MICGen: MICCAI Workshop on Imaging Genetics

Overview

MICGen: MICCAI Workshop on Imaging Genetics (http://micgen.mit.edu) is held on September 14th, 2014, in conjunction with the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference, at the Massachusetts Institute of Technology, Cambridge, MA, USA. It brings together researchers and clinicians from various fields including medical genetics, computational biology and medical imaging, presenting a forum for both fundamental concepts as well as state-of-the-art methods and applications.

Being the first MICCAI workshop in imaging genetics, MICGen includes tutorial sessions introducing the fundamental concepts and challenges of imaging genetics, as well as oral presentations and posters of accepted abstracts presenting novel methods or new applications.

Motivation

Imaging genetics studies the relationships between genetic variation and measurements from anatomical or functional imaging data, often in the context of a disorder. While traditional genetic analyses are successful for deciphering simple genetic traits, imaging genetics can aid in understanding the underlying complex genetic mechanisms of multifaceted phenotypes. Specifically, imaging-based biomarkers are used as an intermediate or alternative phenotype that provides a rich quantitative characterization of disease. As large imaging genetics datasets are becoming available, their analysis poses unprecedented methodological challenges. MICCAI offers an ideal and timely opportunity to bring together people with different expertise and shared interests in this rapidly evolving field. This motivation led to the creation of MICGen, the first MICCAI workshop on imaging genetics.

Organizing Committee

Adrian V. Dalca
Massachusetts Institute of Technology, CSAIL, Cambridge, MA, USA.

Kayhan N. Batmanghelich
Massachusetts Institute of Technology, CSAIL, Cambridge, MA, USA.

Mert R. Sabuncu
A.A Martinos Center for Biomedical Imaging, Charlestown, MA, USA
Mass. General Hospital, Harvard Medical School

MICGen 2014 Program

The program is maintained at http://micgen.mit.edu. Please check there for the latest updates.

Morning Session 1 - Genetics Tutorial

The first morning session will introduce concepts from human genetics, genome wide association studies, epigenomics, system genomics and beyond through talks by Mark Daly and Manolis Kellis. It will serve as the genetic background for the imaging genetics accepted abstracts presented in the afternoon.

8:15 - 8:30	**Welcome Remarks** Adrian Dalca EECS, Massachusetts Institute of Technology
8:30 - 9:10	**Progress in Human Genetics: GWAS and Beyond** Mark Daly Massachusetts General Hospital, Harvard Medical School
9:15 - 9:55	**Regulatory Genomics and Epigenomics of Complex Traits and Human Disease** Manolis Kellis EECS, Massachusetts Institute of Technology

Coffee Break 10:00 – 10:30

Morning Session 2 - Imaging Tutorial

The second morning session will begin with a clinical discussion at the intersection of imaging and genetics by Jordan Smoller. Then, Li Shen and Mert Sabuncu will then introduce concepts in imaging genetics, including computation of imaging phenotypes and current imaging genetics models and directions.

10:30 - 11:00	**Imaging Genetics: Decoding Psychopathology** Jordan Smoller Center for Human Genetic Research, Massachusetts General Hospital
11:05 - 11:35	**Bioinformatics Strategies for Multidimensional Brain Imaging Genetics** Li Shen Radiology and Imaging Sciences, Indiana University School of Medicine
11:40 - 12:00	**Probing Multivariate Associations Between Structural Neuroimaging Phenotypes and Genetic Markers** Mert Sabuncu A.A Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School

Lunch 12:00 – 13:00

Afternoon Session 1 - Accepted Talks and Posters - Group 1

Authors from accepted submissions will give short talks in a roughly 50 minute session, followed by respective posters. This format was chosen to encourage as much useful interaction as possible among participants.

13:00 - 13:50	**5-Minutes Talks** Group 1 (see below)
14:00 - 15:00	**Posters** Group 1 (see below)

Coffee Break 15:00 – 15:30 – All posters (Groups 1 and 2)

Afternoon Session 2 - Accepted Talks and Posters - Group 2

Authors from accepted submissions will give short talks in a roughly 40 minute session, followed by respective posters. This format was chosen to encourage as much useful interaction as possible among participants.

15:30 - 16:10	**5-Minutes Talks** Group 2 (see below)
16:10 - 17:00	**Posters** Group 2 (see below)

Talks/Posters

Group 1

- Correlation and integration of fMRI imaging and SNP data with sparse models
  Yu-Ping Wang

- Detection of genes associated with multiple correlated imaging phenotypes by a sparse group-ridge low-rank regression model
  Dongdong Lin, Hong-Wen Deng, Vince Calhoun and Yu-Ping Wang

- Probabilistic Approach to Joint Modeling of Imaging and Genetics
  Kayhan Batmanghelich, Adrian Dalca, Mert Sabuncu and Polina Golland

- Detecting Gene-Environment Interactions via a Kernel Machine Method
  Tian Ge, Thomas Nichols, Debashis Ghosh, Elizabeth Mormino, Jordan Smoller and Mert Sabuncu

- Fast Heritability Analysis Using Genome-Wide Data via Kernel Machines
  Tian Ge, Thomas Nichols, Avram Holmes, Phil Lee, Joshua Roffman, Randy Buckner, Mert Sabuncu and Jordan Smoller

- Longitudinal 3D MR Spectroscopic Imaging of 2-Hydroxyglutarate in patients with mutant IDH1 glioma
  Ovidiu Andronesi, Franziska Loebel, Wolfgang Bogner, Malgorzata Marjanska, Elizabeth Gerstner, Andrew Chi, Tracy Batchelor, Daniel Cahill and Bruce Rosen

- Hierarchical clustering of whole genome sequence data for forecasting age of Alzheimer's diagnosis
  Rachel Yotter, Xiao Da, Bilwaj Gaonkar, Roman Filipovych and Christos Davatzikos

- Genetic Determinants of Acute Cerebral Infact Volume: Results from a Preliminary Genome-Wide Associaton Study
  Lisa Cloonan, Kelsey Shideler, Cathy R Zhang, Adriana Perilla, Allison Kanakis, Kaitlin Fitzpatrick and Natalia S Rost

- Differential Effect of 17q25 Locus on White Matter Hyperintensity Volume in Patients with Ischemic Stroke
  Cathy R Zhang, Lisa Cloonan, Adrian Dalca, Ramesh Sridharan, Kaitlin Fitzpatrick, Allison Kanakis, Alison M Ayres, Jonathan Rosand, Ona Wu, Polina Goland and Natalia S Rost

Group 2

- Feature Selection and Imaging-Genetics Predictions Using a Sparse, Extremely Randomized Forest Regressor
  Albert Montillo, Shantanu Sharma and Marcel Prastawa

- Predictive Imaging-Genetics Models with Feature Selection and Dimension Reduction Using Sparse Partial Least Squares
  Rui Li, Xiaojie Huang, Shantanu Sharma and Marcel Prastawa

- Investigation of biological pathways involved in brain development in preterm neonates using a multivariate phenotype and sparse regression
  Michelle Krishnan, James Boardman, Matt Silver, Gareth Ball, Serena Counsell, Andrew Walley, A David Edwards and Giovanni Montana

- A Novel Atlas-based Approach to the Detection of Mouse Embryo Ventricular Septal Defects
  Xi Liang, Zhongliu Xie, Asanobu Kitamoto, Masaru Tamura, Toshihiko Shiroishi and Ramamohanarao Kotagiri

- Dopamine-Related Genetic Influences on Cognitive Flexibility
  Hans Melo, Daniel Mueller, Adam Anderson and William Cunningham

- Imaging Genomic Mapping of Tumor Volume MRI Phenotype in Glioblastoma and Correlation with the Survival and Treatment Response
  Ginu A. Thomas, Sanjay Singh, Islam Hassan, Pascal O. Zinn and Rivka R. Colen

- Imaging Genomic Biomarker Signature for MGMT Promoter Methylation Identification
  Ginu A. Thomas, Pascal O. Zinn and Rivka R. Colen

- Introduction to Imaging Genomics in Glioblastoma
  Rivka R. Colen, Ginu A.Thomas and Pascal O. Zinn

Correlation and integration of fMRI imaging and SNP data with sparse models

Yu-Ping Wang[*]

Department of Biomedical Engineering, Tulane University, New Orleans, LA,70118, USA

*corresponding author. wyp@tulane.edu

Aims. We review our recent efforts in developing sparse models for the correlation and integration of brain imaging (e.g., fMRI) and genomics (e.g., SNPs) data. Despite much of work on the integration of imaging and genomic data, there remains a lack of efficient and effective approaches to combine this complementary information (see our review [1]). Current integrative methods have not fully taken advantage of the special characteristics of imaging genomic data (e.g., inter-correlations, small sample size, group structures) and have not incorporated prior knowledge into the model design. To this end, we proposed two novel sparse model based approaches [2-4] for the correlation and integration of fMRI imaging and SNP data, resulting in better detection of risk genes and improved diagnosis of mental illnesses such as schizophrenia (SZ).

Methods. For the correlation of two imaging and genomic data sets X_1 and X_2 with dimensions $n * p$ and $n * q$, the CCA model (see **Fig.1(a)**) identifies a pair of vectors w_1 and w_2 such that $w_1^*, w_2^* = \arg\ max\ corr(X_1 w_1, X_2 w_2)$, where $corr(X_1 w_1, X_2 w_2) = w_1^T X_1^T X_2 w_2$ measures the correlation of $X_1 w_1$and $X_2 w_2$. It is very common that the number of features is greatly larger than the number of observed samples (i.e., $p, q \gg n$) in an imaging or genomic study, making the CCA impractical for use. To overcome such a difficulty, we have proposed a sparse CCA (sCCA) approach [5]to yield w_1 and w_2 sparse. The sCCA is obtained by imposing penalty term on the CCA model as follows: $P_1(w_1) \ll c_1, P_2(w_2) \ll c_2$, where P_1 and P_2 are usually taken to be penalty functions such as the l_1 norm (Lasso) or a combination of l_1 and l_2 nom (elastic net). In addition, we consider group structures [2, 5](e.g., SNPs in a gene) in the data, resulting in a sparse group CCA (sgCCA) model.

For the integration of two data sets, we propose the following integrative model[3, 4]: $Y = [\alpha_1 A_1, \alpha_2 A_2] \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \varepsilon\ = AX + \varepsilon$, where $Y \in R^{m \times 1}$ is the observation vector (phenotypes of the subjects); $A_1 \in R^{m \times n_1}$ and $A_2 \in R^{m \times n_2}$ are the measurements of two different data types (e.g., fMRI and SNPs); $A = [\alpha_1 A_1, \alpha_2 A_2] \in R^{m \times n}$; $\alpha_1 + \alpha_2 = 1$, and $\alpha_1, \alpha_2 > 0$ are the weight factors for the two types of data. $\varepsilon \in R^{m \times 1}$ is the measurement error. In order to overcome the difficulty of small sample problem, we find the approximate solution by introducing the sparsity in X, *e.g.*, $min \sum_{i=1}^{2} \|X_i\|_p$, *re-*

sulting in a sparse model based variable selection (SRVS). We use the model for biomarker selection, as illustrated in **Fig. 2**.

Results. We have tested the sgCCA model on 208 subjects containing both fMRI voxels and SNPs[2] (see **Fig.1(b)**). We detected novel genes susceptible to schizophrenia (SZ). In addition, we identified several brain regions susceptible to SZ such as superior, middle, inferior and medial frontal gyrus, inferior parietal lobule, superior and middle temporal gyrus, thalamus, parahippocampal gyrus, cingulate gyrus. The effects of these brain regions on SZ have been reported by other neuroimaging studies [6], providing additional confidence that these disease relevant brain regions may be affected by those correlated genomic variations.

We also applied the integrative model to the same 208 subjects (92 cases and 116 controls) for the selection of both fMRI imaging and SNPs biomarkers [3, 4], resulting better diagnosis of SZ. **Fig.3(a)** shows the results of identified brain regions with SRVS method in comparison with Li et al.'s sparse regression method [7] (**Fig.3 (b)**); our method tends to find regions with voxels clustered together. In addition, the models all give much higher classification ratios than that of Li et al.'s method [7] and the model with $L_{1/2}$ norm generates the highest classification ratio (**Fig.3 (c)**).

Conclusion. Our results indicate that sparse representation based models provide a powerful and flexible way for the analysis of imaging genomic data, with the following advantages: 1) they can incorporate specific features of imaging genomic data (e.g., group structures); 2) they can overcome the difficulty of analyzing imaging genomic data with small sample but larger number of features, which is often the case in practice.

Acknowledgement: *This is a joint work with Dongdong Lin (Tulane), Honabao Cao (NIH), Vince Calhoun (Mind Research Network). The research has been supported by the NIH.*

References:

[1] Y. P. Wang, "Multiscale Genomic Imaging Informatics," *Ieee Signal Processing Magazine*, vol. 26, pp. 169-+, Nov 2009.

[2] D. Lin, V. Calhoun, and Y. P. Wang, "Correspondence between fMRI and SNP data by group sparse canonical correlation analysis," *Medical Image Analysis*, Nov. 4, in press, 2013.

[3] H. B. Cao, J. B. Duan, D. D. Lin, V. Calhoun, and Y. P. Wang, "Integrating fMRI and SNP data for biomarker identification for schizophrenia with a sparse representation based variable selection method," *BMC Med Genomics*, vol. 6, Nov 11 2013.

[4] H. Cao, J. Duan, D. Lin, Y. Y. Shugart, V. Calhoun, and Y. P. Wang, "Sparse representation-based biomarker selection for schizophrenia

with integrated analysis of fMRI and SNPs," *Neuroimage*, Feb 12 2014.

[5] D. Lin, J. Zhang, J. Li, V. D. Calhoun, H. W. Deng, and Y. P. Wang, "Group sparse canonical correlation analysis for genomic data integration," *BMC Bioinformatics*, vol. 14, p. 245, 2013.

[6] M. E. Shenton, C. C. Dickey, M. Frumin, and R. W. McCarley, "A review of MRI findings in schizophrenia," *Schizophrenia Research*, vol. 49, pp. 1-52, Apr 15 2001.

[7] Y. Li, P. Namburi, Z. Yu, C. Guan, J. Feng, and Z. Gu, "Voxel selection in FMRI data analysis based on sparse representation," *IEEE Trans Biomed Eng*, vol. 56, pp. 2439-51, Oct 2009.

