NETWORK-BASED GENOME WIDE STUDY OF HIPPOCAMPAL IMAGING PHENOTYPE IN ALZHEIMER'S DISEASE TO IDENTIFY FUNCTIONAL INTERACTION MODULES

Xiaohui Yao[1,2], *Jingwen Yan*[1,2], *Shannon Risacher*[1], *Jason Moore*[3], *Andrew Saykin*[1], *Li Shen*[1,2,†]

[1] Radiology and Imaging Sciences, Indiana University, Indianapolis, IN, 46202, USA
[2] Informatics and Computing, Indiana University, Indianapolis, IN, 46202, USA
[3] Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, 19104, USA

ABSTRACT

Identification of functional modules from biological network is a promising approach to enhance the statistical power of genome-wide association study (GWAS) and improve biological interpretation for complex diseases. The precise functions of genes are highly relevant to tissue context, while a majority of module identification studies are based on tissue-free biological networks that lacks phenotypic specificity. In this study, we propose a module identification method that maps the GWAS results of an imaging phenotype onto the corresponding tissue-specific functional interaction network by applying a machine learning framework. Ridge regression and support vector machine (SVM) models are constructed to re-prioritize GWAS results, followed by exploring hippocampus-relevant modules based on top predictions using GWAS top findings. We also propose a GWAS top-neighbor-based module identification approach and compare it with Ridge and SVM based approaches. Modules conserving both tissue specificity and GWAS discoveries are identified, showing the promise of the proposal method for providing insight into the mechanism of complex diseases.

Index Terms— GWAS, tissue-specific network, module identification, hippocampus, Alzheimer's disease

1. INTRODUCTION

Brain imaging genetics is an emerging field that studies how genetic variation influences brain structure and function. Genome-wide association studies (GWAS) have been performed to identify genetic markers such as single nucleotide polymorphisms (SNPs) that are associated with brain imaging quantitative traits (QTs) [1, 2, 3]. These findings, however, have limited power to explain how the identified SNPs interact to influence QTs. Using the biological networks and pathways as prior knowledge, integrative analysis have been performed to discover disease-relevant modules enriched by GWAS findings to examine collective effects of multiple genes, with the potential to enhance statistical power and help biological interpretation [4, 5, 6, 7, 8].

Existing module identification methodologies typically start from assigning GWAS statistics onto a user-specified functional interaction network. After that, candidate modules are formed across the entire network and assessed for whether to be enriched by the GWAS findings. One successful example is the dmGWAS [4], which loads gene-level p-values onto the network as node weights, and then applies dense module searching to identify modules with locally maximized proportion of significant genes. Another example is the network interface miner for multigenic interactions (NIMMI) [5], which scores genes by combining p-values with connectivities and then constructs modules from high weighted genes. Protein interaction network-based pathway analysis (PINBPA) [9] and its extension iPINBPA [8] start from a seed and expand the module by adding neighbors to reach a pre-given statistical significance. These strategies are all bottom-up. The power of the bottom-up strategy could be limited by multiple comparison correction as it examines a large number of candidate modules to identify GWAS enriched ones. Meanwhile the efficiency could also become suboptimal when large-scale networks are present.

Most network-based GWAS of quantitative traits are using tissue-free biological networks such as human PPI network, without considering tissue specificity. The precise functions of genes are highly related to their tissue context, and human diseases result from the disordered interplay of tissue-specific processes [10]. Recently, tissue-specific genome-scale functional interaction networks have been constructed to capture the changing functional roles of

†Correspondence to Li Shen (shenli@iu.edu). This research was supported by NIH R01 LM011360, R01 EB022574, U01 AG024904, RC2 AG036535, R01 AG19771, P30 AG10133, UL1 TR001108, R01 AG 042437, R01 AG046171, and R00 LM011384; DOD W81XWH-14-2-0151, W81XWH-13-1-0259, and W81XWH-12-2-0012; NCAA 14132004; and CTSI SPARC Program at Indiana University, and by NIH R01 LM011360, R01 LM009012, and R01 LM010098 at University of Pennsylvania. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

genes across tissues. Disease-gene associations have been re-prioritized by constructing a support vector machine (SVM) classifier to reorder GWAS results using tissue-specific network data as features, named as NetWAS (network-wide association study). It has been implemented on hippocampal volume in an Alzheimer's disease (AD) study to re-prioritize GWAS results and demonstrated that tissue-specific networks could provide helpful context for understanding complex human diseases [11]. Note that SVM classifier employed in NetWAS requires pre-defined threshold to label GWAS results, and may lose some valuable information embedded in the continuous z-scores corresponding to the GWAS p-values.

