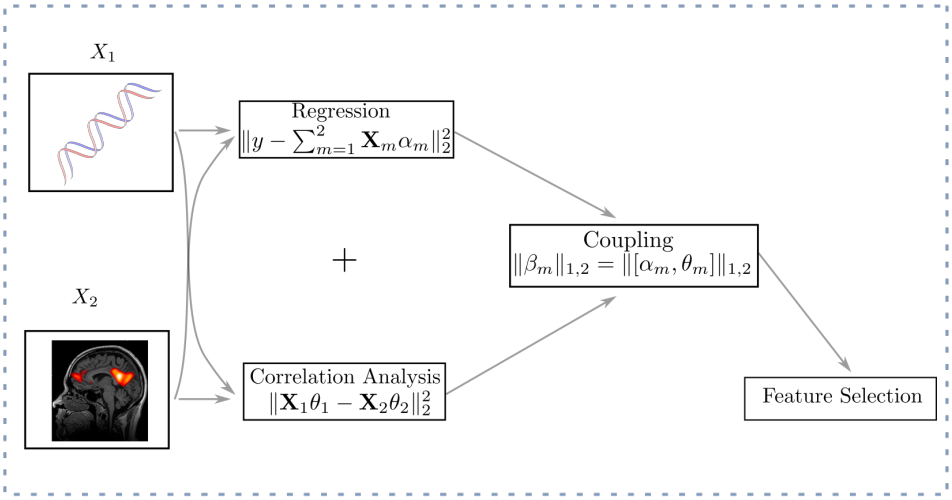

Fig. 1. Schematic view of the MT-CoReg pipeline. From two different datasets $X1$ and $X2$ from different modalities (here SNP and fMRI respectively), we fit both a regression and CCA terms and couple the resulting components (α_m, θ_m) using the $\ell 1/\ell_2$ norm denoted $\|\cdot\|_{1,2}$ here. The ultimate goal is to find discriminative SNP and brain regions that are also co-expressed across modalities.

3.2. *Optimization*

We solve the problem from Eq.(3.3) by optimizing the β_m's alternatively over iterations until convergence, in a similar fashion to Wilms[20] et al. formulation of sCCA. Suppose we have an initial value β_1^* for β_1, and want to estimate β_2. Updating matrix β_2 can be recast into a problem of the following form:

$$\min_{\tilde{\beta}_2} J(\tilde{\beta}_2|\beta_1^*) = \|\tilde{\mathbf{y}}_2 - \tilde{\mathbf{X}}_2\beta_2\|_F^2 + \lambda \sum_{i=1}^{p_2} \|\beta_2^i\|_2 \tag{8}$$

where

$$\tilde{\mathbf{y}}_2 = [\sqrt{(1-\gamma)}\mathbf{y}, \sqrt{\gamma}\mathbf{X}_1\theta_1^*], \quad \tilde{\mathbf{X}}_2 = [\sqrt{(1-\gamma)}\mathbf{X}_2, \sqrt{\gamma}\mathbf{X}_2] \tag{9}$$

Obviously, Eq.(8) is a classical group-lasso regression problem[10] (cf. Eq.(2)). It is easy to show that updating β_1 reduces to solving a similar problem. As a consequence, solving our mixed Lasso/CCA problem from Eq.(3.3) can be briefly summarized as:

1 Initialization: estimate initial values for α_1, β_1, α_2, β_2 using ridge regression and ridge CCA.
2 Assume β_1's value fixed, and update β_2 using Eq.(8).
3 Assume β_2's value fixed, and update β_1 using the adapted version of Eq.(8).
4 Go back to step 2. until convergence