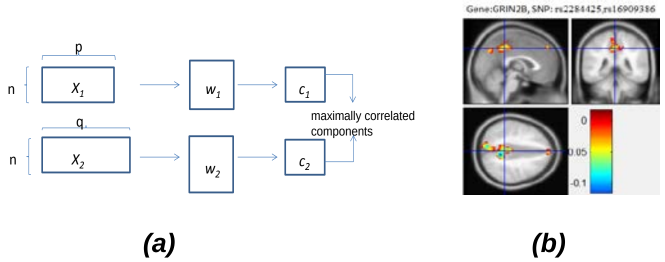

p
n
X_1
w_1
c_1
maximally correlated components
q
n
X_2
w_2
c_2
Gene:GRIN2B, SNP: rs2284425,rs16909386
0
0.05
-0.1
(a)
(b)

***Fig. 1** Identifying correspondence of two datasets X_1 and X_2 using CCA **(a)**, which is used to identify abnormal brain regions associated with genes (GRIN2B) and SNPs (rs2284425, rs16909386) **(b)**.*

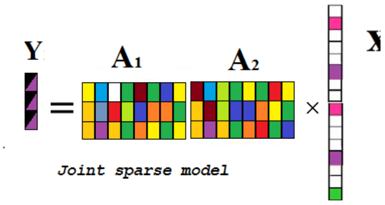


***Fig. 2** Integration of two data sets A_1 and A_2. The sparse model (bottom) can better capture joint information than joint model (top), where the correlative information is represented simultaneously by non-zero entries in a sparse vector.*

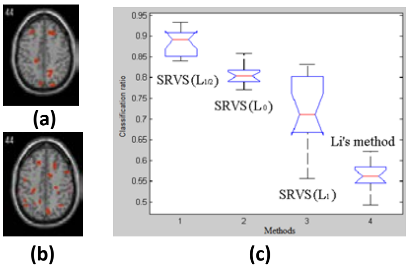


***Fig.3** Our sparse model based variable selection (SRVS) method can better identify abnormal brain region (a) than Li et al.'s method (b) (i.e., yielding more clustered regions that have been validated before), and give better accuracy of classifying SZ from healthy controls (c).*

Detection of Genes associated with multiple correlated imaging phenotypes by a sparse group-ridge low-rank regression model

Dongdong Lin[1,2], Hong-Wen Deng[2,3], Vince D. Calhoun[4,5], Yu-Ping Wang[1,2,3,*]

[1]Department of Biomedical Engineering, Tulane University, New Orleans, LA,70118, USA

[2]Center of Genomics and Bioinformatics, Tulane University, New Orleans, LA,70112, USA

[3]Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, LA, 70118, USA

[4]The Mind Research Network, Albuquerque, NM, 87131, USA

[5]Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, 87131, USA

*corresponding author. wyp@tulane.edu

Aims. Recently, more evidences on polygenicity and pleiotropy have been found in genome-wide association study of complex psychiatric diseases (e.g., schizophrenia), where multiple interacting genetic variants may affect multiple phenotypic traits simultaneously. More heritability of complex diseases is expected to be explained by these factors but few studies have been performed in imaging genetics. It is significant to identify those genetic variants (e.g., SNPs, gene) having pleiotropic effects on multiple imaging-derived quantitative traits (i.e, endophenotype). The high dimensionality of imaging and genetic datasets, however, presents a challenge for statistical analysis. Most statistical methods for pleiotropy currently focus on single loci, which may ignore group effect of genetic variants, leading to the loss of power. Several sparse regularization based models have been recently proposed to reduce the number of features in the model with an optimal feature selection criterion. However, most of them did not provide a significance test for each selected feature. In this work, we propose a new sparse group-ridge low-rank regression model (SGRLR) for exploring the pleiotropic effects of a group of genetic variants on multiple correlated endophenotypes derived from fMRI. In the method, we enforce sparse regularization to reduce the number of features and then construct an effective permutation-based statistic test to evaluate the significance of selected features (e.g., gene or gene set).

Methods. SGRLR is a sparse multivariate regression model, defined by the following formula

$$min_{\boldsymbol{C}}\ ||\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{C}||_F^2+\ \lambda_1\sum_{g=1}^{G}\left\|\boldsymbol{C}_{[g]}\right\|_{gridge}+\lambda_2||\boldsymbol{C}||_*$$

where $\boldsymbol{Y}$ is an image-derived endophenotype matrix, $\boldsymbol{X}$ is a high dimensional genomic measurement matrix which can be divided into $\boldsymbol{G}$ groups (*e.g.*, genes) and $\boldsymbol{C}$ is a coefficient matrix with each row representing the weights of individual SNPs across all endophenotypes. $\left\|\boldsymbol{C}_{[g]}\right\|_{gridge}=\left[\sum_{i\in g}||\boldsymbol{C}_i||_2\right]^2$ is a composite group ridge penalty on the *i*-th submatrix of $\boldsymbol{C}$, denoted by $\boldsymbol{C}_i$. [*g*] indicates the set of row indices (i.e. SNPs) of $\boldsymbol{C}_i$ to form the *g*-th gene. This group ridge penalty uses a lasso penalty to perform SNPs selection within each gene and ridge penalty to identify causal genes simultaneously. Both the enforcement of sparsity at SNP level and the smoothness constraint at group (e.g., gene) level can remove irrelevant SNPs within genes while consider the correlation among genes in the model. $||\boldsymbol{C}||_*$ is a low rank penalty based on nuclear-norm to account for the correlation among multiple endophenotypes by reducing the rank of $\boldsymbol{C}$. Based on the estimation of the model, a statistical test is proposed to perform significance test at both gene and gene set level. An empirical p-value for each gene or gene set can be obtained, which follows a uniform distribution due to the smoothness of ridge regression at gene level. We conducted a simulation to evaluate the performance of SGRLR in terms of the power of detecting causal genes. Then, we compared our method with other sparse multivariate models such as sparse multi-task learning methods with lasso, group lasso, and group lasso combined with low rank penalties.

Results. Fig.1 shows the comparison results using ROC curve based on the average of 50 replications where the curve of SGRLR is drawn by varying significant levels from 0 to 1. At each replication, 20 genes with 5 as casual genes were simulated. Each gene contains 10~100 SNPs generated from PLINK and the number of causal SNP in each causal gene (*Tr*) ranges from 1,2,4,6, as shown in Fig.1. SGRLR shows better performance than other sparse methods, especially when *Tr*=6, i.e., more causal SNPs are included in each causal gene. Group lasso combined with low rank penalty will increase the performance compared to that of using group lasso penalty only. However, group lasso based methods perform worse than SGRLR and lasso model, which may be due to the existence of many non-causal SNPs into causal genes. Evaluation on real data is currently ongoing.

Conclusion. Our proposed SGRLR model outperforms other multivariate sparse models in terms of ROC. SGRLR can consider correlative structure in both genetic variants and imaging traits, showing superior performance for detecting genes in imaging genetic study.

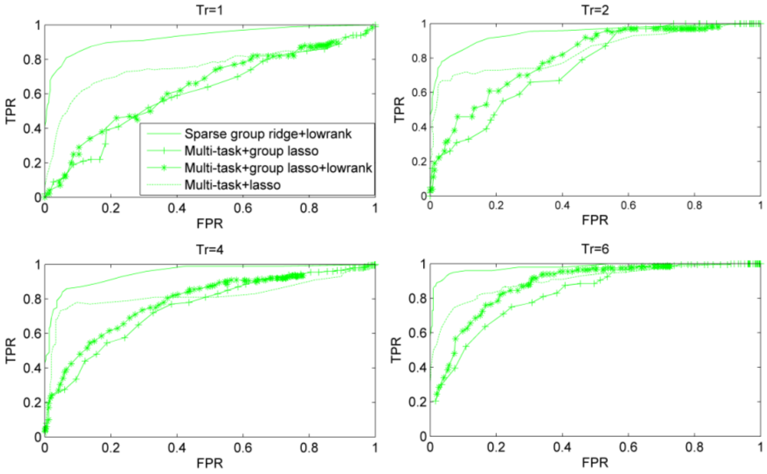


Fig. 1. ROC curves comparing SGRLR with other sparse models: sparse multi-task learning with lasso, group lasso and group lasso combined with low rank penalties.

Probabilistic Approach to Joint Modeling of Imaging and Genetics

Nematollah K. Batmanghelich[1], Adrian V. Dalca[1] Mert R. Sabuncu[2], and Polina Golland[1]

[1] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA,
[2] Martinos Center for Biomedical Imaging, Charlestown, MA
{kayhan,adalca,msabuncu,polina}@csail.mit.edu

Aims

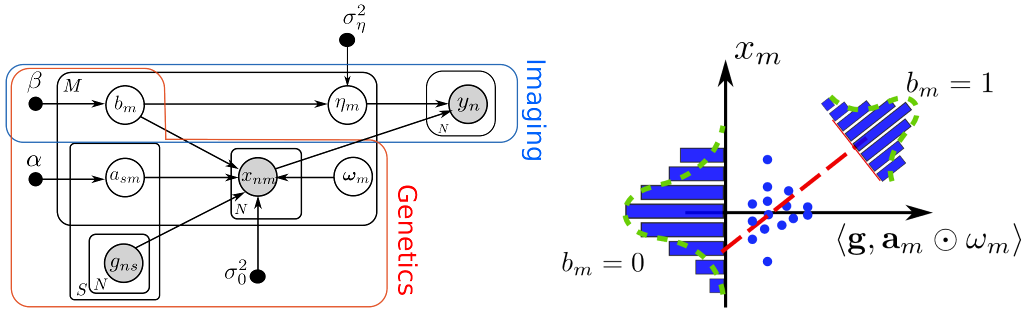
We propose a unified Bayesian framework for detecting genetic variants associated with disease while exploiting image-based features as an intermediate phenotype. Using imaging data for examining genetic associations is a growing new field, but currently the most widely used methods make sub-optimal use of those three types of measurements (*i.e.,* clinical, imaging and genotype) by performing the association test between them separately [4, 5]. In contrast, we propose a probabilistic framework to exploit the connection among all these data modalities simultaneously. Our method ultimately assigns probabilistic measures of clinical relevance to both genetic and imaging biomarkers. We derive an efficient approximate inference algorithm that handles the high dimensionality of imaging and genetic data. We also illustrate the application of the method on Alzheimer's Disease Neuroimaging Initiative (ADNI) data.

Methods

We assume that a study contains N individuals, each with three measurements: 1) genotype, $\mathbf{g}_n \in \mathbb{R}^S$, 2) clinical outcome $y_n \in \{0, 1\}$, and 3) imaging measurements, $\mathbf{x}_n \in \mathbb{R}^M$. The genotype $\mathbf{g}_n$ is the allele count from S locations on the genome. The clinical status y_n indicates Normal vs Alzheimer's. The brain imaging measurements $\mathbf{x}_n$ are volume or thickness of M brain structures, and are usually referred to as the "intermediate phenotype."

The objective is to choose a subset of the intermediate phenotypes (*i.e.,* imaging regions) which are simultaneously relevant to the clinical measurements and the genotype.

The model comprises of two regressions (see Fig.1b): one that selects a subset of imaging regions to predict the clinical phenotype y and the second regression that explains the variations of the selected imaging measurements via a sparse set of genotypes. Therefore, there are two possibilities for each brain region: selected or not (1 or 0). We employ the spike and slab model [1] to model the selected imaging regions. It assigns a binary latent variables to indicate relevance of the genotypes to a phenotype via a regression. If a brain region is not selected,



(a) (b)

Fig. 1. (a) Graphical model of conditional independence between variables. Open circles are the random variables (gray: observed, white: latent) and filled circles are the hyper-parameters. (b) Schematic of modeling in the x-node (imaging phenotype). If a region m is irrelevant to clinical phenotype y (*i.e.*, $b_m = 0$), the normal distribution explains the variation (Null). Otherwise, a regression with a latent mask $\mathbf{a}_m$ from the genotype explains the variation of imaging features inside of the region.

we assume that it has a normal distribution with zero mean and unit variance. The graphical model is illustrated in Fig.1a. Finally, our method produces two quantities that can be used to interpret the importance of the brain regions and SNPs: 1) $\mathbb{P}(b_m|D)$, the posterior probability that brain region m is relevant, and 2) $\mathbb{P}(\mathbf{a}_m|D)$, the posterior relevance of the SNPs in explaining variability in the brain region m.

Results

We apply the method to a subset of samples from the ADNI[3] dataset. For each hemisphere of the brain, cortical thickness of 34 regions spanning the whole cortex was computed in addition to volumes of 26 sub-cortical brain structures. Here, we focus on subset of previously studied SNPs [3].

Fig.2 reports posterior probability of brain regions identified as relevant to Alzheimer disease and the genotype. Fig.3b shows the relevance of SNPs averaged over 1000 draws from the posterior distribution of the brain regions. Fig.3d shows the posterior probability of the relevant SNPs for explaining variability in the entorhinal cortex, which is one the top areas of the brain to be affected in Alzheimer's Disease [2]. Fig.3a and Fig.3c respectively show the log p-value of the GLM when the clinical phenotype y and the average thickness of the entorhinal cortex are used as dependent variables.

[3] Alzheimer's Disease Neuroimaging Initiative

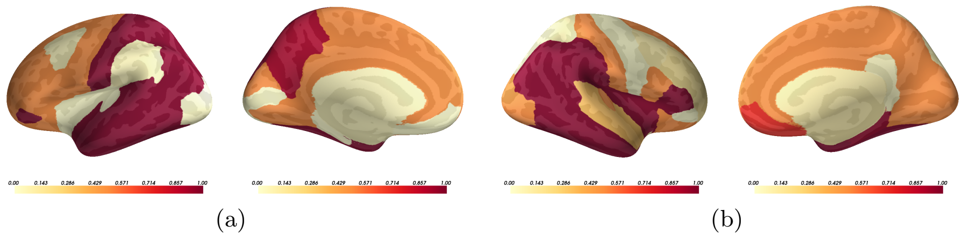


Fig. 2. (a) Posterior probability of inclusion for the brain regions on the left hemisphere in the lateral and medial views. Similarly for right hemisphere (b)

Conclusions

We propose a probabilistic approach that jointly models the clinical measurements, the intermediate phenotype and the genotype. Flexibility of the model allows to define endophenotypes other than imaging features but here we focus on the imaging measurements, namely the cortical thickness and the volume of the sub-cortical regions. The framework exploits the rich information in the image and finds the SNPs that are implicitly relevant to the disease. We derived an efficient inference algorithm that can infer relevant SNPs and answer many other interesting questions about the data.