In this study, we propose and compare two novel module identification frameworks: (i) a machine learning approach that introduces a regression model into NetWAS to take continuous z-scores into account (Ridge regression in this paper); and (ii) a GWAS top-neighbor-based (tnGWAS) searching approach that extracts densely connected modules from top GWAS findings. Ridge and tnGWAS both offer a more efficient, top-down strategy to identify phenotype-relevant modules, while using slightly different hypotheses: (1) Ridge hypothesizes relevant modules are enriched by relatively significant and functionally-relevant genes; and (2) tnGWAS hypothesizes that relevant modules consist of top GWAS findings and their close neighbors. Of note, machine learning methods (e.g., SVM and Ridge) provide re-prioritized gene findings, while tnGWAS does not. We demonstrate the effectiveness of the proposed frameworks by applying them to a hippocampal imaging genetics analysis in the study of AD.

2. MATERIALS AND METHODS

To demonstrate the implementation of Ridge and tnGWAS on imaging QT-relevant module identification, we apply them to hippocampal imaging GWAS in AD. Studies with [^{18}F]FDG-PET have demonstrated that AD is associated with reduced use of glucose metabolism in hippocampus [12, 13]. We propose to identify imaging QT-relevant modules, by integrating a hippocampus-specific functional interaction network and GWAS results of hippocampal FDG measures.

2.1. Imaging data, genotyping data and GWAS

Imaging data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (adni.loni.usc.edu). Pre-processed FDG-PET scans were downloaded from LONI (adni.loni.usc.edu), and [^{18}F]FDG measurements of hippocampus were extracted based on the MarsBaR AAL atlas. Genotype data were also obtained from LONI. 989 non-Hispanic Caucasian participants with both FDG and genotype data available were studied. Association between the average FDG measure in the hippocampal region at the baseline and 5,574,300 SNPs was examined by GWAS using PLINK[14]. To facilitate the subsequent network-based analysis, a gene-level p-value was determined as the second smallest p-value of all SNPs located in $\pm$20K bp of the gene [15]. A number of 17,881 protein-coding gene p-values were obtained.

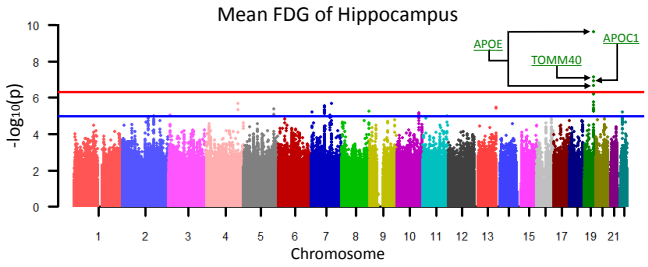

Fig. 1. Manhattan plot. Blue and red lines correspond to the p-values of 5e-5 and 5e-7 respectively.

2.2. Hippocampus functional interaction Network

Genome-wide functional interaction networks for specific human tissues and cell types have been generated to specialize protein functions and interactions of specific human tissues [10]. A hippocampus-specific functional interaction network was downloaded from GIANT (http://giant.princeton.edu/). Interactions among 17,881 protein-coding genes was extracted after being mapped by the GWAS results.

2.3. Alzheimer's disease documented genes

A list of 66 documented AD risk genes were collected to evaluate the re-prioritization results from three resources: 24 susceptibility genes from a large meta-analysis of AD [3], 15 AD-relevant genes from the Online Mendelian Inheritance in Man Disease database (OMIM), and 40 significant candidates from the AlzGene database (http://www.alzgene.org/).

2.4. Module identification framework

Two top-down module identification approaches were proposed, machine learning based and GWAS top-neighbor (tnGWAS) based. Below we describe their details.

Machine learning based GWAS re-prioritization: Following [10], we trained an SVM model using hippocampus-specific network connectivity as features and the significance status based on nominal p=0.01 as labels to re-prioritize GWAS results. In addition to SVM, we trained a ridge regression (Ridge) model using also the network data as features while using z-scores converted from p-values as responses. Different from classification which required a pre-defined threshold, regression approaches utilize the complete information from the continuous z-scores.

We trained SVM and Ridge models using interactions between a subset of genes C and all genes as features, and the z-scores converted from the gene-level p-values of C as responses (positive or negative labels for SVM). To balance the

training data, set C was constructed from the combination of significant gene set A and one third of randomly selected nonsignificant gene set B, where p=0.01 was used as nominal significance. Genes were re-prioritized according to their predictions (Ridge) or distances from separating hyperplane (SVM). Re-prioritized results offered a more flexible way to analyze functional associations at different scales.

