Bioinformatics, 33(20), 2017, 3250–3257
doi: 10.1093/bioinformatics/btx344
Advance Access Publication Date: 20 July 2017
Original Paper

Systems biology

Tissue-specific network-based genome wide study of amygdala imaging phenotypes to identify functional interaction modules

Xiaohui Yao[1,2], Jingwen Yan[1,2], Kefei Liu[2], Sungeun Kim[2,3], Kwangsik Nho[2], Shannon L. Risacher[2], Casey S. Greene[4], Jason H. Moore[5], Andrew J. Saykin[2] and Li Shen[1,2,*], for the Alzheimer's Disease Neuroimaging Initiative[†]

[1]Department of BioHealth Informatics, Indiana University School of Informatics & Computing, Indianapolis, IN 46202, USA, [2]Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN 46202, USA, [3]Department of Electrical and Computer Engineering, SUNY Oswego, NY 13126, USA, [4]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA and [5]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed.
[†]See ADNI Acknowledgements.

Associate Editor: Alfonso Valencia

Received on November 30, 2016; revised on April 16, 2017; editorial decision on May 25, 2017; accepted on May 26, 2017

Abstract

Motivation: Network-based genome-wide association studies (GWAS) aim to identify functional modules from biological networks that are enriched by top GWAS findings. Although gene functions are relevant to tissue context, most existing methods analyze tissue-free networks without reflecting phenotypic specificity.

Results: We propose a novel module identification framework for imaging genetic studies using the tissue-specific functional interaction network. Our method includes three steps: (i) re-prioritize imaging GWAS findings by applying machine learning methods to incorporate network topological information and enhance the connectivity among top genes; (ii) detect densely connected modules based on interactions among top re-prioritized genes; and (iii) identify phenotype-relevant modules enriched by top GWAS findings. We demonstrate our method on the GWAS of [^{18}F]FDG-PET measures in the amygdala region using the imaging genetic data from the Alzheimer's Disease Neuroimaging Initiative, and map the GWAS results onto the amygdala-specific functional interaction network. The proposed network-based GWAS method can effectively detect densely connected modules enriched by top GWAS findings. Tissue-specific functional network can provide precise context to help explore the collective effects of genes with biologically meaningful interactions specific to the studied phenotype.

Availability and implementation: The R code and sample data are freely available at http://www.iu.edu/shenlab/tools/gwasmodule/

Contact: shenli@iu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) have been performed to identify genetic markers such as single nucleotide polymorphisms (SNPs) that are associated with common diseases. In brain imaging genetics, an emerging field that studies how genetic variation influences brain structure and function, GWAS also have discovered genes susceptible to brain imaging quantitative traits (QTs) (Lambert *et al.*, 2013; Saykin *et al.*, 2015; Shen *et al.*, 2014). Each identified imaging QT locus (iQTL), however, often has a small effect size and is hard to be individually interpreted. These iQTLs can potentially interact with one another to jointly have an impact on QTs. To address this challenge, integrative analysis of GWAS data with prior-knowledge has gained recent attention to test collective effects of multiple genes on targeted phenotypes. Using biological networks and pathways as prior knowledge, construction and identification of functionally interacted network modules have been performed to discover phenotype-relevant network modules enriched by the GWAS findings. This promising strategy can potentially enhance the statistical power of the GWAS and help biological interpretation (Akula *et al.*, 2011; Hirschhorn, 2009; Ideker and Krogan, 2012; Jia *et al.*, 2011; Wang *et al.*, 2015).

Existing module identification studies typically search for disease- or QT-relevant modules by mapping GWAS statistics onto a functional interaction network. After that, candidate modules are formed across the entire network and evaluated on whether they are enriched by the GWAS findings. A successful example is dense module GWAS (dmGWAS) (Jia *et al.*, 2011), which first loads gene-level p-values onto human protein–protein interaction network as node weights, then applies dense module searching strategy to identify modules that locally maximize the proportion of genes with small enough *P*-values. Network interface miner for multigenic interactions (NIMMI) is another network-based GWAS approach (Akula *et al.*, 2011), where phenotype-relevant modules are constructed from high-scored genes and their scores are computed by combining GWAS *P*-values with node weights calculated based on their network connectivity. The integrative protein-interaction-network-based pathway analysis (iPINBPA) method is also a network-based GWAS approach (Wang *et al.*, 2015) and is an extension of the original PINBPA (Baranzini *et al.*, 2009). It starts from a seed and expands the module by adding one neighbor at a time to reach an aggregate score meeting a given statistical significance. *Note that all these approaches employ a bottom-up strategy that examines a large number of candidate modules in order to identify enriched ones, and their efficiencies could become suboptimal when large-scale networks are present.*