Bibliography

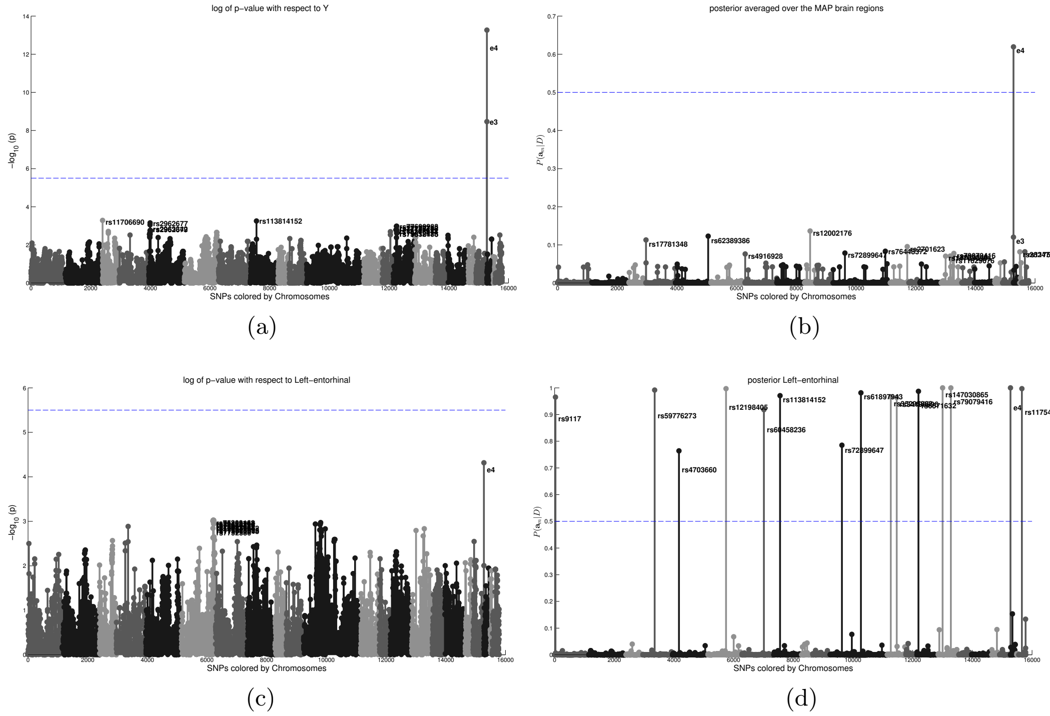
[1] Carbonetto, P., Stephens, M.: Scalable Variational Inference for Bayesian Variable Selection in Regression, and its Accuracy in Genetic Association Studies. Bayesian Analysis 7, 73–108 (2012), `http://dx.doi.org/10.1214/12-BA703`

[2] Khan, U.A., Liu, L., Provenzano, F.A., Berman, D.E., Profaci, C.P., Sloan, R., Mayeux, R., Duff, K.E., Small, S.A.: Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical Alzheimer's disease. Nature neuroscience 17(2), 304–311 (2014)

[3] Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A.L., Bis, J.C., Beecham, G.W., et al.: Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. Nature genetics (2013)

[4] Stein, J., Hua, X., Lee, S., Ho, A., et al.: Voxelwise genome-wide association study (vGWAS). Neuroimage 53(3), 1160–1174 (Nov 2010), `http://dx.doi.org/10.1016/j.neuroimage.2010.02.032`

[5] Vounou, M., Janousova, E., Wolz, R., Stein, J.L., et al.: Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. Neuroimage 60(1), 700–716 (Mar 2012)



(a)

(b)

(c)

(d)

Fig. 3. (a) Negative log p-value of the GLM when the diagnosis y is used as the dependent variable. (b) Posterior relevance of SNPs averaged over 1000 draws from the posterior distribution of the brain regions. Horizontal lines are 0.05 after Bonferroni corrections (left figures) and 0.5 for the posterior probabilities (right figures). On average the results of (a) and (b) are very similar and only `APOE` passes the significant levels. However, our probabilistic model provides the posterior relevance $\mathbb{P}(b_m|D)$ corresponding for each region as well. We can then conduct a post-hoc analysis by computing the posterior probability of SNPs for the relevant regions (*i.e.,* for regions with $\mathbb{P}(b_m|D) > 0.5$). The posterior probability of SNPs $\mathbb{P}(\mathbf{a}_m|D)$ for region m quantifies the relevance of each SNP to that region. For example, (d) shows the posterior probability of each SNP for the left entorhinal cortex which has $\mathbb{P}(b_m|D) > 0.5$; many SNPs exhibit a posterior higher than 0.5 which makes them implicitly relevant to the disease. (c), negative log p-value of GLM using average thickness of the entorhinal cortex as the dependent variable; no SNPs passes the significant threshold of 0.05 after Bonferroni correction.

Detecting Gene-Environment Interactions via a Kernel Machine Method

Tian Ge[1,2], Thomas E. Nichols[3], Debashis Ghosh[4], Elizabeth C. Mormino[5], Jordan W. Smoller[2,6,†], Mert R. Sabuncu[1,7,†], and for the Alzheimers Disease Neuroimaging Initiative[⋆]

[1] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital / Harvard Medical School, Charlestown, MA 02129, USA

[2] Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA

[3] Department of Statistics & Warwick Manufacturing Group, The University of Warwick, Coventry CV4 7AL, UK

[4] Department of Statistics, The Pennsylvania State University, PA 16802, USA

[5] Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

[6] Broad Institute of MIT and Harvard, Cambridge, MA 02138, USA

[7] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

† Contributed equally

tge1@nmr.mgh.harvard.edu; msabuncu@nmr.mgh.harvard.edu

1 Aims

To date, imaging genetics studies have mostly focused on discovering isolated gene effects, typically ignoring potential interactions with environmental variables. However, identifying significant gene-environment (G×E) interactions is critical in revealing the true relationships among genetic, environmental, and phenotypic variables, and shedding light on disease mechanisms. Here we present a powerful and flexible method for detecting G×E interactions.

2 Methods

We propose a semiparametric kernel machine based approach to detect G×E interactions. The kernel machine framework has been widely used in association studies between a collection of single nucleotide polymorphisms (SNPs), and complex diseases or imaging phenotypes [1–5]. To jointly model the genetic and

[⋆] Data used in preparation of this article were obtained from the Alzheimers 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.

environmental variables, and their interactions, we extend the classical kernel machine model [2], and include three appropriately selected kernels in the model; one for genetic variants, one for environmental factors, and a third one, which is the Hadamard product of the genetic and environmental kernels, for the interaction effect. An identical-by-state (IBS) genetic kernel provides a biologically-informed way to capture epistasis in a set of SNPs, and model their joint effect on the phenotype. Examining collective contribution of SNPs can provide improved reproducibility, better biological interpretability, and increased power relative to univariate methods. A linear environmental kernel allows for jointly modeling the effect of multiple environmental variables. By using a connection to linear mixed effects models, the interaction effect can be efficiently tested by a variance component score test [2, 6].

We apply the method to data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), to detect the interaction effect between 20 candidate late-onset Alzheimer's disease (AD) risk genes, as identified by a recent meta-analysis of genome-wide association studies (GWAS) with 74,046 individuals [7], and cerebral amyloid beta (Aβ) deposition, as measured by the global standardized uptake value ratio (SUVR) index value in florbetapir F18 (AV45) positron emission tomography (PET) images, on an AD probability score derived from structural brain Magnetic Resonance Imaging (sMRI) scans using the Relevance Voxel Machine (RVoxM) algorithm [8]. The AD score quantifies the likelihood of AD-associated atrophy based on the pattern of cortical thinning, and is correlated with the disease stage.

3 Results

Gene *BIN1* ($p = 0.0005$) was identified to have significant interaction with the Aβ levels (Table 1), after controlling for age, gender, education, and diagnosis (healthy control, mild cognitive impairment, AD).

4 Conclusions

We have presented a kernel machine based approach for detecting G×E interactions, which offers a flexible framework to model epistatic effects, accommodate multiple environmental factors, and test for interactions between the two sets of variables, producing more interpretable and powerful results compared to classical univariate approaches. The gene *BIN1* we identified was thought to be a strong genetic determinant of AD susceptibility. The expression of *BIN1* was found to increase AD risk by modulating tau pathology [9, 10], indicating possible mechanisms of the progression of late-onset AD in the context of Aβ accumulation.

References

1. Li, S., Cui, Y.: Gene-centric gene–gene interaction: A model-based kernel machine method. Ann Appl Stat. 6, 1134–1161 (2012)

Table 1. The 20 candidate risk genes for late-onset Alzheimer's disease identified by [7], the final number of SNPs located on them, and the p-value for the interaction effect between the candidate genes and the Aβ level on logit transformed AD probability scores, using the proposed kernel machine method. p-values that survive multiple testing corrections are highlighted in bold.

Gene	**Chr**	**# of SNPs**	***p*-value**	**Gene**	**Chr**	**# of SNPs**	***p*-value**
ABCA7	19	78	0.7508	*EPHA1*	7	38	0.9800
APOE	19	4	0.1868	*FERMT2*	14	156	0.2127
BIN1	2	203	**0.0005**	*INPP5D*	2	463	0.3705
CASS4	20	77	0.9992	*MEF2C*	5	194	0.9868
CD2AP	6	297	0.5681	*MS4A6A*	11	13	0.6262
CD33	19	14	0.1031	*PICALM*	11	266	0.9682
CELF1	11	73	0.9857	*PTK2B*	8	318	0.7043
CLU	8	27	0.4456	*SLC24A4*	14	638	0.9966
CR1	1	147	1.0000	*SORL1*	11	206	0.9604
DSG2	18	51	0.6429	*ZCWPW1*	7	34	0.9941

2. Liu, D., Lin, X., Ghosh, D.: Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics. 63, 1079–1088 (2007)
3. Kwee, L.C., Liu, D., Lin, X., et al.: A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet. 82, 386–397 (2008)
4. Wu, M.C., Kraft, P., Epstein, M.P., et al.: Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 86, 929–942 (2010)
5. Ge, T., Feng, J., Hibar, D.P., et al.: Increasing power for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures. NeuroImage. 63, 858–873 (2012)
6. Lin, X.: Variance component testing in generalised linear models with random effects Biometrika. 84, 309–326 (1997)
7. Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., et al.: Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 45, 1452–1458 (2013)
8. Sabuncu, M.R., Van Leemput, K.: The relevance voxel machine (RVoxM): A self-tuning bayesian model for informative image-based prediction. IEEE Trans Med Imaging. 31, 2290–2306 (2012)
9. Kingwell, K.: Alzheimer disease: BIN1 variant increases risk of Alzheimer disease through tau. Nat Rev Neurol. 9, 184 (2013)
10. Chapuis, J., Hansmannel, F., Gistelinck, M., et al.: Increased expression of BIN1 mediates Alzheimer genetic risk by modulating tau pathology. Mol Psychiatry. 18, 1225–1234 (2013)

Fast Heritability Analysis Using Genome-Wide Data via Kernel Machines

Tian Ge[1,2], Thomas E. Nichols[3], Avram J. Holmes[4], Phil H. Lee[2], Joshua L. Roffman[5], Randy L. Buckner[1], Mert R. Sabuncu[1,6,†], and Jordan W. Smoller[2,7,†]

[1] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital / Harvard Medical School, Charlestown, MA 02129, USA
[2] Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA
[3] Department of Statistics & Warwick Manufacturing Group, The University of Warwick, Coventry CV4 7AL, UK
[4] Department of Psychology, Yale University, New Haven, CT 06520, USA
[5] Department of Psychiatry, Massachusetts General Hospital / Harvard Medical School, Boston, MA 02114, USA
[6] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[7] Broad Institute of MIT and Harvard, Cambridge, MA 02138, USA
† Contributed equally
tge1@nmr.mgh.harvard.edu; jsmoller@hms.harvard.edu

1 Aims

Reliable and rapid tools to estimate the heritability of imaging measurements are necessary for prioritizing brain phenotypes for genetic association analyses [1]. Classical estimates of heritability require twin or pedigree data, which are costly and difficult to acquire. Genome-wide Complex Trait Analysis (GCTA) [2] can provide heritability estimates without twin data, but is computationally demanding and thus can only be applied to a handful of a priori selected measurements or regions of interest (ROIs). Here we present a flexible and computationally efficient method for high-dimensional heritability analysis, using whole-genome single nucleotide polymorphism (SNP) data from unrelated individuals.

2 Methods

We propose an efficient method for quantifying the statistical significance of heritability using genome-wide data. The approach relies on a variance component score test for linear mixed effects models [3, 4], and has a strong connection to the nonparametric kernel machine methods for SNP-set based association studies [5–7]. Each entry of the kernel matrix, also known as the genetic relationship matrix (GRM) in GCTA, is a measure of the genetic similarity between pairs of individuals computed from all SNPs over the genome to match the definition of narrow-sense heritability. The fast non-iterative score test makes computationally demanding analyses, such as a voxel-/vertex-wise test for significant heritability or permutation test, possible.

We apply our approach to the structural images and whole-genome SNP data from 1,464 unrelated young healthy adults (18-35 years old) with non-Hispanic Caucasians of European ancestry, as part of the Harvard/MGH Brain Genomics Superstruct Project (GSP) [8].

3 Results

An ROI heritability analysis of the average thickness and surface area measurements within cortical regions defined by the Desikan-Killiany atlas [9] shows that the score test produces almost identical p-values as GCTA (Fig. 1), but is thousands of times faster (10 mins versus 30 ms), making high-dimensional heritability mapping and permutation inferences possible. The vertex-wise p-value map of the heritability of cortical thickness shows similar pattens with previous observations in twin studies (Fig. 2, upper panel) [10–13]. In our analyses, surface area measurements are less heritable than cortical thickness in general, and show distinct spatial profiles (Fig. 2, lower panel). Surface-based clustering of vertices on the cortical thickness map using a cluster-forming threshold $p = 0.001$ identifies four clusters with FWE-corrected significance using permutation-based cluster-size inferences (Fig. 3). Posthoc heritability analyses on average cortical thickness measurements within theses significant clusters show high heritability of these cortical regions (Table 1).

4 Conclusions

We have proposed a fast and accurate statistical test for significant heritability using genome-wide SNP data from unrelated individuals, and an accompanying permutation procedure that can produce accurate and flexible permutation inferences for arbitrary statistics of interest. Our proposed approach has the potential for large-scale heritability screening, three-dimensional heritability profiles construction, and optimally choosing brain phenotypes under genetic control.

References

1. Thompson, P.M., Ge, T., Glahn, D.C., et al.: Genetics of the connectome. NeuroImage. 80, 475–488 (2013)
2. Yang, J., Lee, S.H., Goddard, M.E., et al.: GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 63, 76–82 (2011)
3. Lin, X.: Variance component testing in generalised linear models with random effects Biometrika. 84, 309–326 (1997)
4. Liu, D., Lin, X., Ghosh, D.: Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics. 63, 1079–1088 (2007)

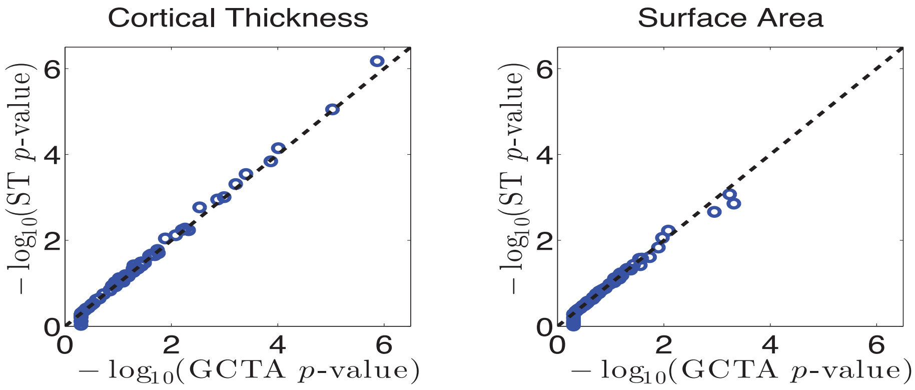


Fig. 1. The excellent concordance between the score test and GCTA.