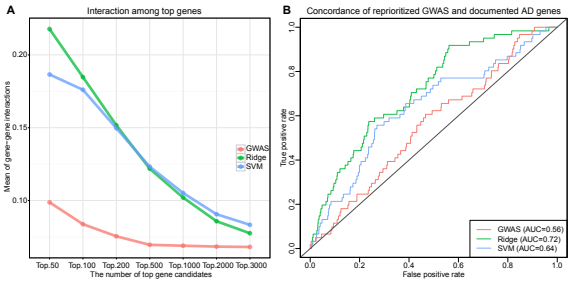

Fig. 2. Performance evaluation of re-prioritized results. (A) Mean of interactions among top predictions. (B) ROC curves.

To demonstrate the performance of re-prioritization, we accessed the mean interactions and the area under receiver-operator characteristic (ROC) curve (AUC) of re-prioritized genes from Ridge and SVM with original GWAS using 66 documented AD candidates as gold standard positives.

tnGWAS: Starting from a set of significant GWAS findings, tnGWAS includes their immediate neighbors in the result. tnGWAS hypothesizes that QT-relevant functional modules consist of top GWAS findings and their close neighbors. We extracted the interaction matrix containing connectivity measures between significant GWAS findings and all the genes, and identified genes highly interacted with $\geq$1 significant gene. In the experiment, we applied gene p-value $\leq$1e-7 to select significant GWAS findings, and interaction weight $\geq$0.3 to define strong connectivity. This yielded 4 significant genes and 120 highly interacted neighbors. In practice, we can include more top predictions and take more GWAS top neighbors to obtain larger scale candidate modules.

Identification of GWAS enriched modules: Machine learning based approaches were designed to yield top gene findings not only enriched by GWAS results but also densely connected; while tnGWAS was to identify top GWAS findings together with their immediate neighbors. For module identification, both frameworks offered a list of candidates for us to detect GWAS-enriched modules. We clustered top genes from above to firstly identify candidate modules. Since one gene could perform functions in multiple pathways, we employed the *Link Clustering* algorithm [16] on top genes to detect communities as clusters of links instead of nodes. The resulting candidate modules could be overlapping. Top GWAS findings were used to assess the enrichment of candidate module, while significantly enriched ones were identified as phenotype-relevant modules.

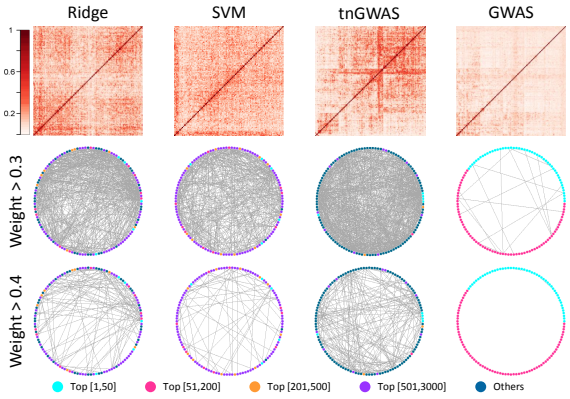

Fig. 3. Comparison of top 124 findings from Ridge, SVM, tnGWAS and original GWAS. Heatmaps show the complete interaction matrix of top predictions. Circular networks show interactions after filtering weak connections. Nodes in circular network are colored based on their ranks in GWAS result.

Different from previous bottom-up methods, these top-down strategies examine only a small number of candidate modules that are both highly connected and GWAS enriched, and thus can potentially help increase the statistical power.

Functional annotation: To assess the functional relevance of identified modules, we tested their over-representation on specific neurobiological functions and signalling pathways. We analyzed functional annotation using KEGG pathways and Gene Ontology Biological Process (GO-BP).

3. RESULTS

3.1. GWAS of hippocampal QT

GWAS was performed to examine the association between SNPs and the hippocampal FDG measure. Four SNPs were identified as significant using p$\leq$5E-7 (see Fig. 1 for the Manhattan plot), including two within *APOE*, one within *TOMM40* and one within *APOC1*. After mapping the SNPs to 17,881 protein coding regions, four genes were identified to be significant: *APOC1, APOE, PVRL2* and *TOMM40*.