3.3. *Parameter selection*

Solving problem from Eq.(3.3) requires the estimation of two parameters, λ and γ, which respectively control the weights of the sparsity and the co-expression regularization terms.
The choice of sparsity parameter λ for this type of problems is known to display a high sensitivity[21]. In order to make the searching process more robust, we chose to let the sparsity level of the solution control the tuning parameter value[22,23]. Consider a column vector $\beta \in \mathbb{R}^p$ (e.g. a column of β from Eq.): let us denote $|\beta|_\kappa$ the κ-th ($\kappa \in \mathbb{N}^+$) largest absolute magnitude of β. We can define a correspondence between λ and κ by making sure that for each iteration, we have $\lambda \in [|\beta|_\kappa, |\beta|_{\kappa+1}]$. The selection can be looked for around the sample size (i.e. $\kappa = n$ for the entire estimation process), which helps drastically stabilize the estimation process in practice.
As for the estimation of γ, we chose to rely on a technique introduced by Sun et al.[24] based on variable selection stability. Its main goal is to select a given tuning parameter so that the associated variable selection method (in our case, the model from Eq.(3.3)) is stable in terms of the features it selects. In this framework, the training set is split in two halves using resampling (bootstrap resampling in our case). The variable selection method is then applied to each of the subsamples along a grid of candidate values for the parameter. Kappa selection criterion[25] is then used to measure the degree of agreement between the two sets of variables obtained for a given parameter value. This process is then repeated a number of times, and an approximated measure of selection consistency is derived. The parameter value for which this consistency is the highest (after correction for the number of non-zeros elements retained) is the one kept for the estimation.

3.4. *MT-CoReg VS. CoReg*

As mentioned earlier in Section 3.1, in their CoReg model from Eq.(5) Gross et al.[19] did not separate the solution vectors β_m into two components. We then propose to illustrate the behavior of both models (Eq.(5) and Eq.(3.3)) on a toy dataset.
We generated $M = 2$ data matrices $\mathbf{X}_1, \mathbf{X}_2$ such that $p_1 = p_2 = 30$ and $n = 50$ observations. We used a latent variable model to simulate cross-correlated components so that columns $p = [1, ..5] \cup [10, ..15]$ of $\mathbf{X}_1, \mathbf{X}_2$ are mutually co-expressed. We further use columns $p = [10, ..15] \cup [20, ..25]$ to generate a phenotype vector $\mathbf{y}$ such that $\mathbf{y}_i \in \{-1; 1\}$. With such setup, columns $p = [10, ..15]$ correspond to both non-zeros values in the true regression and canonical coefficients. Furthermore, let us point out that these non-zero values are diferent (canonical coefficients' amplitude is lower than the regression ones). This setup can be seen in the first row of Fig.(2, *Truth*), where the blue and red curves are the values taken by the canonical and regression coefficients respectively. Resulting estimates for sCCA, Lasso, CoReg[19] as well as proposed method MT-CoReg can also be seen in Fig.(2). In such scenario, while CoReg model assumes that regression and canonical coefficients have identical values, MT-CoReg has a wider scope and allows a finer joint estimation of both components types.

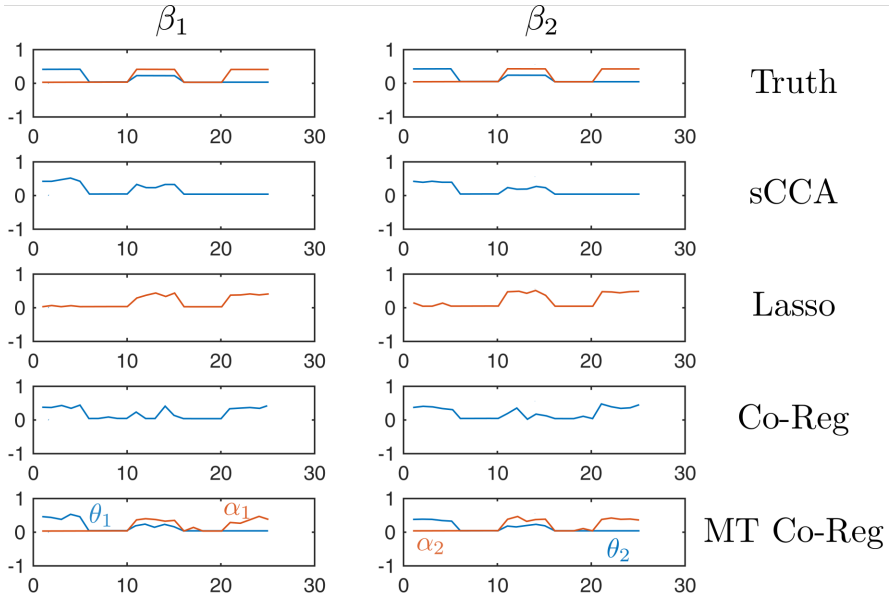

Fig. 2. Resulting estimates β_1, β_2 on the toy dataset. (*Truth*) blue and red curves are the values taken by the true canonical and regression coefficients respectively. Solutions obtained with sCCA, Lasso, CoReg[19] and proposed method MT-CoReg are displayed. Notice that columns columns $p = [10, ..15]$ correspond to both non-zeros values in the true regression and canonical coefficients, although their amplitudes are different. By relaxing the assumption that regression and canonical coefficients have identical values, MT-CoReg allows a finer joint estimation of both components types compared to CoReg.