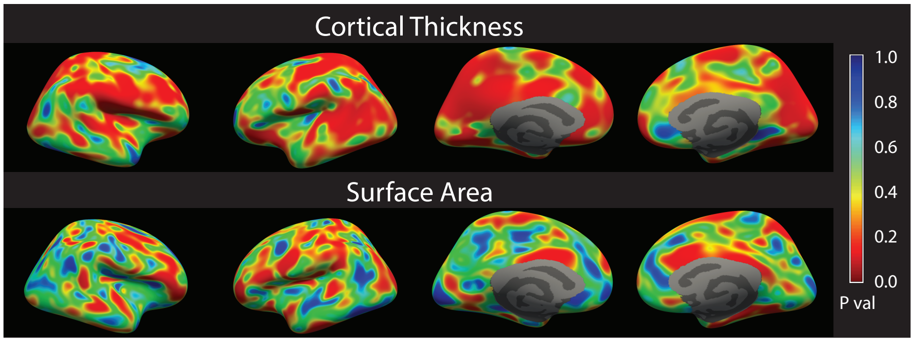


Fig. 2. Vertex-wise p-value maps of the heritability of cortical thickness and cortical surface area, constructed by the score test.

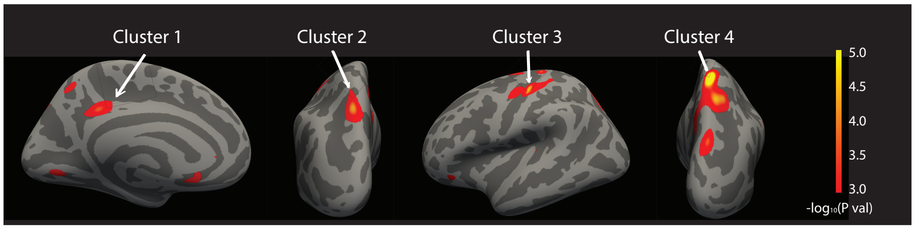


Fig. 3. Clusters identified on the spatial heritability map of cortical thickness with a cluster-forming threshold $p = 0.001$. Four clusters are identified as family-wise error corrected (FWEc) significant in size.

Table 1. Statistics of significant clusters identified on the spatial heritability map of cortical thickness with a cluster-forming threshold $p = 0.001$. The number of vertices, family-wise error corrected (FWEc) p-value of the cluster size, and the posthoc heritability estimate of the average cortical thickness within each cluster are shown.

Cluster ID	# Vertices	FWEc cluster-size p-val	GCTA $\hat{h}^2$	GCTA SE	GCTA p-val
Cluster 1	603	0.048	0.982	0.245	3.42×10^{-5}
Cluster 2	796	0.016	0.905	0.236	4.03×10^{-5}
Cluster 3	1841	0.002	1.000	0.249	2.19×10^{-6}
Cluster 4	2656	0.001	1.000	0.240	1.18×10^{-7}

5. Kwee, L.C., Liu, D., Lin, X., et al.: A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet. 82, 386–397 (2008)
6. Wu, M.C., Kraft, P., Epstein, M.P., et al.: Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 86, 929–942 (2010)
7. Ge, T., Feng, J., Hibar, D.P., et al.: Increasing power for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures. NeuroImage. 63, 858–873 (2012)
8. Holmes, A.J., Lee, P.H., Hollinshead, M.O., et al.: Individual differences in amygdala-medial prefrontal anatomy link negative affect, impaired social functioning, and polygenic depression risk. J. Neurosci. 32, 18087–18100 (2012)
9. Desikan, R.S., Ségonne, F., Fischl, B., et al.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest NeuroImage. 31, 968–980 (2006)
10. Eyler, L.T., Chen, C., Panizzon, M.S., et al.: A comparison of heritability maps of cortical surface area and thickness and the influence of adjustment for whole brain measures: a magnetic resonance imaging twin study. Twin Res Hum Genet. 15, 304–314 (2012)
11. Lenroot, R.K., Schmitt, J.E., Ordaz, S.J., et al.: Differences in genetic and environmental influences on the human cerebral cortex associated with development during childhood and adolescence. Hum Brain Mapp. 30, 163–174 (2009)
12. Joshi, A.A., Lepore, N., Joshi, S.H., et al.: The contribution of genes to cortical thickness and volume. Neuroreport. 22, 101 (2011)
13. Yoon, U., Fahim, C., Perusse, D., et al.: Lateralized genetic and environmental influences on human brain morphology of 8-year-old twins. NeuroImage. 53, 1117–1125 (2010)

Longitudinal 3D MR Spectroscopic Imaging of 2-Hydroxyglutarate in patients with mutant IDH1 glioma undergoing radiochemotherapy

Ovidiu C. Andronesi[1], Franziska Loebel[2], Wolfgang Bogner[3], Malgorzata Marjanska[4], Elizabeth Gerstner[5], Andrew S. Chi[5], Tracy T. Batchelor[5], Daniel P. Cahill[2] and Bruce R. Rosen[1]

[1]Martinos Center for Biomedical Imaging, Dept. of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114; [2]Dept. of Neurosurgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114; [3]MR Center of Excellence, Dept. of Radiology, Medical University Vienna, 1090 Vienna, Austria; [4]Center for Magnetic Resonance Research, University of Minnesota, Minneapolis, MN 55455; [5]Pappas Center of Neuro-Oncology, Dept. of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114.

Purpose

The hallmark metabolic alteration of mutant IDH gliomas is the production of 2-hydroxyglutarate (2HG) [1], and high levels of this metabolite may play a central role in downstream effects of gene mutation [2]. Hence, 2HG may be an ideal biomarker for both diagnosing IDH mutations and monitoring response to treatment. 2HG can be measured in-vivo by magnetic resonance spectroscopy [3-6] and there is significant interest in developing methodology that performs reliably in patients. Here we present results obtained with a new 3D MR spectroscopic imaging (MRSI) sequence that maps reliably 2HG over the entire volume of the tumor during treatment.

METHOD AND MATERIALS

A robust and efficient 3D MRSI sequence for 2HG imaging was newly developed by integrating three highly optimized modules: i) J-difference spectral editing MEGA-LASER [5], ii) spiral spectroscopic imaging, and iii) real-time motion and shim correction. J-difference spectral editing can disambiguate the detection of brain metabolites such as GABA, glutamate and glutamine (Glx), and 2HG by removing overlapping signals. However, difference methods are susceptible to subtraction errors caused by subject movement and scanner instability. Using a double-echo EPI volume navigator we performed real-time correction of motion, update dynamically the shims and scanner frequency, and reacquire the excitations that are corrupted [7]. 3D brain coverage was obtained with a weighted stack of spirals. The acquisition parameters of the 3D MRSI sequence were: TR=1.6s, TE=68ms, FOV=200x200x200 mm^3, 20mm isotropic voxels, acquisition matrix 10x10x10 zero-filled to 16x16x16, NA=20, acquisition time TA=9:55 min:s. Spectra were fitted with LCModel software

[9] and metabolic maps were obtained from the fitted signal. All experiments were performed on a whole-body 3T MR scanner. 3D MRSI was performed in 9 patients with mutant IDH1 gliomas (WHO grades II-IV) who were consented with an approved IRB protocol. In all patients a baseline scan was done before starting adjuvant treatment. Adjuvant treatment included radiotherapy and/or chemotherapy. The post-treatment scan was done within 1-3 months after end of radio/chemotherapy.

RESULTS

Detectable levels of 2HG were measured in all patients that did not have gross total resection of tumor. 3D metabolic maps were obtained for 2HG and several other important metabolites for assessing brain tumors, such as total choline (Cho), N-acetyl-aspartate (NAA), glutamate and glutamine (Glx), and lactate (Lac). Four patients had marked decrease (30-50%) in the levels of 2HG and the remainder showed 10-20% reduction of 2HG (Figure 1).

CONCLUSION

We demonstrate for the first time that 3D imaging of 2HG is clinically feasible in patients with IDH1 mutated gliomas. Quantification of 2HG levels in a cohort of mutant IDH glioma patients shows measurable changes during treatment. 3D mapping of 2HG and other metabolites is important to capture tumor heterogeneity and reduce variability in longitudinal studies. 2HG imaging could be used to differentiate true-/pseudo-response and true-/pseudo-progression in mutant IDH glioma patients.

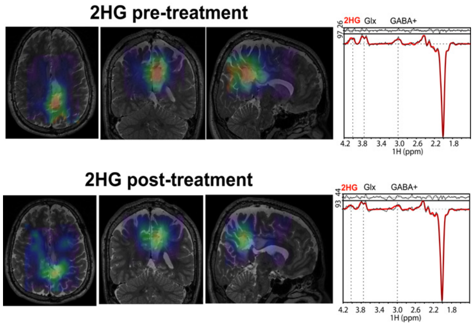


Fig. 1. 2HG metabolic maps obtained longitudinally during treatment in a mutant IDH glioma patient with a navigated adiabatic spiral 3D MRSI sequence. 3D maps of 2HG provide selectivity and specificity for the spatial extent of tumor. Marked decrease (40%) of 2HG levels are found post-treatment compared to pre-treatment levels in this patient. Maps are scaled to the same intensity levels. Spectra are shown in the right most figures with the 2HG signal indicated at 4ppm.

References:

[1] Dang L et al, Nature 462:739-52 (2009); [2] Turcan S et al, Nature 483:479-83, (2012); [3] Pope WB et al, J Neuroonc. 107:197-205 (2012); [4] Choi C et al, Nature Med 18:624-29 (2012); [5] Andronesi OC et al, Science Transl Med 4:116ra4 (2012); [6] Choi C et al, Proc ISMRM 2013, #509; [7] Bogner W et al, Neuroimage, 88C:22-31 (2013); [8] Andronesi OC et al, JMR 203:283-93 (2010); [9] Provencher S, MRM 30:672-79 (1993);

Acknowledgments: Funding 1K22CA178269-01 (NCI/NIH), Burroughs-Wellcome Career Award, 1P50CA165962-01, Harvard-MIT Bridge Project. **Contact:** Ovidiu Andronesi, ovidiu@nmr.mgh.harvard.edu.

Hierarchical clustering of whole genome sequence data for forecasting age of Alzheimer's diagnosis

Rachel A. Yotter, Xiao Da, Bilwaj Gaonkar, Roman Filipovych, Christos Davatzikos,
for the Alzheimer's Disease Neuroimaging Initiative

Center for Biomedical Image Computing and Analytics, University of Pennsylvania,
Philadelphia, PA

1 Aims

We sought to forecast the age of Alzheimer's disease (AD) diagnosis using a polygenic score from whole genome sequence (WGS) data. We first validated a clustering algorithm to reduce data dimensionality by ~60-fold, using the highly predictive APOE gene. We then identified additional genes that were predictive of AD onset. These genes were then combined in a polygenic model to forecast AD onset. In non-AD subjects, we sought to identify structural, lifestyle, and neuropsychological differences in subjects that delayed genetically predetermined AD onset by at least three years.

2 Methods

443 Caucasian WGS participants were selected from the ADNI database [1]. Of these subjects, 101 had obtained a clinical diagnosis of AD (Table 1). Data included genotyping data, demographic and neuropsychological data, APOE genotype, diagnosis information, a measure of AD-like atrophy (SPARE-AD) [2], and voxel-wise approximations of atrophy from MRI.

WGS data for 39 genes were extracted [3-4]. We clustered WGS data for each gene by first performing hierarchical clustering using Manhattan distance, then formed clusters using the default parameters in the dynamic tree cut package [5]. To validate the clustering approach, we compared it to forecasts made from the APOE allele status or raw WGS data. All forecasts used 10-fold cross-validation and the same subgroups. We then identified other genes by predicting residual error after forecasting with the APOE gene.

We used our polygenic model to forecast age of AD diagnosis in non-AD subjects. Subjects who had delayed AD by three or more years were compared to an age-matched non-delayed group.

3 Results

Using clustered WGS data with support vector regression, we achieved a much higher correlation than that achieved from APOE allele status (Table 2; Figure 1). We then identified the ABCG5 gene to be independently predictive of age of AD diagnosis (Table 3). The final polygenic model had a typical error of 6.89 years.

Delayed non-AD individuals exhibited many AD-like features (Table 4). However, they have significantly lower BMI that is difficult to explain from pre-AD weight loss alone [6] (Figure 2). The delayed group had significantly more atrophy bilaterally in the cingulate cortices and temporal gyri, as well as in the uncus, insula, and precentral gyrus in the right hemisphere.

4 Conclusions

We have developed a new approach to cluster WGS data, and have successfully achieved a low error in forecasting the age of AD diagnosis. Non-AD subjects who had delayed AD onset exhibited significantly different characteristics than an age-matched non-delayed group. This result suggests that a lower BMI may be protective.

Table 1. Group demographics (mean ± standard deviation [range])

	All (N=443)	**AD (N=101)**	**non-AD (N=342)**
Age (years)	73.19 ± 7.16 [55 – 91]	74.56 ± 7.79 [55 – 90] *	72.78 ± 6.96 [55 – 91] *
Sex	246 males, 197 females	65 males, 36 females	181 males, 161 females
Education (years)	16.3 ± 2.6 [9 – 20]	16.0 ± 2.8 [9 – 20]	16.4 ± 2.6 [9 – 20]
APOE ε4	196 positive (44%)	67 positive (66%) **	129 positive (38%) **
MMSE	27.68 ± 2.34 [19 – 30]	25.44 ± 2.64 [20 – 30] **	28.34 ± 1.77 [19 – 30] **
Modified ADAS-Cog	11.02 ± 6.96 [1 – 38]	19.89 ± 6.05 [9 – 38] **	8.39 ± 4.68 [1 – 26] **
SPARE-AD	-0.48 ± 1.11 [-2.14 – 1.96]	0.82 ± 0.71 [-1.5 – 1.96] **	-0.87 ± 0.88 [-2.14 – 1.79] **
BMI (kg/m^2)	26.95 ± 4.75 [15 – 51]	25.65 ± 4.09 [15 – 43] **	27.34 ± 4.87 [19 – 51] **
Systolic Blood Pressure (mm Hg)	135.5 ± 16.8 [83 – 190]	134.6 ± 16.2 [100 – 179]	135.8 ± 17.0 [83 – 190]
Diastolic Blood Pressure (mm Hg)	74.3 ± 9.9 [49 – 100]	73.3 ± 9.8 [49 – 98]	74.6 ± 9.9 [50 – 100]

* $p < 0.05$ ** $p < 0.01$for AD and non-AD groups

Table 2. Forecasting age at which clinical diagnosis of AD is received (n=101).