3.2. Machine learning based re-prioritization

As mentioned earlier, top predictions from machine learning based re-prioritization would conserve both dense functional interaction and strong phenotype-relevance. Since tnGWAS did not assign ranks to top neighbors, we compared the top predictions from Ridge and SVM with original GWAS to assess their re-prioritization performances. Mean statistics of functional interactions and AUC were assessed on different scales of top predictions and shown in Fig. 2.

From Fig. 2(A), both Ridge and SVM yielded much stronger connectivity than GWAS. Dense interactions among top predictions demonstrated the advantage of network-based

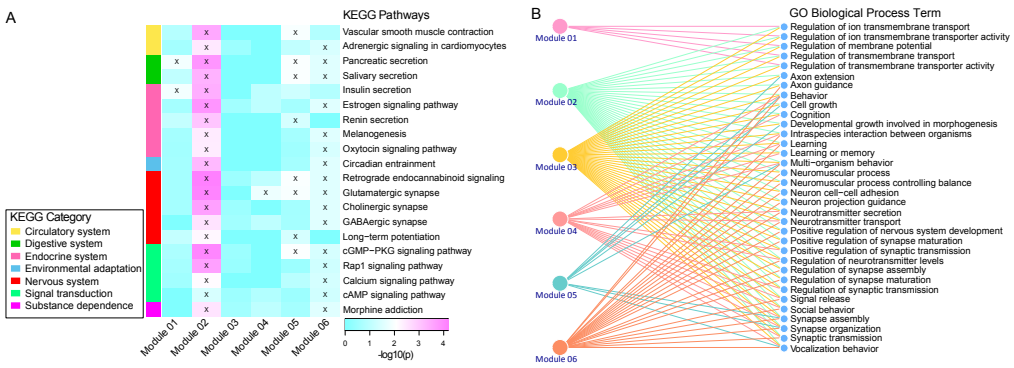

Fig. 4. Functional annotation of modules from Ridge.

integration. From Fig. 2(B), Ridge and SVM gained higher AUC than original GWAS, indicating the AD-relevance of top predictions by these new approaches. These support the idea that strong relationships exist between gene and phenotype, and that functionally-relevant genes are more likely to be interacted [17, 18, 19]. Ridge performed better than SVM in both evaluations, suggesting the value of the continuous z-scores over the significance status.

3.3. Hippocampus-relevant top predictions

We compared the functional connectivity of top findings among two machine learning-based methods, tnGWAS, and original GWAS. For a fair comparison, we focused on top 124 findings, since 124 is the number of top findings from tnGWAS (see section 2.4). Fig. 3 showed the heatmaps of connectivity and interaction networks using different thresholds where genes were colored by their original GWAS ranks.

Both heatmaps and networks demonstrate much denser interactions yielded by Ridge, SVM and tnGWAS than original GWAS. tnGWAS, due to the inclusion of immediate neighbors, gains the densest interaction. Top predictions from Ridge and SVM are also densely connected. In addition, they contain more top GWAS findings than tnGWAS (i.e., more nodes were colored by top GWAS findings). These observations reflect the different hypotheses behind the two strategies described earlier. Machine learning approaches seem to perform better as a whole as they integrate GWAS results and the tissue-specific network in a better fashion.

3.4. Hippocampus-relevant modules

We focus on top 124 predictions from Ridge given its top performance among four approaches. We preprocessed the functional connectivity network among these 124 genes to keep interactions with weights ≥ 0.2, and performed link clustering on this network. 21 modules were identified as candidates after removing those with < 10 genes. Six out of 21 were significantly enriched by top 50 GWAS findings; see Table 1. Functional annotation was applied to further examine functional relevance of identified modules. Fig. 4 shows (A) the KEGG pathway and (B) GO-BP enrichment results. All modules except Module 03 have significantly enriched pathways, some of which are related to neurodegenerative diseases (e.g., signal transduction like calcium signaling pathway had shown abnormality in many neurodegenerative disorders like AD [20]). Fig. 4(B) shows GO-BP terms that are significantly enriched by more than 2 modules. We could also find a large number of BP terms related to neurological system process (e.g., cognition), behavior (e.g., learning or memory), neurological system process (e.g., neuromuscular process), all of which had direct or indirect relationships with neurodegenerative diseases.

Table 1. Details of the identified modules from Ridge.