4. Experiments

In this section, we evaluate the proposed estimator from Eq.3.3. Performances will be assessed in terms of feature selection relevance on both simulated and real data.

4.1. *Results on synthetic data*

For our first test, we simulate both fMRI and SNP datasets. Similar to the toy dataset from Section 3.4, we start by generating explanatory variables $\alpha_1^*, \alpha_2^* \in \mathbb{R}^{900}$ for both genomic and brain imaging data. The first 100 components of α_1^*, α_2^* are drawn from Normal distribution, while the rest is set to zero. The total number of observations is set to $n = 200$. Genomic values are coded as 0 (no minor allele), 1 (one minor allele), and 2 (two minor allele). We first define a minor allele frequency η drawn from a uniform distribution $\mathcal{U}([0.2, 0.4])$. The i-th SNP is then generated from a binomial distribution $\mathcal{B}(2, \eta_i)$. For the imaging data, voxels values were drawn from a Gaussian distribution $\mathcal{N}(0, I_p)$. Finally, binary phenotype $\mathbf{y}$ data are generated from $\mathcal{B}(1, d_i)$, where $d_i = \dfrac{exp(5\sum_{m=1}^{M} \mathbf{X}_m \alpha_m^*)}{1 + exp(5\sum_{m=1}^{M} \mathbf{X}_m \alpha_m^*)}$. Furthermore, we add 100 additional variables to the problem that will play the role of cross-correlated variables. Two canonical vectors $\theta_1^*, \theta_2^* \in \mathbb{R}^{100}$ are drawn from Normal distribution. Cross-correlated SNP are drawn from $\mathcal{B}(2, \text{logit}^{-1}(-a_i + \text{logit}(\eta_i)))$ where a is issued from $\mathcal{N}(\theta_1^* \mathbf{y}, I_{100})$, while cross-correlated voxels are drawn from $\mathcal{N}(\theta_2^* \mathbf{y}, I_{100})$. The final dataset is made of $n = 200$ observations of $p = 1000$ variables for both SNP and fMRI. Each of these datasets is made of explanatory and cross-

correlated components. A common way to assess the performance of a model when it comes to feature selection is to measure the true positive rate (TPR) and false positive rate (FPR). TPR reflects the proportion of variables that are correctly identified, while FDR reflects the proportion of variables that are incorrectly selected by the model. We apply MT-CoReg to 100 random generation of the dataset described above. The tuning parameter γ from Eq.(3.3) that weights the CCA term against the regression one is optimized through a grid search over $\{[0] \cup [10^{-1+\ell/20}]; \ell = 0, .., 20\}$. We plotted TPR values against FDR ones in Fig.(3) for two different cases. In the first (left) subfigure are displayed TPR/FDR values relative to non-zero components of α_1^*, α_2^* for $\gamma = 0$ (i.e. classical Lasso), $\gamma = \gamma$(C.S.) where the weight value is determined using consistency selection (C.S.) scheme described in Section 3.3, and $\gamma = 1$ (i.e. classical sCCA). We can observe that although classical regression seems to perform slightly better for really low FDR values, MT-CoReg is quickly catching up around $FDR \approx 0.15$. sCCA, on the other hand, has a low selection power. The second (bottom) figure displays TPR/FDR values relative to non-zero components of θ_1^*, θ_2^*, i.e. the cross-correlated components. We can observe that MT-CoReg performs as well as sCCA, while Lasso is unable to properly select the components of interest. It is encouraging to see that MT-CoReg takes the best of both methods and seems to properly select the components we are interested in. It seems to confirm our hypothesis that using a mix of both terms may lead to an improved feature selection accuracy. In the next section, we apply the same method to a real dataset of fMRI and SNP data.