Data	**Method**	**Correlation**	**p-value**
APOE e4 positive (binary)	*Linear regression*	0.165	0.09
APOE e4 count	*Linear regression*	0.254	0.01
APOE e3/e4 counts	*Linear regression*	0.339	0.0005
WGS data	*Linear regression*	0.380	8.9e-05
WGS data	*Support vector machine*	0.392	5.1e-05
Clustered WGS data	*Linear regression*	0.423	1.1e-05
Clustered WGS data	*Support vector machine*	0.485	2.8e-05
Clustered APOE + ABCG5	*Support vector machine*	0.538	6.3e-09

Table 3. Group demographics (mean ± standard deviation [range]). The delayed group was at least three years older than their forecasted age of AD diagnosis, still without a diagnosis. The non-delayed group was matched for age and gender to the delayed group.

	Delayed (N=69)	**Not delayed (N=71)**
Age (years)	80.04 ± 4.9 [70 – 91]	78.65 ± 2.73 [75 – 86]
Sex	46 males, 23 females	37 males, 34 females
Education (years)	16.2 ± 2.9 [9 – 20]	16.3 ± 2.4 [11 – 20]
APOE ε4	47 positive (68%) **	10 positive (14%) **
Forecasted AD onset (years)	72.3 ± 4.5 [64.7 – 83.7] **,‡	80.9 ± 4.6 [73.7 – 100.2] **,‡
MMSE	27.49 ± 2.35 [19 – 30] *	28.41 ± 1.74 [21 – 30] *
Modified ADAS-Cog	11.33 ± 5.02 [2 – 25] **,†	7.97 ± 4.27 [1 – 23] **,†
SPARE-AD	-0.29 ± 0.99 [-2.1 – 1.5] *	-0.64 ± 0.84 [-1.9 – 1.8] *
BMI (kg/m^2)	25.44 ± 3.14 [19 – 34] **,†	27.45 ± 3.84 [19 – 40] **,†
Systolic Blood Pressure (mm Hg)	137.0 ± 19.1 [83 – 187]	140.1 ± 16.8 [102 – 178]
Diastolic Blood Pressure (mm Hg)	75.2 ± 9.1 [51 – 90] *,†	71.8 ± 10.5 [57 – 98] *,†

unadjusted: * $p < 0.05$; ** $p < 0.01$; adjusted for APOE: † $p < 0.05$; ‡ $p < 0.01$

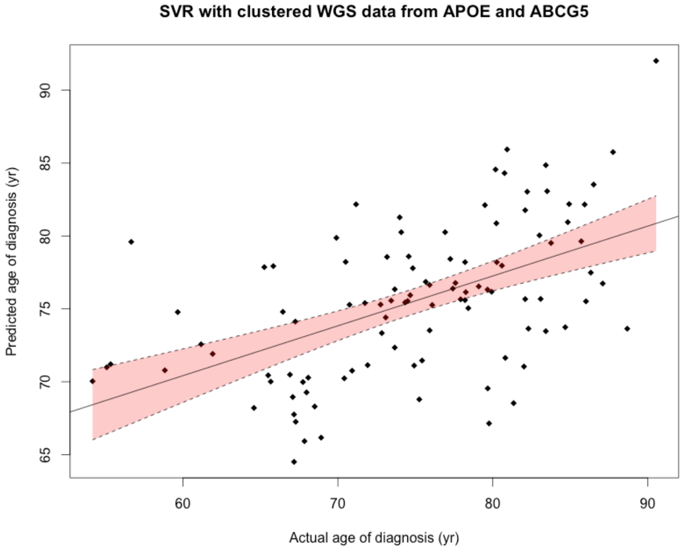


Fig. 1. Clustered WGS data from the ABCG5 is significantly predictive of the residual error after forecasting by APOE. A forecast using a combination of APOE and ABCG5 clustered data provides the most accurate forecast.

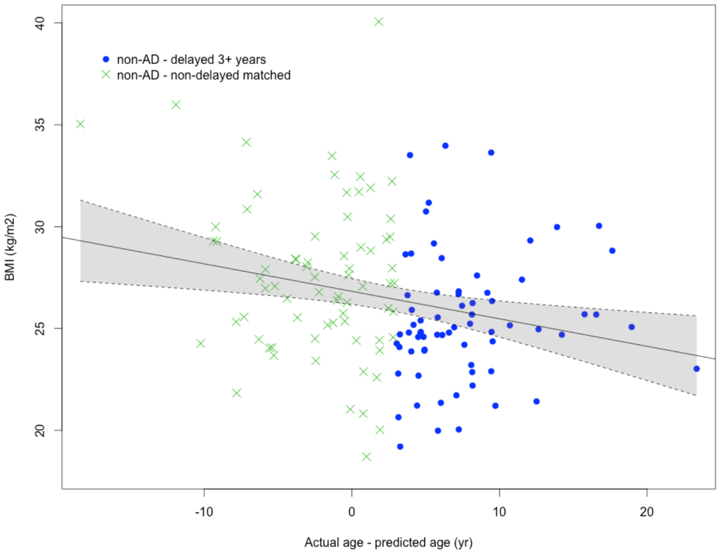


Fig. 2. Non-AD subjects who reached 3 years or longer after their genetically predicted age of AD diagnosis without obtaining a diagnosis had significantly lower BMI than an age- and gender-matched non-AD group. This relationship held even after correcting for APOE e4 allele counts.

References

1. http://www.loni.ucla.edu/ADNI; accessed 14 December 2012.
2. Davatzikos, C., Xu, F., An, Y., Fan, Y., Resnick, S. 2009. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. Brain 132(8), 2026-35.RAVENS.
3. Pico, A., Salomonis, N., Hanspers, K., Kelder, T., Conklin, B., Evelo, C., Willighagen, E., Kutmon, M., Sklar, S. Statin Pathway. http://www.wikipathways.org/index.php/Pathway:WP430; accessed 2 May 2014.
4. AlzGene – Top Results. 2011. http://www.alzgene.org/; access 2 May 2014.
5. Langfelder, P., Zhang, B., Horvath, S. 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24(5), 719-20. doi:10.1093/bioinformatics/btm563.
6. Johnson, D.K., Wilkins, C.H., Morris, J.C. 2006. Accelerated weight loss may precede diagnosis in alzheimer disease. Archives of Neurology 63(9), 1312-7. doi:10.1001/archneur.63.9.1312.

Genetic Determinants of Acute Cerebral Infact Volume:
Results from a Preliminary Genome-Wide Association Study

Lisa Cloonan, BA[1]; Kelsey Shideler, MA[1]; Cathy R Zhang, MA[1]; Adriana Perilla, MA[1]; Allison Kanakis, MD[1]; Kaitlin Fitzpatrick, BSc[1]; Natalia S Rost, MD, MPH[1]

[1] Department of Neurology, Massachusetts General Hospital, Harvard Medical School

Abstract. Introduction: Ischemic stroke is the leading cause of adult disability in the United States and the second cause of mortality worldwide. Severity of stroke is closely linked to the size of cerebral infarction, which also contributes to poor post-stroke outcomes.

Aim: We sought to determine the common genetic variants associated with acute cerebral infarct volume detected on diffusion weighted imaging (DWI) in a prospectively collected, hospital-based acute ischemic stroke (AIS) cohort.

Methods: AIS patients >18 years of age admitted to Massachusetts General Hospital Emergency Department, with an admission brain MRI scan and blood sample donated for genetic analysis, were included in this analysis.

Neuroimaging Analysis

Natural log-transformed DWI infarct volumes were measured using a validated semi-automated method. The hyperintense signal of an acute infarct on DWI was cross-referenced with the hypointensity on the apparent diffusion coefficient (ADC) sequence. An intersection between the region of interest (ROI) corresponding to the infarct outline on DWI and the intensity threshold ROI matching the DWI hyperintensity was manually corrected by an expert operator to generate a DWI volume (DWIV) in a final step of the analysis (Figure 1).

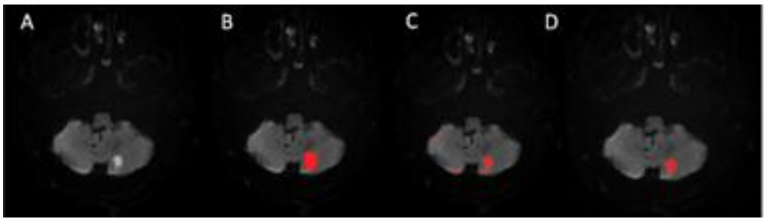
Genetic Analysis

We conducted a genome wide association study (GWAS) on 593 subjects with AIS and available DWIV. Standard per single nucleotide polymorphism (SNP) and per subject genotyping QC measures were implemented and unobserved SNPs were imputed. An association analysis of ln(DWIV) was adjusted for age, sex and principal components (PC) 1 and 2. GWAS threshold for significance was set at a nominal p-value $<5x10^{-5}$, given the pilot nature of this analysis.

Results: The mean age of the cohort was 65 years (±15) and 66% were men and 90% were white. The mean ln (DWIV) was 0.97 (±1.8). The QQ plot demonstrated genomic inflation rate < 10% (Figure 2). There were 71 SNPs associated with DWIV at $p<5\ x10^{-6}$ (Figure 3).

Conclusions: The preliminary genome-wide analysis of acute cerebral infarct volume, measured on clinical MRI of patients with AIS as DWIV, demonstrated 71 common SNPs that were associated with DWIV at a nominal genome-wide significance threshold. Future studies are warranted to replicate and validate these genetic loci.

Fig. 1. DWI cerebral infarct semi-automated analysis process



A. DWI sequence B. Infarct Outline C. Intensity Threshold D. Manually Correct Infarct

Fig. 2. QQ plot for natural log transformed DWI cerebral infarct volumes

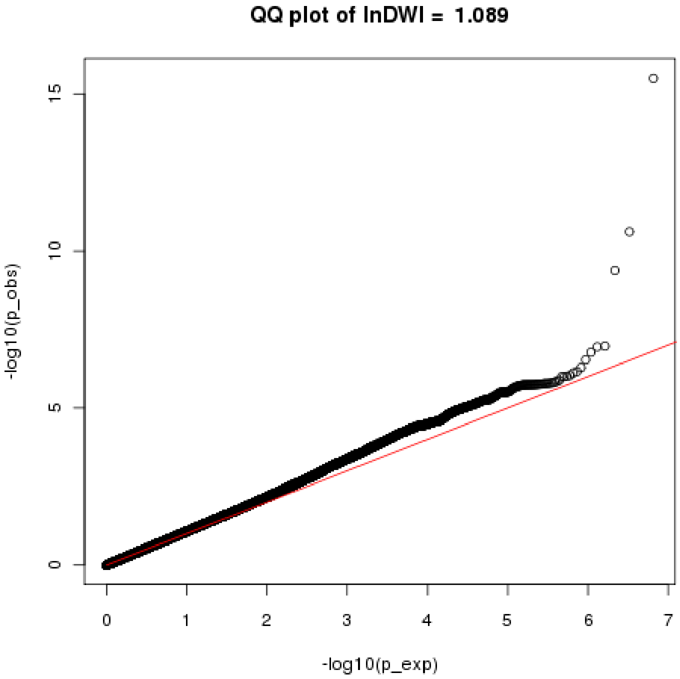
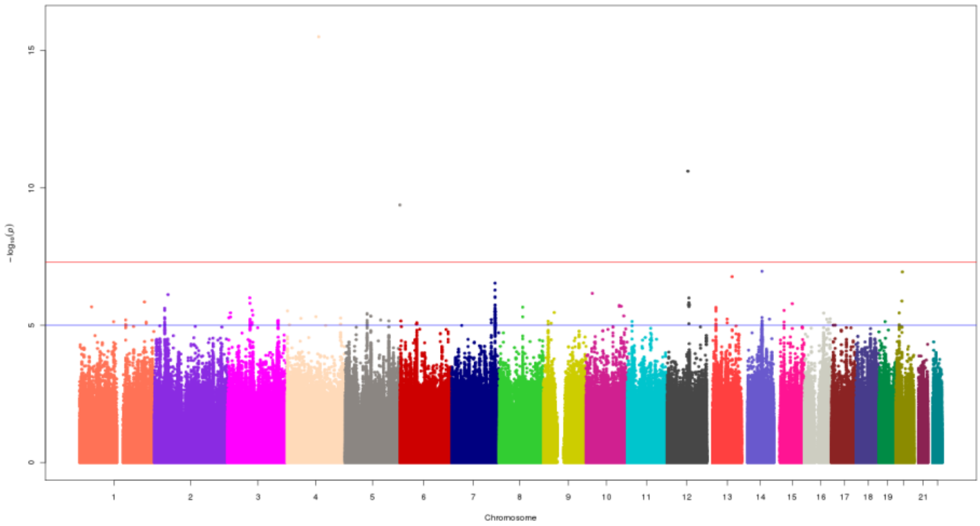


Fig. 3. Manhattan plot for the association results of DWI infarct volume



Differential Effect of 17q25 Locus on White Matter Hyperintensity Volume in Patients with Ischemic Stroke

Cathy R Zhang, MA[1]; Lisa Cloonan, BA[1]; Adrian Dalca, MS[2]; Ramesh Sridharan, MS[2]; Kaitlin Fitzpatrick, BSc[1]; Allison Kanakis, MD[1]; Alison M Ayres, BA[1]; Jonathan Rosand, MD, MSc, MSc[1, 3]; Ona Wu, PhD[1, 4]; Polina Golland, PhD[2]; Natalia S Rost, MD, MPH[1]

[1] Department of Neurology, Massachusetts General Hospital, Harvard Medical School
[2] Computer Science and Artificial Intelligence Lab, MIT
[3] Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School
[4] Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School

Abstract. Introduction: White matter hyperintensity (WMH) as seen on T2 fluid attenuated inversion recovery (FLAIR) MRI is a rarefaction of white matter that is an established risk factor for stroke, independent of traditional vascular risk factors. Age is highly correlated with WMH volume (WMHV), but much variation in WMHV among ischemic stroke (IS) patients of similar ages remains unexplained. The 17q25 locus has been reported to be associated with WMHV in IS subjects, though heterogeneity within the IS population may limit the detected effect size.

Aims: To compare the effect size of the 17q25 locus between subgroups of IS subjects, identified using a regression mixture model.

Methods: Clinical characteristics as well as laboratory and radiographic data were ascertained on admission for IS in all consecutive patients $\geq$ 18 years. WMHV was measured using a previously validated, semi-automated method (Figure 1), normalized for head size, and natural log-transformed for genome-wide association analyses. Subjects were classified into groups with different rates of WMHV progression using the regression mixture model of age against normalized WMHV (nWMHV). Standard genotyping quality control procedures for quantitative trait analysis were applied; unobserved genotypes were imputed and subjected to standard post-imputation quality control filters. Association analysis of nWMHV for the 17q25 SNPs within each group was adjusted for age, sex, and principal components 1 and 2.

Results: Regression mixture modelling of age versus nWMHV identified 5 clusters (Figure 2). Association analysis of the 17q25 SNPs showed increasing effect sizes with increasing rate of progression of WMHV of the IS subgroup (Table 1).