Almost all the network-based GWAS are using tissue-free interaction networks such as the human PPI network without taking tissue specificity into consideration. The precise functions of genes are highly related to their tissue context, and human diseases often result from the disordered interplay of tissue-specific processes (Greene *et al.*, 2015). Recently, tissue-specific genome-wide functional interaction networks have been constructed in order to identify the changing functional roles of genes across tissues (Greene *et al.*, 2015). One application of tissue specific networks is to re-prioritize disease-gene associations by constructing a support vector machine (SVM) classifier to re-rank GWAS results based on tissue-specific network information. This strategy is named as NetWAS, and has been applied to analyze hippocampus volume in Alzheimer's disease and demonstrated that tissue-specific networks could provide helpful context for understanding complex human diseases (Song *et al.*, 2016). *Note that SVM classification requires a pre-defined threshold to partition GWAS P-values into significant and nonsignificant groups, and important information embedded in the continuous spectrum of these P-values get lost during the procedure.*

With the above observations, we expand the NetWAS work into a new framework to achieve two goals at one time: (i) introduce regression models in addition to classification models for re-prioritizing GWAS results with network information; (ii) use the re-prioritized results to identify GWAS-enriched network modules. In short, we propose an innovative phenotype-relevant module identification method by integrating GWAS data and tissue-specific network with effective machine learning models. First, in addition to traditional NetWAS using SVM, we re-prioritize GWAS results by constructing two regression models (support vector regression and ridge regression) using tissue-specific functional interaction network as features and continuous GWAS *P*-values as responses. We then extract densely connected modules from top NetWAS findings based on their functional interactions. Finally, GWAS findings are used to test the enrichment significance on these candidate modules to identify phenotype-relevant ones.

Compared with traditional GWAS-based module identification methods and SVM-based NetWAS, *the novelty of the proposed new framework is threefold*: (i) Our framework expands the NetWAS scope from re-prioritizing GWAS findings to module identification. (ii) Our framework introduces regression models into NetWAS to embrace the complete coverage of the continuous *P*-value spectrum. (iii) Our framework offers a more efficient, top-down strategy to identify phenotype-relevant network modules, given that the top findings from NetWAS are designed to be both GWAS-enriched and densely connected.

To show the effectiveness of the proposed framework, we compare support vector regression (SVR) and ridge regression (Ridge) with SVM to illustrate that continuous GWAS *P*-values supply more valuable information than binary significant/non-significant labels. We also compare the NetWAS re-prioritized results with original GWAS findings to show that the former is more densely connected than the latter. Identified modules are further tested for functional association by KEGG pathway, Gene Ontology Biological Process and Online Mendelian Inheritance in Man (OMIM) disease databases, to demonstrate that tissue-specific networks may provide helpful context for understanding the mechanisms behind complex diseases.

2 Materials and methods

To demonstrate the proposed NetWAS-based method for identifying phenotype-relevant functional interaction modules, we apply it to the amygdala imaging genetic analysis in the study of Alzheimer's disease (AD). The amygdala is located in the medial temporal lobe region of the brain and has been implicated in emotional processes, survival instincts and aspects of memory, especially for emotional components. Analyses on amygdala have indicated that it is prominently related to AD and its progression (Fjell *et al.*, 2010; Palmer *et al.*, 2007; Poulin *et al.*, 2011) and has been used to assist the clinical diagnosis of AD (Tang *et al.*, 2014). Studies on fluorodeoxyglucose [^{18}F]FDG-PET have demonstrated different usage patterns of glucose metabolism in amygdala between AD and healthy control subjects (Johnson *et al.*, 2012).

2.1 Imaging data, genotyping data and GWAS

The imaging and genotyping data used for GWAS were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI)

Table 1. Participant characteristics: HC = Healthy Control; SMC = Significant Memory Concern; EMCI = Early Mild Cognitive Complaint; LMCI = Late Mild Cognitive Complaint; AD = Alzheimer's Disease

Subject	HC	SMC	EMCI	LMCI	AD
Number	244	86	280	247	132
Gender (M/F)	124/120	34/52	159/121	146/101	79/53
Age (mean±std)	74.02±5.72	71.86±5.61	71.16±7.29	72.31±7.63	73.32±7.34
Education (mean±std)	16.44±2.66	16.85±2.63	16.06±2.66	16.24±2.81	16.19±2.72

database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Preprocessed [^{18}F]FDG-PET scans were downloaded from the LONI website (adni.loni.usc.edu), then aligned to each participant's same visit scan and normalized to the Montreal Neurological Institute (MNI) space as $2 \times 2 \times 2$ mm voxels. FDG measurements of amygdala (left and right) were further extracted based on the MarsBaR AAL atlas. Genotype data of both ADNI-1 and ADNI-GO/2 phases were also obtained from LONI, and quality controlled, imputed and combined as described in (Kim *et al.*, 2013). 989 non-Hispanic Caucasian participants (Table 1) with complete baseline FDG amygdala measurements were studied.