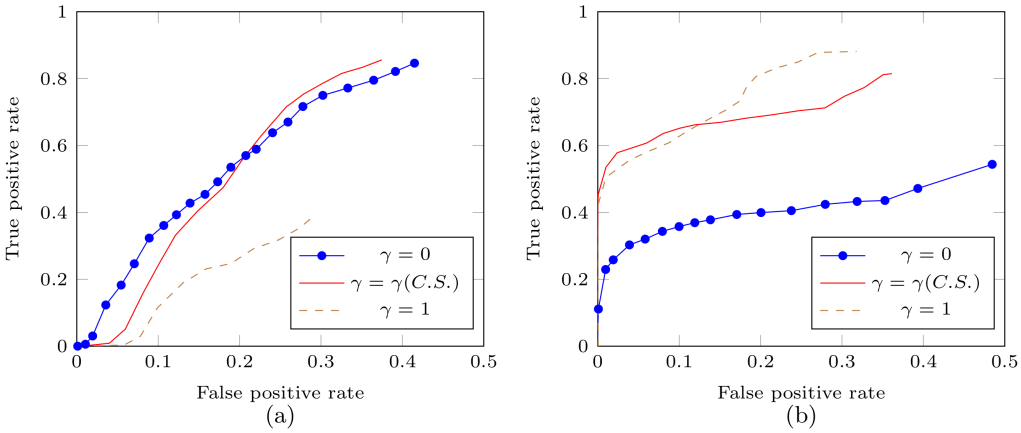

Fig. 3. TPR against FDR values averaged over 100 simulations for different γ values. Fixing $\gamma = 0$ amounts to using Lasso regression, while $\gamma = 1$ is equivalent to classical sparse CCA. γ(C.S.) is the ROC curve obtained while using consistency selection (C.S.) scheme described in section 3.3 to automatically estimate γ. (a) values for the selection of first 100 components (i.e. the explanatory components) only (b) values for the selection of the last 100 components (i.e. the cross-correlated components). It can be seen that a non-trivial weight combinaison for γ seems to be taking the best of the two methods that are Lasso ($\gamma = 0$) and CCA ($\gamma = 1$).

4.2. *Results on real imaging genetics data*

4.2.1. *Data acquisition*

Both SNP and fMRI acquisition were conducted by the Mind Clinical Imaging Consortium (MCIC) for 214 subjects, including 92 schizophrenia patients (age: 34 ± 11, 22 females) and 116 controls (age 32 ± 11, 44 females). Schizophreniac were diagnosed based on DSM-IV-TR citeria. Controls were free of any medical, neurological of psychiatric illnesses.

fMRI were acquired during a sensor motor task with auditory simulation. Data were pre-processed with SPM5, spatially normalized and resliced, smoothed, and analyzed by multiple regression considering the stimulus and their temporal derivatives plus an intercept term as regressors. For each patient, a stimulus-on vs. stimulus-off contrast image was extracted. 116 ROIs were extracted based on the aal brain atlas, which resulted in 41236 voxels left for analysis. SNP data were obtained from blood sample using Illumina Infinium HumanOmni1-Quad array covering 1,140,419 SNP loci. After standard quality control procedures using PLINK software package [a], a final dataset spanning $777,635$ SNP loci was available. Each SNP was categorized into three clusters based on their genotype and was represented with discrete numbers: 0 (no minor allele), 1 (one minor allele) and 2 (two minor alleles). SNPs with $> 20\%$ missing data were deleted and missing data were further imputed. SNPs with minor allele frequency $< 5\%$ were removed. This procedure yielded a final set of $129,145$ SNPs.