Conclusions: The effect size of the 17q25 locus increased from the group with the slowest rate of progression of WMHV to the group with the fastest rate of WMHV progression. Because the slowest WMHV progressors may share a similar rate of progression with the stroke-free adults, the increasing effect size with WMHV progression demonstrates that reducing heterogeneity in the IS cohort may increase detectable effect size of SNPs associated with WMHV.

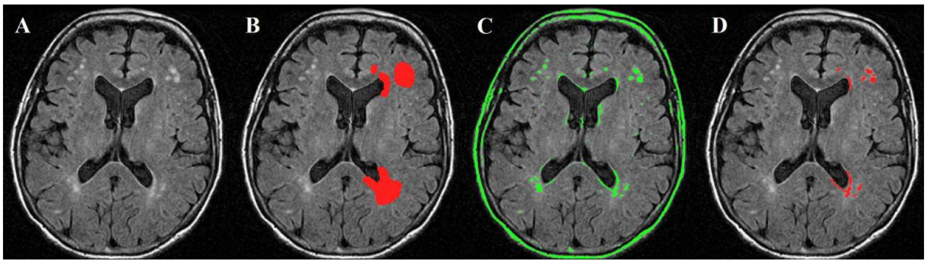


Fig. 1. Magnetic resonance imaging-based volumetric analysis of white matter hyperintensity volume using MRIcro software. An axial T2-FLAIR sequence with WMH is presented in panel (A). In panel (B), a region-of-interest map has been drawn over the WMH on the right side. In panel (C), the hyperintensities on the scan have been automatically highlighted using signal intensity thresholding. Taking the intersection of the region-of-interest and the hyperintensities and then manually editing as necessary produces the final scan with the highlighted and quantified WMHV in panel (D).

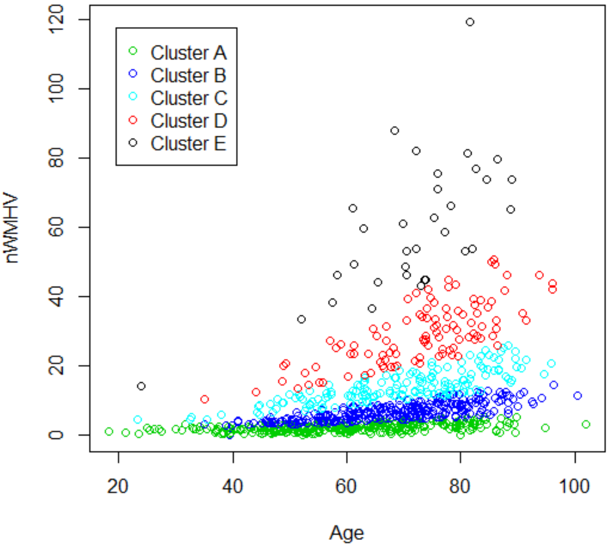


Fig. 2. Regression mixture modelling of age versus nWMHV identifies five subgroups.

Table 1. Effect sizes of the 17q25 locus SNPS on nWMHV within each IS subgroup.

	SNP					
Group	rs3744028	rs9894383	rs11869977	rs936393	rs3744017	rs1055129
A	-0.068	-0.078	-0.093	-0.088	-0.073	0.085
B	0.170	0.266	0.222	0.350	0.325	0.017
C	0.239	0.496	0.389	0.366	0.187	0.151
D	0.812	0.765	1.087	0.568	0.264	1.055
E	1.847	3.315	3.342	3.377	3.274	3.115

Feature Selection and Imaging-Genetics Predictions Using a Sparse, Extremely Randomized Forest Regressor

Albert Montillo, Shantanu Sharma, and Marcel Prastawa

GE Global Research, Niskayuna, NY 12309

Aims: We propose a sparse extension of the extremely randomized forest (ERF) [2] nonlinear regressor by embedding it in a model reduction framework providing it with sparsity to reduce model complexity and reduced variance. The method enjoys few tunable parameters and is readily scalable to large data through parallelization. We demonstrate the utility of the method on two cases entailing joint modeling of genetic and image features with cognitive scores in Alzheimer's disease. In the first case ($\sim 10^3$) genetic SNP features are combined with the trimmed mean summary statistic of voxel shrinkage in 38 cortical and subcortical structures upon nonlinear registration to a reference brain. In the second case the SNP features are combined with quartile summary statistics of the shrinkage in a subset of 17 structures. In both the method identifies clinically relevant features by assigning feature importance scores and the final model using only relevant features achieves high prediction accuracy.

Method: We construct a model reduction framework consisting of a hierarchical cross-validation in which each fold of an outer k-fold cross-validation contains a complete q-fold inner cross-validation. The outer divides the data into train and test sets allowing for model evaluation, while the inner divides each outer train set into new q-fold train and validation sets. Similar to recursive feature elimination [3], only features whose importance is greater than the mean importance is retained for the next iteration until the validation error no longer diminishes. However for the ERF, OOB predictions [1] are unavailable therefore we use mean decrease in node impurity to compute features importance. Additionally, each run of the inner and outer cross-validation folds are repeated n=10 and m=4 times respectively with random training data shuffling to reduce variance. In the inner cross-validation feature importances are averaged across repetitions, while the outer computes optimal feature set size from votes cast by the inner cross-validation.

Results and Conclusions: We applied the proposed approach to a subset of the ADNI [5] imaging genetics data containing 30 normal and 18 AD subjects. For genomic features we normalized the $< R, \theta >$ tuples from 427 SNPs associated with AD (i.e. whose p-value $\leq 10^{-3}$) [4] that are contained in the ADNI2 GWAS panel and have resolved R and θ values. To form image features we computed the log Jacobian of the mapping between the subject's T1 MRI and a reference template. Our first imaging-genetic dataset combining our genomic features with imaging features computed as the trimmed mean (10%) of the voxel Jacobians in 38 cortical and subcortical regions defined as part of the Freesurfer atlas. Our second dataset combines the genomic features with summary quartile (Q1, Q2, Q3) measures of the Jacobian distribution of a subset of 17 structures.

We applied our method to predict AVLT [1] for both datasets. Similar RMSE prediction errors were are achieved (Fig. 1), though the best anatomical region identification occurred using quartile measures. The anatomical regions assigned high importance are shown in Fig. 2 and Fig. 3. Importances for the individual SNP R and Θ components are shown in Fig. 4a and Fig. 4b. The results from our approach show promising capabilities for sparse feature selection and prediction, We look forward to applying it to additional datatsets and extending its capabilities.

References

[1] Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)

[2] Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning 36(1), 3–42 (2006)

[3] Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46(1-3), 389–422 (Mar 2002)

[4] Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M.L., Pahwa, J.S., Moskvina, V., Dowzell, K., Williams, A., et al.: Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease. Nature genetics 41(10), 1088–1093 (2009)

[5] Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., et al.: The Alzheimer's Disease neuroimaging initiative: A review of papers published since its inception. Alzheimer's & Dementia 9(5), e111–e194 (2013)

[1] Auditory Verbal Learning Test

Features	Trimmed mean	Quartiles
Imaging-Genetics	4.86	4.47
Imaging	4.49	4.46
Genetics	4.56	4.53

Fig. 1: Regression performance measured as RMS error using 4-fold cross-validation. In terms of RMSE, both trimmed mean and quartiles give similarly good performance.

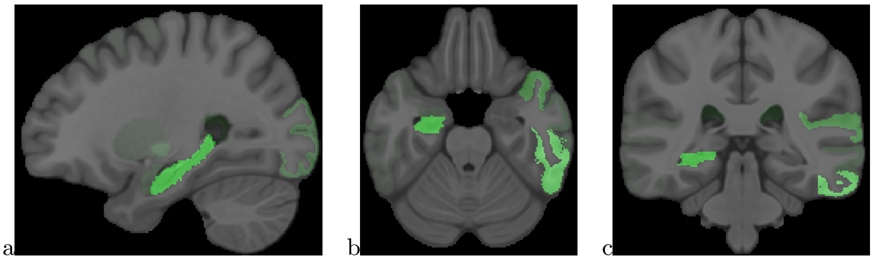


Fig. 2: Automatically assigned region importances in green for AVLT prediction on ICBM template using quartile Jacobian measures. (a) Sagittal highlights hippocampus in center, while axial, coronal views (b,c) highlight hippocampus, inferior and superior temporal regions. (asymmetry from training on different structures per hemisphere.)

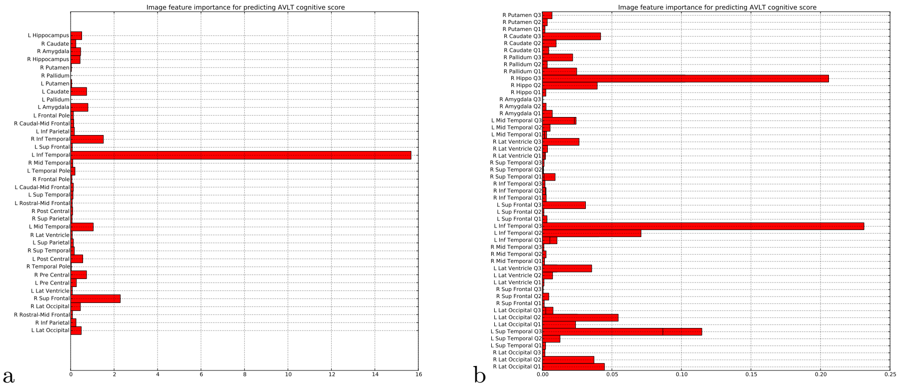


Fig. 3: Importances for the imaging regions. Using trimmed mean of log Jacobian (a) yields only temporal region with high importance while using quartile measures (b) assigns high importance to hippocampus and temporal regions.

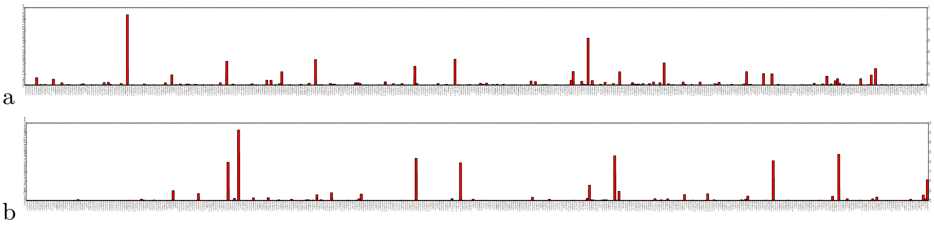


Fig. 4: SNP importances for (a) genomic R measure and (b) genomic Θ measure.

Predictive Imaging-Genetics Models with Feature Selection and Dimension Reduction Using Sparse Partial Least Squares

Rui Li, Xiaojie Huang, Shantanu Sharma, and Marcel Prastawa

GE Global Research, Niskayuna, NY 12309

Aims: We propose an extension to the sparse Partial Least Squares (PLS) regression framework using image regularization terms for sparsity and smoothness. Our method leverages the wealth of information present in the disease phenotype at the voxel level, and is extendable and scalable to more comprehensive imaging genetics data. We use this method on the joint modeling of image features and genetics variants that associate with cognitive scores in Alzheimer's disease. The proposed method enables the parsimonious modeling of the image and genetics feature which are high dimensional ($\sim 10^5$) as it reduces the feature dimensionality. Furthermore, discovery of clinically relevant features is enabled via importance weighting of individual features.

Method: We extend the sparse PLS regression framework [1], predicting cognitive scores Y from imaging-genetic features X by solving the optimization problem:

$$\min_{w,c} -\kappa w^T M w + (1-\kappa)(c-w)^T M (c-w) + \lambda_1 |c|_1 + \lambda_2 |c|_2 + \lambda_3 \int_{z\in\Omega} |\nabla c_I(z)|\, dz \ \text{ s.t. } w^T w = 1$$

where w and c are the original and surrogate direction vectors and $M = X^T Y Y^T X$. Here, $\kappa \in (0, 0.5]$ controls the concavity of the objective function and the closeness of w and c. The L_1 regularization term $|c|_1$ promotes sparsity on w governed by penalty weights λ_1 that can be set differently for imaging and genetic features. The L_2 penalty $|c|_2$ deals with potential singularity of M. The penalty $|\nabla c_I|$ encourages smoothness on the weights of the voxel image features.

Results and Conclusions: We used the imaging genetics data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [3], using a subset of 30 normal controls and 18 AD subjects. We applied the proposed method to voxel-wise image data of the log Jacobian of the mappings between the subjects and a reference template. We combined these imaging features with the set of 761 AD-associated SNPs (p-value $\leq 10^{-3}$) from [2] pruned to 427 SNPs by removing sites missing in ADNI2 GWAS panel (Illumina HumanOmniExpress BeadChip) and or with unresolved R and θ values over the training set. The normalized $< R, \theta >$ tuples from these 427 SNPs were used as the genomic features for each subject.

The proposed method provides predictions of cognitive scores (ADAS [1] and AVLT [2]) using imaging-genetics, with Fig. 1 showing the algorithm's perfor-

[1] Alzheimer's Disease Assessment Scale

[2] Auditory Verbal Learning Test

mance and behavior with increasing sparsity penalty. Fig. 2 shows the projection of the imaging genetics features X onto a 2D space formed by the first two PLS components. Fig. 3 shows the feature relevance weights for the prediction of cognitive scores. Preliminary results using our high dimensional data analysis framework are promising. Our framework can be readily extended, for example with explicit modeling of imaging-genetics interactions.

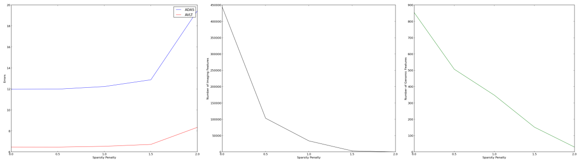


Fig. 1. Performance and behavior of our proposed algorithm with increasing sparsity penalty. Left: error measures for the Alzheimer's Disease Assessment Scale (ADAS) and Auditory Verbal Learning Test (AVLT) cognitive scores. Middle: The number of relevant voxel imaging features (i.e., having non-zero weights). Right: The number of relevant genomic features. The number of selected imaging features decrease exponentially, while the number of selected genomic features decrease linearly.

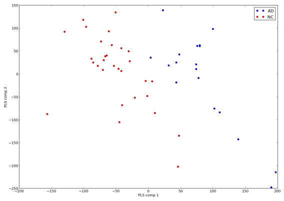


Fig. 2. Projection of imaging-genetics features onto the first two PLS components, showing good separation between AD subjects (blue) and normal controls (red) .

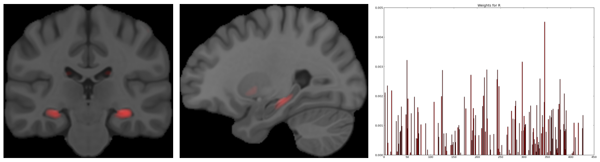


Fig. 3. Imaging-genetics feature relevance for prediction. Left and middle: voxel relevance weights (red) overlaid on the template provided by the International Consortium on Brain Mapping. The coronal (left) and sagittal (middle) views show the hippocampus being highlighted. Right: SNP relevance weights for the genomic R measure.