Ridge	Module ID	# of genes	GWAS Enrichment p-value (corrected)
Hippocampus	Module 01	21	2.68E-03
	Module 02	89	4.84E-04
	Module 03	26	7.85E-05
	Module 04	11	4.21E-02
	Module 05	22	3.10E-03
	Module 06	11	4.21E-02

4. DISCUSSIONS AND CONCLUSIONS

We have proposed two top-down module identification frameworks: machine learning based and tnGWAS. Both approaches integrate tissue specific functional interaction network with GWAS data to identify phenotype-relevant modules. Different from previous network-based module identification strategies, we start our search from the whole network to extract GWAS-relevant and highly interacted ones. Machine learning based approaches re-prioritize GWAS results, which can facilitate various relevant analyses. Subsequent enrichment assessment considers both tissue and GWAS specificities of the identified modules. Possible future directions include: (1) extending tnGWAS to re-rank identified top-neighbors using their GWAS statistics and interactions; and (2) applications to other tissues and omics data.

5. REFERENCES

[1] A. J. Saykin, L. Shen, et al., "Genetic studies of quantitative mci and ad phenotypes in adni: Progress, opportunities, and plans," *Alzheimers Dement*, vol. 11, no. 7, pp. 792–814, 2015.

[2] L. Shen, P. M. Thompson, et al., "Genetic analysis of quantitative phenotypes in ad and mci: imaging, cognition and biomarkers," *Brain Imaging Behav*, vol. 8, no. 2, pp. 183–207, 2014.

[3] J. C. Lambert, C. A. Ibrahim-Verbaas, et al., "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease," *Nature Genetics*, vol. 45, no. 12, pp. 1452–U206, 2013.

[4] P. L. Jia, S. Y. Zheng, et al., "dmgwas: dense module searching for genome-wide association studies in protein-protein interaction networks," *Bioinformatics*, vol. 27, no. 1, pp. 95–102, 2011.

[5] N. Akula, A. Baranova, et al., "A network-based approach to prioritize results from genome-wide association studies," *Plos One*, vol. 6, no. 9, 2011.

[6] J. N. Hirschhorn, "Genomewide association studies–illuminating biologic pathways," *N Engl J Med*, vol. 360, no. 17, pp. 1699–701, 2009.

[7] T. Ideker and N. J. Krogan, "Differential network biology," *Mol Syst Biol*, vol. 8, pp. 565, 2012.

[8] L. L. Wang, T. Matsushita, et al., "Pinbpa: Cytoscape app for network analysis of gwas data," *Bioinformatics*, vol. 31, no. 2, pp. 262–264, 2015.

[9] S. E. Baranzini, N. W. Galwey, et al., "Pathway and network-based analysis of genome-wide association studies in multiple sclerosis," *Human Molecular Genetics*, vol. 18, no. 11, pp. 2078–2090, 2009.

[10] C. S. Greene, A. Krishnan, et al., "Understanding multicellular function and disease with human tissue-specific networks," *Nat Genet*, vol. 47, no. 6, pp. 569–76, 2015.

[11] A. Song, J. Yan, et al., "Network-based analysis of genetic variants associated with hippocampal volume in alzheimer's disease: a study of adni cohorts," *BioData Min*, vol. 9, pp. 3, 2016.

[12] K. Ishii, T. Soma, et al., "Comparison of regional brain volume and glucose metabolism between patients with mild dementia with lewy bodies and those with mild alzheimer's disease," *Journal of Nuclear Medicine*, vol. 48, no. 5, pp. 704–711, 2007.

[13] L. Mosconi, W. H. Tsui, et al., "Multicenter standardized f-18-fdg pet diagnosis of mild cognitive impairment, alzheimer's disease, and other dementias," *J of Nuclear Medicine*, vol. 49, no. 3, pp. 390–398, 2008.

[14] S. Purcell, B. Neale, et al., "Plink: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, no. 3, pp. 559–75, 2007.

[15] D. Nam, J. Kim, and ohters, "Gsa-snp: a general approach for gene set analysis of polymorphisms," *Nucleic Acids Res*, vol. 38, pp. W749–54, 2010.

[16] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761–U11, 2010.

[17] K. A. Pattin and J. H. Moore, "Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases," *Human Genetics*, vol. 124, no. 1, pp. 19–29, 2008.

[18] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases," *Nature Reviews Genetics*, vol. 10, no. 6, pp. 392–404, 2009.

[19] M. Emily, T. Mailund, et al., "Using biological networks to search for interacting loci in genome-wide association studies," *European Journal of Human Genetics*, vol. 17, no. 10, pp. 1231–1240, 2009.

[20] I. Bezprozvanny, "Calcium signaling and neurodegenerative diseases," *Trends Mol Med*, vol. 15, no. 3, pp. 89–100, 2009.