4.2.2. *Significance analysis*

In order to achieve a stable feature selection process, we follow Lin[8] and perform $N = 100$ random samplings out of the 214 total subjects, where for each time 80% are used for training and parameter selection, while the remaining 20% are used for evaluation. At the $k-th$ random sampling, we can calculate a set of solution vectors $\hat{\beta}_m^k, m \in \{1, 2\}$. It is then possible to define a measure of relevance p_m^i for the i-th feature in the m-th dataset such that: $p_m^i = \frac{1}{N}\sum_{k=1}^{N} I(\hat{\beta}_m^k(i) \neq 0)$ where $i = 1, .., d_m$ is the feature index and $I(\cdot)$ is the indicator function. We can then rank each SNP and voxel based on their associated relevance measure and apply a cut-off threshold of 0.3 (c.f. Lin[8]). After applying this significance test, we were left with a subset of 43 SNP spanning 30 genes and 6 ROI with a number of selected voxels over 5.
We display in Table.1 the list of each of the 43 selected SNP, as well as their associated genes. Some of them have been identified by other similar studies[8,26,27] such as CNTNAP2, GLI2, GRIK3, NOTCH4, SUCLG2, GABRG2. Others have been identified from well-known databases[28] such as GRIK4 or HTR4. We display in Table.2 the list of the selected ROI as well as the corresponding voxel count for each one of them. ROI for which less than 5 voxels were selected where dismissed. Once again, it is encouraging to note that each of the selected ROI (3, 7, 11, 40, 51, 100 from aal.) have been identified in similar studies[8,29] on the same dataset. Other studies pointed out both functionnal or structural differences in the middle occipital gyrus[30] and the parahippocampal gyrus[31] for schizophrenic patients. Finally, a detailled slice view of the selected voxels can be seen in Fig.(4).

[a] http://pngu.mgh.harvard.edu/ purcell/plink

Table 1. List of selected SNP and their associated genes.

SNP ID	Gene name	SNP ID	Gene name	SNP ID	Gene name	SNP ID	Gene name
rs3856465	ATP6V1C2	rs11607732	GRIK4	rs815533	CACNA2D3	rs10748732	HPSE2
rs12333931	CNTNAP2	rs12332417	HTR4	rs2373347	CNTNAP2	rs13359903	HTR4
rs2407264	CYSLTR2	rs7725785	HTR4	rs9535112	CYSLTR2	rs11875988	LIPG
rs6567629	DHRSX	rs12454370	LIPG	rs858341	ENPP1	rs9787820	LRRC4C
rs16842460	EPHB1	rs17819648	MAML2	rs11927660	FGF12	rs3134797	NOTCH4
rs17599845	FHIT	rs3134799	NOTCH4	rs10926254	FMN2	rs394657	NOTCH4
rs4659573	FMN2	rs1009708	PDE2A	rs11060822	FZD10	rs7111188	PDE2A
rs12824777	FZD10	rs17016738	RARB	rs2963094	GABRG2	rs12101383	SMAD6
rs10831614	GALNTL4	rs7030433	SMARCA2	rs7602673	GLI2	rs573700	SPRY2
rs6753202	GPD2	rs9849270	SUCLG2	rs1392744	GRIK3	rs1105880	UGT1A6
rs10502240	GRIK4	rs17863787	UGT1A6				

Table 2. List of selected ROI (from aal.) and associated voxel count.

ROI ID (aal.)	ROI name	voxels nb.
51	Left middle occipital gyrus	13
7	Left middle frontal gyrus	11
11	Left middle frontal gyrus, orbital part	9
100	Right lobule VI of cerebellar hemisphere	9
3	Left superior frontal gyrus	8
40	Right parahippocampal gyrus	7

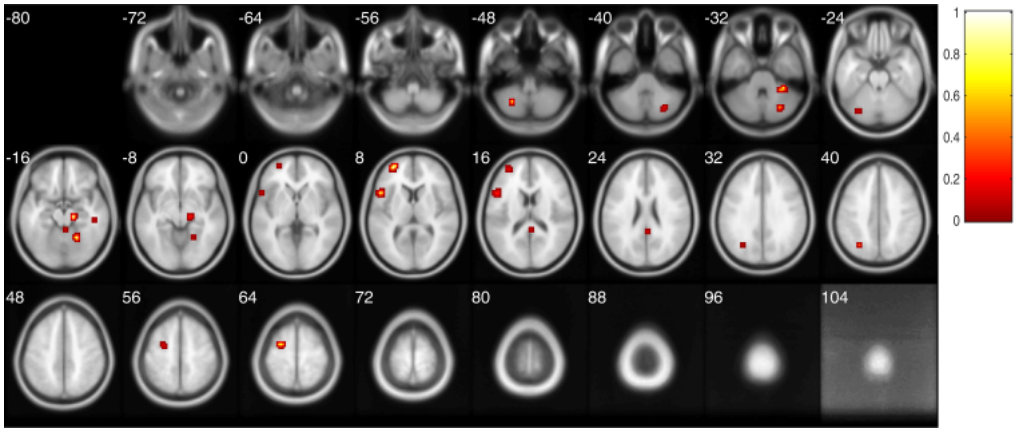

Fig. 4. Slice view of the selected voxels (without thresholding using cluster size) and their significance.