References

1. H. Chun and S. Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.
2. D. Harold, R. Abraham, P. Hollingworth, R. Sims, A. Gerrish, M. L. Hamshere, J. S. Pahwa, V. Moskvina, K. Dowzell, A. Williams, et al. Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease. *Nature genetics*, 41(10):1088–1093, 2009.
3. M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, E. Liu, et al. The Alzheimer's Disease neuroimaging initiative: A review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111–e194, 2013.

Investigation of biological pathways involved in brain development in preterm neonates using a multivariate phenotype and sparse regression

Michelle L Krishnan[1], James Boardman[2], Matt Silver[3], Gareth Ball[1], Serena Counsell[1], Andrew J Walley[4], A David Edwards[1], Giovanni Montana[1]

[1]Centre for the Developing Brain, King's College London; [2]Neonatology, Royal Infirmary of Edinburgh; [3]London School of Hygiene and tropical Medicine; [4]Molecular Genetics and Genomics, NHLI, Imperial College London

1 Background and Aims

The incidence of preterm birth is increasing steadily [1], with a high proportion of survivors experiencing adverse motor, cognitive and psychiatric sequelae [2]. Diffusion tensor imaging (DTI) provides measures of white matter microstructure that are correlated with neurodevelopmental outcome [3] and highly heritable [4]. Joint modeling of multivariate imaging and genetic data, leveraging prior biological knowledge of functional pathways, increases power to detect associations in complex disease [5,6]. We aim to identify biological pathways through which premature birth impacts the microstructure of white matter in neonates.

2 Methods

3-Tesla MR images and saliva were acquired for 72 preterm infants (mean gestational age (GA) 28^{+4} weeks, mean postmenstrual age (PMA) at scan 40^{+3} weeks). FA maps were constructed from 15-direction DTI, and Tract Based Spatial Statistics [7] was used to obtain a group white matter skeleton varying with degree of prematurity, adjusting for PMA at scan (Fig.1). The phenotype was reduced to its three principal components, explaining 47% of the total variance. Salivary DNA was extracted and genotyped using Illumina HumanOmniExpress-12 arrays. Pathways sparse reduced-rank regression (PsRRR) [6] was used to jointly model the voxel-wise effects of genome-wide SNPs grouped into 186 KEGG pathways (λ 0.99, 100 subsamples, 20 iterations, 2000x10 and 4000x10 model fits). The PsRRR method reduces bias due to pathway linkage disequilibrium and size by adjusting pathway weightings in the regression model according to the empirical bias in pathway selection frequencies, obtained by fitting the group LASSO model with a null response.

3 Results

High-ranking pathways associated with a brain endophenotype impacted by prematurity include a preponderance of mechanisms related to lipid metabolism and vesicular transport (13/30, Table 1). The highest ranked pathways have corresponding low selection probabilities in the null model (Fig. 2). Two of the top three pathways (peroxisome proliferator-activated receptor (PPAR) metabolism and alpha-linoleic acid metabolism) include the gene fatty acid desaturase (FADS2), which has been recently associated with changes in brain microstructure in a candidate study with this cohort [8]. Another pathway in the top three (glycine serine and threonine metabolism) is a suggested link between lipid and amino acid metabolism in various tissues including brain [9].

4 Conclusions

Biological pathways associated with a quantitative multivariate imaging endophenotype of prematurity suggest an important role for lipid metabolism. FADS2 might be driving pathway selection as it is a member of two highly ranked, relatively small pathways involving lipid metabolism. Derivatives of alpha-linoleic acid protect cells from free-radical mediated oxidative stress, and promote differentiation of immature brain cells including oligodendrocytes and neuroblasts through PPAR-γ activation, thus conferring neuroprotection and enhancing myelination [10]. The effects of FADS2 variants may also be reflected in higher-level developmental measures, as they may impact childhood IQ by moderating dietary influences [11].

Figures and Tables

Fig. 1. Group white matter DTI skeleton, showing (blue) voxels that vary due to gestational age at birth, adjusting for post-menstrual age at scan. Axial views superior to inferior, left to right.

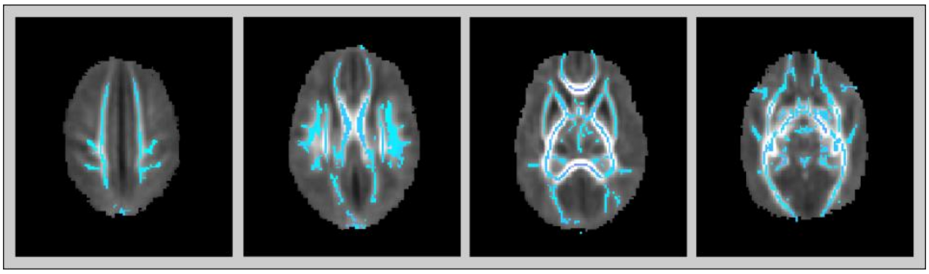
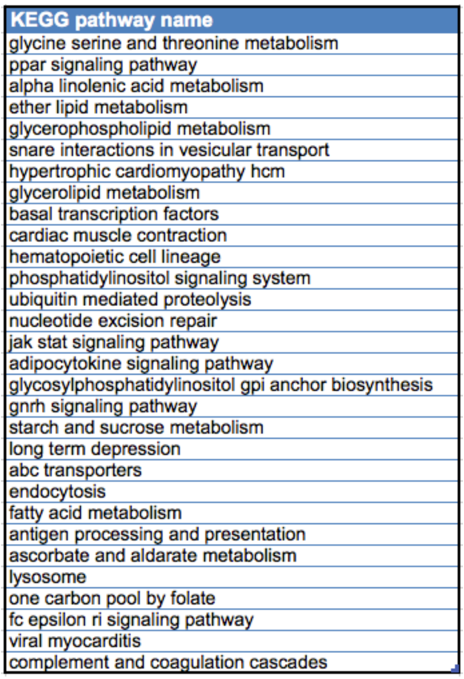
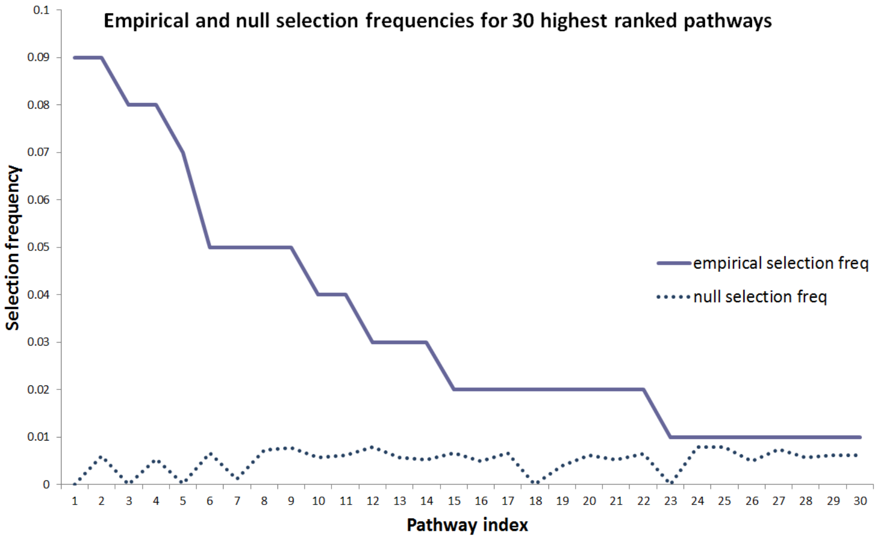


Table 1. Top 30 KEGG pathways, ranked by pathway selection frequency.



KEGG pathway name
glycine serine and threonine metabolism
ppar signaling pathway
alpha linolenic acid metabolism
ether lipid metabolism
glycerophospholipid metabolism
snare interactions in vesicular transport
hypertrophic cardiomyopathy hcm
glycerolipid metabolism
basal transcription factors
cardiac muscle contraction
hematopoietic cell lineage
phosphatidylinositol signaling system
ubiquitin mediated proteolysis
nucleotide excision repair
jak stat signaling pathway
adipocytokine signaling pathway
glycosylphosphatidylinositol gpi anchor biosynthesis
gnrh signaling pathway
starch and sucrose metabolism
long term depression
abc transporters
endocytosis
fatty acid metabolism
antigen processing and presentation
ascorbate and aldarate metabolism
lysosome
one carbon pool by folate
fc epsilon ri signaling pathway
viral myocarditis
complement and coagulation cascades

Fig. 2. The highest ranked pathways have a low selection frequency in the null model.



References.

1. Blencowe, H., Cousens, S., Oestergaard, M.Z., et al: National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. Lancet, 379:2162–2172 (2012)
2. Johnson, S., Marlow, N.: Preterm birth and childhood psychiatric disorders. Pediatr. Res. 69:11R–18R (2011)
3. Counsell, S.J., Edwards, A.D., Chew, A.T.M., et al: Specific relations between neurodevelopmental abilities and white matter microstructure in children born preterm. Brain, 131:3201–3208 (2008)
4. Geng, X., Prom-Wormley, E.C., Perez, J., Kubarych, T., Styner, M., Lin, W., Neale, M.C., Gilmore, J.H.: White matter heritability using diffusion tensor imaging in neonatal brains. Twin Research and Human Genetics, 15(3): 336-350 (2012)
5. Silver, M. and Montana G.: Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. Stat. Appl. Genet. Mol. Biol., 11(1): 7 (2012)
6. Silver, M., Janousova E., Hua X., Thompson, P.M., Montana, G.: Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. Neuroimage, 63(3): 1681-1694 (2012)
7. Smith, S.M., et al.: Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. Neuroimage, 31(4): 1487-505 (2006)
8. Boardman, J.P., Walley, A., Ball, G., Takousis, P., Krishnan, M.L., Hughes-Carre, L., Aljabar, P., Serag, A., King, C., Merchant, N., Srinivasan, L., Froguel, P., Hajnal, J., Rueckert, D., Counsell, S., Edwards, A.D.: Common Genetic Variants and Risk of Brain Injury After Preterm Birth. Pediatrics, 133(6): e1655-e1663 (2014)
9. Li, P., Kim, S.W., Li, X., Datta, S., Pond, W.G., Wu, G.: Dietary supplementation with cholesterol and docosahexaenoic acid affects concentrations of amino acids in tissues of young pigs. Amino Acids, 37(4): 709-716 (2009)
10. Minghetti, L., Salvi, R., Lavinia Salvatori M., Antonietta Ajmone-Cat, M., De Nuccio, C., Visentin, S., Bultel-Poncé, V., Oger, C., Guy, A., Galano, J.M., Greco, A., Bernardo, A., Durand, T.: Nonenzymatic oxygenated metabolites of α-linolenic acid B1- and L1-phytoprostanes protect immature neurons from oxidant injury and promote differentiation of oligodendrocyte progenitors through PPAR-γ activation. Free Radic. Biol. Med., 73C:41-50 (2014)
11. Steer, C.D., Davey Smith, G., Emmett, P.M., Hibbeln, J.R., Golding, J.: FADS2 polymorphisms modify the effect of breastfeeding on child IQ. PLoSONE, 5(7):e11570 (2010)

A Novel Atlas-based Approach to the Detection of Mouse Embryo Ventricular Septal Defects

Xi Liang[1,3⋆], Zhongliu Xie[2,3⋆], Asanobu Kitamoto[3,6], Masaru Tamura[4,5],
Toshihiko Shiroishi[5], and Ramamohanarao Kotagir[1]

[1] University of Melbourne, Melbourne, Australia
[2] Imperial College London, London, UK
[3] National Institute of Informatics, Tokyo, Japan
[4] RIKEN BioResource Center, Tsukuba, Japan
[5] National Institute of Genetics, Shizuoka, Japan
[6] The Graduate University for Advanced Studies (SOKENDAI), Tokyo, Japan

The goal of International Mouse Phenotyping Consortium [1] is to study the over 23,000 mouse genes by knocking them out one-by-one and raising the genetically engineered mouse lines for comparative analysis against the wild-type, with respect to morphology, metabolism and other biological traits (a.k.a. "phenotype"). Large amounts of knock-out mice have been raised, leading to a strong demand for a high-throughput phenotyping technology. The traditional means via time-consuming histological analysis is clearly unsuitable in this scenario. Medical imaging technologies such as CT and MRI therefore have been used to develop more efficient phenotyping approaches.

Existing work [2–5] primarily rests on volumetric analysis for phenotype detection, which however generally fails when features are subtle, such as the ventricular septal defects (VSD) in the heart. More sophisticated VSD detection approaches include measuring the cavities of cardiac and vascular structures [6], or diameters of great arteries and semilunar valves [7], based on a semi-automated segmentation and 3D reconstruction framework, however still unable to meet the high-throughput requirement due to manual labor involved. This study proposes, to the best of our knowledge, the first automated VSD detection system for mouse embryos.

VSD indicates the presence of a hole in the ventricular septum, i.e. the wall dividing the left and right ventricles of the heart, and is probably the most common congenital cardiac anomaly. Our algorithm starts with the creation of a normal average mouse atlas using all the wild-type images, followed by registration of the target mutant images to the atlas whereby labels are back propagated to perform heart segmentation accordingly. Then the left and right ventricles are further segmented with the additional use of a region growing technique, and VSD detection is completed by checking whether there is an overlap between the two ventricle segmentations, as shown in Fig. 1.

Our approach was validated on a database of 15 mouse embryo images: 3 controls and 12 mutants, where VSD is present in 3 mutant cases. All the data was produced at Japan National Institute of Genetics based on the C57BL/10 mouse line, imaged using μ-CT at 14.5 days postcoitum (dpc). The detection system

[⋆] These authors contributed equally to this work

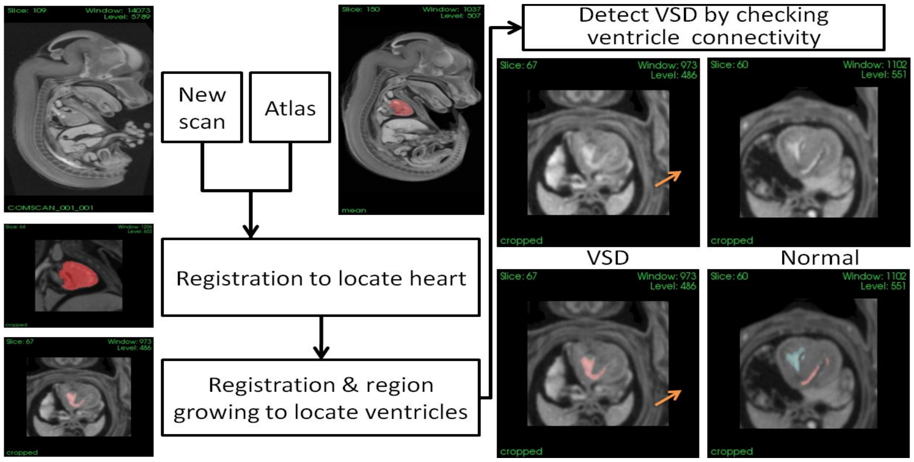


Fig. 1. Overview of the detection algorithm

showed an overall accuracy of 91.7%, with a sensitivity of 66.7% and specificity of 100% in the experiments, i.e. 2 out of the 3 VSD cases were successfully identified and none of the remaining 9 cases were misidentified with the presence of VSD. This pilot study demonstrates the potential of our algorithm to establish an effective high-throughput phenotyping framework for VSD detection, and may also be extended to other similar phenotype detection domains.