Associations between amygdala measures and SNPs (allelic dosage) were examined by performing GWAS using PLINK (Purcell *et al.*, 2007), where a linear regression model with sex, age and education as covariates was employed. To facilitate the subsequent network-based analysis, a gene-level *P*-value was determined as the 2nd smallest *P*-value of all SNPs located in ±20K bp of the gene (Nam *et al.*, 2010). In addition, 10 GWAS permutations were performed to illustrate that only the original GWAS data yielded promising findings.

2.2 Amygdala-specific functional interaction network

Genome-wide functional interaction networks for specific human tissues and cell types had been generated to specialize protein functions and interactions of specific human tissues by integrating a collection of datasets covering thousands of experiments contained in more than 14 000 distinct publications (Greene *et al.*, 2015). The genome-scale maps provided a detailed portrait of protein functional interactions in specific human tissues and cell lineages ranging from B lymphocytes to the whole brain. Amygdala tissue-specific genome-wide interaction network was downloaded from the Genome-scale Integrated Analysis of gene Networks in Tissues (GIANT) website (http://giant.princeton.edu/). A functional interaction network among genes was extracted after mapping to GWAS results. The weights of interactions range from 0 to 1, where larger measures represent stronger interactions.

2.3 Alzheimer's disease risk genes

A list of documented AD risk genes were collected to evaluate the re-prioritization results from multiple machine learning models. Here we integrated totally 66 AD-relevant genes collected from three resources: 24 susceptibility genes from a large meta-analysis of AD (Lambert *et al.*, 2013), 15 AD-relevant genes from Online Mendelian Inheritance in Man Disease database (OMIM), and 40 significant candidates from the AlzGene database (http://www.alzgene.org/).

2.4 Module identification method

Our proposed phenotype-relevant module identification method is a top-down approach integrating tissue-specific functional interaction network and GWAS results. We hypothesize that GWAS significant findings are enriched among nominally significant and functional-relevant genes. Below, we describe the details of the proposed method. See Figure 1 for the workflow.

2.4.1 NetWAS re-prioritization of GWAS results

Following (Song *et al.*, 2016), we re-prioritized GWAS results by integrating the amygdala-specific functional interaction network using SVM-based NetWAS. Briefly, the functional network connectivity matrix was used as feature data and significant/non-significant status based on the nominal $P < 0.01$ was used as class label.

In addition to SVM, we trained two separate regression models, support vector regression (SVR) and ridge regression (Ridge). In both models, we used the functional network connectivity matrix as feature data and continuous GWAS *P*-values as responses. SVR, different from SVM, does not require a pre-defined threshold to convert *P*-values to a binary variable indicating significant/non-significant status. SVR is designed to find a hyperplane that has a deviation of at most ε from the actual data. Ridge is a widely used linear regression approach using the L_2-norm based regularization to stabilize the result.

To train SVM, SVR and Ridge models, we first selected a set of genes with *P*-value < 0.01, denoted as $\mathbb{A}$, then randomly partitioned the remaining genes (i.e. *P*-value ≥ 0.01) into five equal groups $\mathbb{B}^{(t)}, t = 1, \ldots, 5$. We combined $\mathbb{A}$ with each $\mathbb{B}^{(t)}$ to construct gene set $\mathrm{C}^{(t)}$ for model training. That is, gene-level *P*-values of $\mathrm{C}^{(t)}$ were used as responses (positive/negative labels for SVM), while interactions between genes from $\mathrm{C}^{(t)}$ and all genes from the functional network were used as features. In experiments, *we employed* $-\log(p)$ *values instead of original P-values as regression response*. For the prediction part, the features are the entire interaction network across all genes. Five models $M^{(t)}$, $t = 1, \ldots, 5$ were trained for each method $\in$ {SVM, SVR, Ridge} and then applied to predict the responses for all genes. Finally, genes were re-prioritized based on their mean predictions (SVR and Ridge) or distances from hyperplane (SVM) across five sets of results. See Supplementary Materials for detailed implementation.