4.2.3. *Quantitative analysis*

In this section, we try to analyze the results of MT-CoReg using some quantitative metrics. We can first turn our attention to the Sum of Squared Errors (SSE) values obtained on the testing set during our tests. Histograms of SSE distributions for different γ values (i.e.

Lasso, MT-CoReg and sCCA) can be seen in Fig.(5,left): unsurprisingly, Lasso and MT-CoReg produce the lowest RSS values, while sCCA does not fit the phenotype. If we now look at Fig.(5,middle) where distributions of Pearson's correlation on the testing set are displayed for the same 3 strategies, we can see that MT-CoReg produces a better selection than Lasso in terms of cross-correlation. This seems to confirm our intuition that MT-CoReg makes the best of both Lasso and CCA by producing a solution that is good fit to the phenotype while selecting co-expressed features across modalities.
Distribution of γ values produced by the consistency selection scheme described in Section 3.3 can be seen in Fig.(5,right). Most of these values fall into the range $[0; 0.4]$, with a peak in $[0.2; 0.3]$. It does appear, at least in term of feature consistency selection, that a non-zero weight for the CCA term in Eq.(3.3) leads to improved performances.

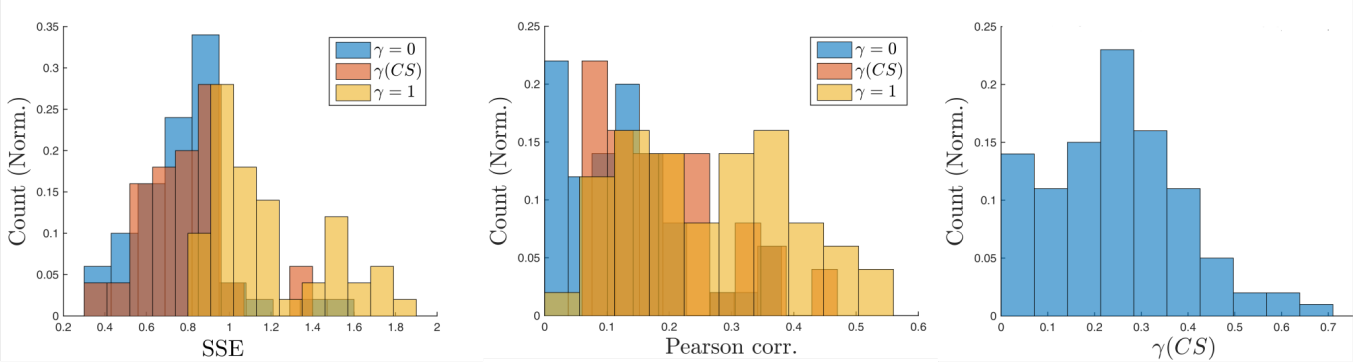

Fig. 5. Frequency distribution of RSS values (on the test set) for $N = 100$ sub-sampling of the original set of observations.

5. Conclusions

The main contributions of this paper can be summarized as follows. First, we proposed a novel variable selection approach using a CCA-like regularization term in order to enforce co-expression between modalities. Secondly, we present an efficient algorithm to solve this problem, as well as strategies to estimate the tuning parameters. On top of that, a series of experiments on both synthetical and real datasets were conducted, allowing us to evaluate the performances of the proposed method. We identified two sets of SNP and voxels in which a number of them have been previously reported to have potential relationship with the risk of schizophrenia. Further exploration of the optimization scheme (alternate estimations) as well as the selection of regularization parameter λ (see Section 3.3) will be needed in the future.

6. Acknowledgments

The authors wish to thank the NIH (NSF EPSCoR#1539067) for their partial support.