References

1. International Mouse Phenotyping Consortium (IMPC), www.mousephenotype.org
2. Zamyadi, M., Baghdadi, L., Lerch, J., Bhattacharya, S., Schneider, J., Henkelman, R., Sled, J.: Mouse embryonic phenotyping by morphometric analysis of MR images. Physiological Genomics, 42, 89-95 (2010)
3. Cleary, J., Modat, M., Norris, F., Price, A., Jayakody, S., Martinez-Barbera, J., Greene, N., Hawkes, D., Ordidge, R., Scambler, P., et al.: Magnetic resonance virtual histology for embryos: 3D atlases for automated high-throughput phenotyping. NeuroImage, 54, 769-778 (2011)
4. Wong, M., Dorr, A., Walls, J., Lerch, J., Henkelman, R.: A novel 3D mouse embryo atlas based on micro-CT. Development, 139, 3248-3256 (2012)
5. Norris, F., Modat, M., Cleary, J., Price, A., McCue, K., Scambler, P., Ourselin, S., Lythgoe, M.: Segmentation propagation using a 3D embryo atlas for high-throughput MRI phenotyping: comparison and validation with manual segmentation. Magnetic Resonance in Medicine (2012)
6. Schneider, J., Bamforth, S., Farthing, C., Clarke, K., Neubauer, S., Bhattacharya, S.: Rapid identification and 3D reconstruction of complex cardiac malformations in transgenic mouse embryos using fast gradient echo sequence magnetic resonance imaging. Journal of Molecular and Cellular Cardiology., 35, 217-222 (2003)
7. Weninger, W., Maurer, B., Zendron, B., Dorfmeister, K., Geyer, S.: Measurements of the diameters of the great arteries and semi-lunar valves of chick and mouse embryos. Journal of Microscopy, 234, 173-190 (2009)

Dopamine-Related Genetic Influences on Cognitive Flexibility

Hans Melo,[1*] Daniel J Mueller,[2,3] William A Cunningham,[1,4] Adam Anderson.[5]

[1]Psychology Department, University of Toronto; [4]Rotman School of Management, University of Toronto; [5]Department of Human Ecology, Cornell University; [3]Department of Psychiatry, University of Toronto; [2]Centre for Addiction and Mental Health.

*Correspondence at hans.melo@mail.utoronto.ca

Objective: Cognitive flexibility has been broadly defined as the ability to adapt one's cognitive resources to engage in the immediate demands of the environment. For example, a change of environment might force an individual to reframe its problem-solving strategy. Failure to adapt may have negative consequences for an individual and underpin a variety of mental disorders. Previous research suggests that cognitive flexibility relies on dopaminergic dynamics involving the basal ganglia, anterior cingulate cortex, and prefrontal cortex. However, little is known with regards to how genes modulating dopaminergic function in these regions might affect cognitive flexibility. The aim of this study was to explore the influence of key genetic polymorphisms on the neural mechanisms underlying cognitive flexibility using functional magnetic resonance imaging (fMRI).

Methods: 70 healthy individuals (34 male; mean age 20) performed an Embedded Figured task while inside an fMRI scanner. The task asked participants to judge whether a simple shape was embedded in a more complex abstract figure. fMRI activity was collected using a 3T GE MRI scanner. T2*-weighted images were collected using a SPRL sequence (TR=2s, TE=30ms, 3x3x3 mm). Brain imaging analyses were performed using FSL (www.fmrib.ox.ac.uk/fsl) and implemented mixed-effect models in R. Activations were considered significant if exceeded P < .0001 (uncorrected). Saliva samples were collected from all participants for genomic DNA extraction and analyzed for a number of genetic variants associated with dopamine function including DAT1(SLC6A3), DRD2(C957T), DRD4(exon III), DARP-32, COMT(Val158Met). Additionally, participants completed a number of self-report measures including the positive and negative affect survey (PANAS) a week prior to the experiment. Ongoing analyses focuses on gene-gene interactions.

Results: The study revealed genotype-related differences in cognitive flexibility. A polymorphism in the DARPP-32 gene (rs907094), associated with striatal dopaminergic function, was predictive of overall performance in the embedded figures task. fMRI analyses revealed recruitment of several brain regions including anterior cingulate cortex, lateral prefrontal cortex and insula. Critically, activation of the right insula was only present in trials requiring enhanced cognitive flexibility.

Conclusions: We provide evidence of dopamine-related genetic influences on cognitive flexibility. Our results point to the role of the DARPP-32 in supporting the frontal dopaminergic mechanisms associated with cognitive flexibility.

Acknowledgements: Supported by CIHR grant #490407 to William A Cunningham, John Templeton Foundation grant #490225 to Adam Anderson, and Varela award # 494241 to Hans Melo.

Imaging Genomic Mapping of Tumor Volume MRI Phenotype in Glioblastoma and Correlation with the Survival and Treatment Response

Ginu A. Thomas[1], Sanjay Singh[1], Islam Hassan[1], Pascal O. Zinn[1] and Rivka R. Colen[1]

[1]The University of Texas MD Anderson Cancer Center

1 Aims

The search for an effective therapy for Glioblastoma(GBM)continues despite the recent discoveries of new molecular targets and pathways. MRI is a noninvasive diagnostic modality previously validated to be able to perform robust radio-genomic (imaging genomic) screens for uncovering potential novel targets.Thus,we seek to provide comprehensive image genomic analysis in GBM using quantitative MRI enhancing volume and large scale gene and micro-RNA expression profiles and correlating with survival.

2 Materials & Methods

We identified 99 treatment naive GBM patients from the The Cancer Genome Atlas (TCGA) who had MR imaging data available in the The Cancer Imaging Archive (TCIA). This data was randomized into training and validation sets. Gene expression profiles for these patients were correlated with tumor volumes derived from contrast enhancement volume in MRI to identify specific genes and gene networks associated with high contrast enhancement.Both sets were further characterized by Kaplan-Meier survival statistics and GBM molecular sub type-specific distribution.

3 Results

The top up-regulated and down-regulated gene in both the discovery and validation sets were identified in those patients with high CE volume.Patients with high CE volume demonstrated gene and miRNA signatures associated with angiogenesis. These patients also had poor survival.

4 Conclusions

Here,we present our imaging screening method for molecular cancer subtypes and genomic correlates of contrast enhancing volume.Our findings also have potential therapeutic significance since successful targeting of those genes,miRNA and pathways involved in the growth of high contrast enhancing tumor volumes will improve therapy and patient survival in GBM.

Imaging Genomic Biomarker Signature for MGMT Promoter Methylation Identification

Ginu A. Thomas[1], Pascal O. Zinn[1] and Rivka R. Colen[1]

[1]The University of Texas MD Anderson Cancer Center

1 Aims

To create an imaging biomarker signature in order to identify those Glioblastoma (GBM) patients with MGMT promoter methylation.

2 Materials & Methods

We identified 86 treatment-naïve patients from The Cancer Genome Atlas (TCGA) who had MGMT methylation status and pretreatment MRI from The Cancer Imaging Archive (TCIA). Qualitative VASARI imaging features for these 86 patients were assessed by 3 independent neuroradiologists and consensus was reached. Quantitative volumetric analysis was done in the 3D Slicer software 3.6(http://www.slicer.org) using segmentation module. Fluid Attenuated Inversion Recovery (FLAIR) was used for segmentation of the edema and post-contrast T1 weighted imaging (T1W1) for segmentation of enhancement (defined as tumor) and necrosis. Each qualitative and quantitative feature was correlated to MGMT methylation status both independently and as groups and subgroups. Multiple classification models were created via regression modeling and partition analysis using various combinations of variables.

3 Results

An imaging biomarker signature was created that predicted MGMT promoter methylation status. Multiple qualitative and quantitative MRI features correlated with MGMT methylation status. The logistic regression model with combinations of quantitative volumetric variables, clinical variables and the qualitative variable 'diffusion' could predict MGMT methylation with an AUC of 0.847 with a sensitivity of 82% and a specificity of 83.8%.

4 Conclusions

MGMT methylation status plays an important role in patient predictive and prognostic stratification of patients with GBM. The identification of a non-invasive biomarker signature as a surrogate for MGMT methylation can help stratify patients in specific therapy and predict response versus non response to therapy. An imaging genomic signature can be expected to promote a more robust personalized approach to patient care and accelerate drug development and clinical trials.

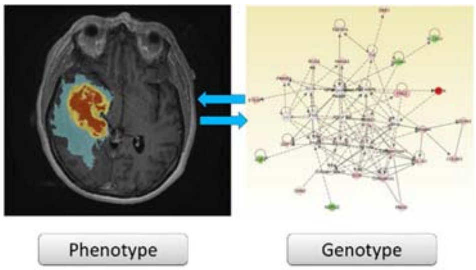
An Introduction to Imaging Genomics in Glioblastoma

Rivka R.Colen[1], Ginu A.Thomas[1], Pascal O.Zinn[1]

[1]The University of Texas MD Anderson Cancer Center

1 Aims

Imaging genomics (radio genomics) is a new field which links the imaging traits (radiophenotypes) with gene-expression profiles and the underlying genomic composition in tumors such as Glioblastoma (GBM). Large dataset analysis and interpretation for cancer requires a high cost, time and manpower. This new field can provide cost-effective biomarkers that can accurately reflect underlying molecular cancer compositions. This is the first large-scale quantitative imaging genomic mapping study in GBM.



2 MATERIALS & METHODS

In this retrospective study, we identified 78 treatment-naïve GBM patients from The Cancer Genome Atlas (TCGA) who had:

Gene and microRNA expression profiles: Using Affymetrix level 1 mRNA and Agilent level 2 microRNA data, a total of 13,628 genes and 555 microRNAs (1,510 hybridization probes) were analyzed for each patient.

Pretreatment MR-neuroimaging from TCIA: Using 3D Slicer software 3.6, FLAIR was used for segmentation of the edema, and post-contrast T1 weighted imaging (T1W1) for segmentation of enhancement (defined as tumor) and necrosis (Figures 1 & 2).

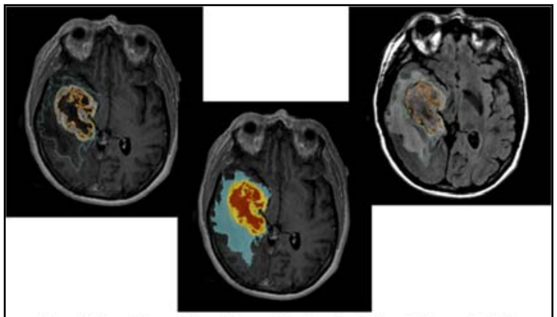


Figure 1. Tumor Segmentation. 65 year old male patient with a right temporal GBM. Segmentation of tumor edema (blue), enhancement (yellow) and necrosis (red) was performed and edema volume was obtained.

Genomic data was analyzed for significance and differential fold regulation and expression using Comparative Marker Selection (CMS) (Broad Institute, MIT) and Ingenuity Pathway Analysis (IPA) and then associated with the imaging characteristics.

For the first part of the study, we categorized patients into high and low FLAIR volume groups for analysis and comparison. The Kaplan Meier method was used to calculate overall- and progression-free survival between the two groups. Mean gene and microRNA correlations were calculated using R square statistics.

3 RESULTS

Gene expression analysis identified preferentially up-regulated genomic events in the high vs. low FLAIR volumes groups. A total of 53 mRNAs and 5 microRNAs were identified and were analyzed by IPA. The top upregulated gene was PERIOSTIN (POSTN), and the top downregulated miRNA was miR-219, which is predicted to bind to POSTN (Fig. 3a, b). They were concordant with the underlying biological processes of edema/invasion, necrosis, and enhancing tumor MRI phenotypes. Kaplan Meier analysis demonstrated that these resulted in significantly decreased survival (P=0.0008) and shorter time to disease progression (P=0.0009). In some cases, the gene expression was a stronger prognostic variable than the molecular subtype (as defined by Verhaak and colleagues) (P=0.0001) (Figure 3.c).

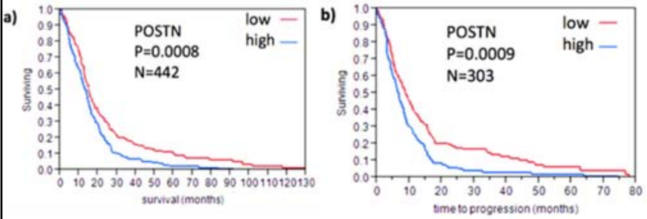


Figure 3a & b. Kaplan Meier curves for Periostin. (a) Time to death and (b) Progression free survival.

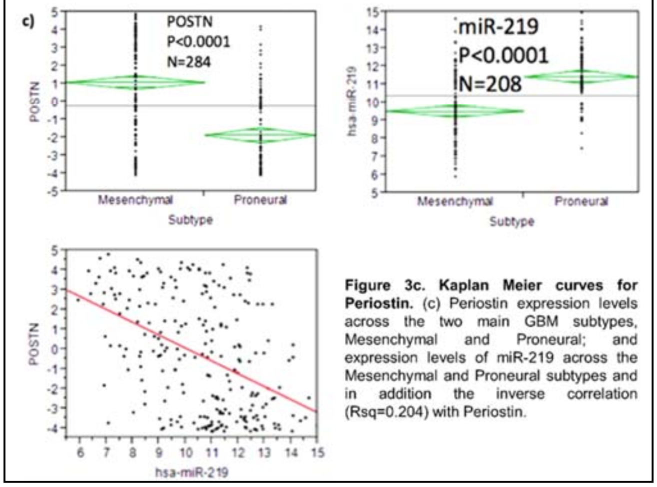


Figure 3c. Kaplan Meier curves for Periostin. (c) Periostin expression levels across the two main GBM subtypes, Mesenchymal and Proneural; and expression levels of miR-219 across the Mesenchymal and Proneural subtypes and in addition the inverse correlation (Rsq=0.204) with Periostin.

4 CONCLUSION

Imaging genomics can provide relevant clinical decision making data by linking imaging phenotypes to underlying genotypes in GBM patients and vice versa. MRI FLAIR volumes provide a screening method for molecular cancer subtypes and genomic correlates of cellular invasion. Imaging genomic mapping can be used to discover biologically meaningful genes and microRNA that can be used for development of therapeutic drugs, identifying candidates with target genes, and predicting prognosis and drug response.