To demonstrate the effectiveness of the patterns discovered from the real data, we also trained these models on permuted GWAS results using the same strategy. We used the area under the receiver operating characteristic (ROC) curve (AUC) to compare the re-prioritization performance obtained from the original GWAS data with those from permuted GWAS data. Similar to (Song *et al.*, 2016), ROC curves and AUCs were calculated using 66 documented AD candidates as gold standard positives (See Supplementary Materials for more details). In addition, mean statistics of functional

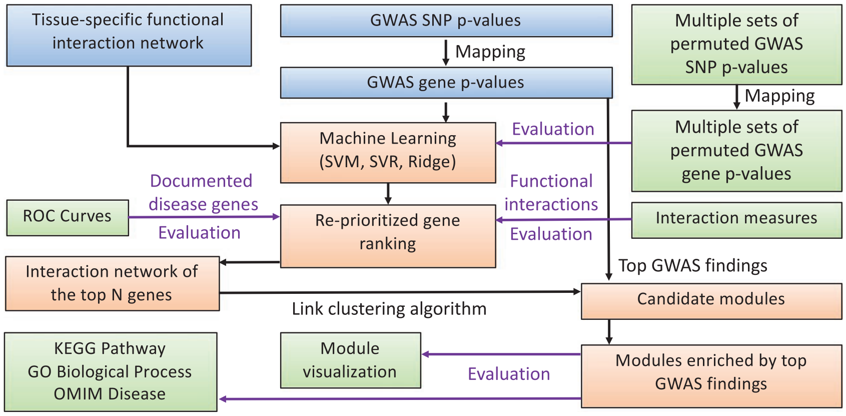
Tissue-specific functional interaction network
GWAS SNP p-values
Mapping
GWAS gene p-values
Multiple sets of permuted GWAS SNP p-values
Mapping
Multiple sets of permuted GWAS gene p-values
Evaluation
Machine Learning (SVM, SVR, Ridge)
Documented disease genes
Functional interactions
ROC Curves
Evaluation
Re-prioritized gene ranking
Evaluation
Interaction measures
Interaction network of the top N genes
Top GWAS findings
Link clustering algorithm
Candidate modules
KEGG Pathway
GO Biological Process
OMIM Disease
Module visualization
Evaluation
Modules enriched by top GWAS findings

Fig. 1. The workflow for identifying functional interaction modules from the tissue-specific network using GWAS findings

interaction measures among top genes were used to estimate the functional relevance of these genes.

2.4.2 Identification of GWAS-enriched modules

The goal of the NetWAS re-prioritization is twofold: (i) The original GWAS gene ranking is used to supervise the training of the classification and regression models and ensure that the top genes in the re-prioritization remain GWAS-enriched; (ii) tissue-specific functional interaction connectivity matrix is used as data to train the models and encourage genes with similar interactions to be re-prioritized with similar ranks. Thus NetWAS is designed to yield top gene findings that are both GWAS-enriched and densely connected; and these top genes become the candidates for us to identify GWAS-enriched network modules.

We performed clustering on these top genes to first identify candidate modules. Since one gene could play roles in multiple pathways or functional modules, we applied the Link Clustering algorithm (Ahn *et al.*, 2010) to detect communities as groups of links rather than nodes. The resulting candidate modules consisted of only top NetWAS genes and could overlap each other. After that, top GWAS findings were used to test each candidate module. Only those modules significantly enriched by the GWAS results were identified as phenotype-relevant ones. See Supplementary Materials for details.

As mentioned earlier, many existing network-based GWAS approaches employ a bottom-up strategy that examines a large number of candidate modules in order to identify enriched ones, and their efficiencies could become suboptimal when large-scale networks are present. Our module identification approach proposed above overcomes this limitation. On one hand, it examines only a small number of candidate modules generated from clustering the top NetWAS findings. On the other hand, the NetWAS strategy is designed to yield promising candidate modules with strong potential to be densely connected and phenotype-relevant.

2.4.3 Functional evaluation and visualization

To determine the functional relevance of the identified modules, we tested whether genes from each module were overrepresented for specific neurobiological functions, signaling pathways or complex neurodegenerative diseases. We performed three types of functional annotation analyses using KEGG pathway, Gene Ontology Biological Process (GO-BP) and OMIM disease databases respectively. For identified modules, they could be visualized directly or extended to include neighboring genes in the tissue-specific functional interaction network. We selected one example module and visualized it as well as its extension using GIANT (http://giant.princeton.edu/) to show its dense functional interactions.