References

1. D. Lin, J. Zhang, J. Li, H. He, H.-W. Deng and Y.-P. Wang, *Multi-omic Data Integration* , p. 126 (2015).

2. R. Tibshirani, *Journal of the Royal Statistical Society. Series B (Methodological)* , 267 (1996).
3. C. M. Lewis, D. F. Levinson, L. H. Wise, L. E. DeLisi, R. E. Straub, I. Hovatta, N. M. Williams, S. G. Schwab, A. E. Pulver, S. V. Faraone *et al.*, *The American Journal of Human Genetics* **73**, 34 (2003).
4. S. R. Sutrala, D. Goossens, N. M. Williams, L. Heyrman, R. Adolfsson, N. Norton, P. R. Buckland and J. Del-Favero, *Schizophrenia research* **96**, 93 (2007).
5. M. E. Shenton, C. C. Dickey, M. Frumin and R. W. McCarley, *Schizophrenia research* **49**, 1 (2001).
6. S. A. Meda, M. Bhattarai, N. A. Morris, R. S. Astur, V. D. Calhoun, D. H. Mathalon, K. A. Kiehl and G. D. Pearlson, *Schizophrenia research* **104**, 85 (2008).
7. H. Cao, J. Duan, D. Lin, Y. Y. Shugart, V. Calhoun and Y.-P. Wang, *Neuroimage* **102**, 220 (2014).
8. D. Lin, V. D. Calhoun and Y.-P. Wang, *Medical image analysis* **18**, 891 (2014).
9. É. Le Floch, V. Guillemot, V. Frouin, P. Pinel, C. Lalanne, L. Trinchera, A. Tenenhaus, A. Moreno, M. Zilbovicius, T. Bourgeron *et al.*, *Neuroimage* **63**, 11 (2012).
10. M. Yuan and Y. Lin, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49 (2006).
11. B. Xin, Y. Kawahara, Y. Wang, L. Hu and W. Gao, *ACM Transactions on Intelligent Systems and Technology (TIST)* **7**, p. 60 (2016).
12. B. Jie, D. Zhang, B. Cheng and D. Shen, *Human brain mapping* **36**, 489 (2015).
13. U. Brefeld, T. Gartner, T. Scheffer and S. Wrobel, 137 (2006).
14. H. Hotelling, *Biometrika* **28**, 321 (1936).
15. D. M. Witten and R. J. Tibshirani, *Statistical applications in genetics and molecular biology* **8**, 1 (2009).
16. J. Chen, F. D. Bushman, J. D. Lewis, G. D. Wu and H. Li, *Biostatistics* **14**, 244 (2013).
17. L. Du, H. Huang, J. Yan, S. Kim, S. L. Risacher, M. Inlow, J. H. Moore, A. J. Saykin, L. Shen, A. D. N. Initiative *et al.*, *Bioinformatics* , p. btw033 (2016).
18. Springer, *A novel structure-aware sparse learning algorithm for brain imaging genetics* 2014.
19. S. M. Gross and R. Tibshirani, *Biostatistics* **16**, 326 (2015).
20. I. Wilms and C. Croux, *Biometrical Journal* **57**, 834 (2015).
21. E. Parkhomenko, D. Tritchler and J. Beyene, *Statistical Applications in Genetics and Molecular Biology* **8**, 1 (2009).
22. J. Duan, J.-G. Zhang, H.-W. Deng and Y.-P. Wang, *PloS one* **8**, p. e59128 (2013).
23. Z. Xu, X. Chang, F. Xu and H. Zhang, *IEEE Transactions on neural networks and learning systems* **23**, 1013 (2012).
24. W. Sun, J. Wang and Y. Fang, *Journal of Machine Learning Research* **14**, 3419 (2013).
25. J. Cohen, *Psychological bulletin* **70**, p. 213 (1968).
26. D. Lin, H. He, J. Li, H.-W. Deng, V. D. Calhoun and Y.-P. Wang, 9 (2013).
27. J. Sun, P.-H. Kuo, B. P. Riley, K. S. Kendler and Z. Zhao, *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **147**, 1173 (2008).
28. P. Jia, J. Sun, A. Guo and Z. Zhao, *Molecular psychiatry* **15**, 453 (2010).
29. D. Lin, H. Cao, V. D. Calhoun and Y.-P. Wang, *Journal of neuroscience methods* **237**, 69 (2014).
30. S. Singh, S. Modi, S. Goyal, P. Kaur, N. Singh, T. Bhatia, S. N. Deshpande and S. Khushu, *Journal of biosciences* **40**, 355 (2015).
31. M. J. Escartí, M. de la Iglesia-Vayá, L. Martí-Bonmatí, M. Robles, J. Carbonell, J. J. Lull, G. García-Martí, J. V. Manjón, E. J. Aguilar, A. Aleman *et al.*, *Schizophrenia research* **117**, 31 (2010).