3 Results

We applied our NetWAS-based module identification framework, using amygdala-specific functional interaction network, to the GWAS findings of the FDG-PET measures in the left and right amygdala regions in an AD study. We compared the performances of different machine learning models, as well as those using the original and permuted GWAS results. We evaluated the functional relevance of the identified modules and discussed their relationships with neurobiological or neurodegenerative funtions and diseases. Below we report and discuss our results.

3.1 GWAS of amygdala QTs

GWAS were performed to examine genetic associations between 5 574 300 SNPs and FDG measures in the left and right amygdalas. Using $P \leq 5E\text{-}8$ as the threshold, nine SNPs were identified to be significantly associated with the average FDG-PET measure in the left amygdala (see Supplementary Fig. S1 in the Supplementary Materials for the Manhattan plot), including two within the *APOE* gene (rs429358 with $P = 1.99E\text{-}11$, rs769449 with $P = 3.28E\text{-}09$), one within the *SDK1* gene (rs148359108 with $P = 2.02E\text{-}09$), one between the *APOE* and *APOC1* gene (rs10414043 with $P = 8.56E\text{-}09$), and five within the *APOC1* gene (rs7256200 with $P = 8.56E\text{-}09$, rs12721051 with $P = 1.11E\text{-}08$, rs56131196 with $P = 1.11E\text{-}08$, rs4420638 with $P = 1.11E\text{-}08$ and rs73052335 with $P = 3.50E\text{-}08$). No significant findings were identified on the right side.

After mapping the 2nd smallest SNP-level *P*-values to genes (Nam *et al.*, 2010) using hg19 gene annotation, gene-based *P*-values were obtained for 24 766 genes and transcripts. Using $P \leq (0.05/24\,766) = 2.02E\text{-}6$ as the threshold, the *APOC1*, *APOE*, *PVRL2*, *TOMM40* and *APOC1P1* genes were identified to be significantly associated with the average FDG-PET measure in the left amygdala. Note that *PVRL2* and *APOC1P1* were identified since some of significant SNPs were within ±20K bp of their boundaries. In addition, we compared the 2nd smallest strategy with three other gene-based association approaches (see Supplementary Materials).

All the findings except *SDK1* are either from or proximal to the *APOE* region, which is the best known genetic risk region in AD. *SDK1*, which is located in 7p22.2 and encodes protein sidekick-1 (a member of the immunoglobulin superfamily), shows an association with the average FDG-PET measure of left amygdala, including a significant hit at the SNP level (rs148359108 with $P = 2.02E\text{-}09$), and a nearly significant one at the gene level ($P = 3.35E\text{-}06$). *SDK1* was shown to specifically phosphorylate 14-3-3ζ at serine 58 (Hamaguchi *et al.*, 2003), where the latter played an important role in amygdala cell death (Jeong *et al.*, 2010). *SDK1* also showed high expression in medial amygdala relative to other tissues from the Allen Brain Atlas Adult Mouse Brain Tissue Gene Expression Profiles dataset (http://www.brain-map.org/). The connection between *SDK1* and AD-related amyloid and glucose metabolism markers in the amygdala region warrants further investigation.

3.2 NetWAS re-prioritization

Amygdala-specific functional interaction network among 25 825 nodes was downloaded from GIANT, with interaction weights ranging from 0 to 1. There were totally 20 168 nodes used in our analysis after matching GWAS genes and transcripts with those from the network. After preprocessing, we obtained an amygdala-specific genome-wide functional interaction matrix with size of 20 168 × 20 168 and two lists of 20 168 gene-level *P*-values for left and right amygdala QTs respectively. In addition, GWAS were performed 10 times on permuted data for each of the bilateral amygdala measures. The same procedure was applied to the permuted data as the real data, in order to demonstrate that only the GWAS findings from the real data yielded promising results.

Five sets of regression predictions by SVR and Ridge or classification decision values by SVM (i.e. distances from the separating hyperplane) were obtained from running these machine learning models using functional interaction connectivity matrix as the feature data and the GWAS results as regression responses or classification labels. For each model, genes were re-prioritized based on their average regression predictions or classification decision values across five experiments, on both original and permuted GWAS results.

As we hypothesized, top predictions would conserve both strong functional interaction and high phenotype-relevance (i.e. AD-relevance in this work, given amygdala FDG measures as promising AD biomarkers). We compared the re-prioritization performances of three machine learning models and GWAS using both original and permuted data.

Figure 2(A, B) shows the ROC curves and the AUC performances. *For the original data*, the re-prioritization results of all three NetWAS models demonstrated much higher concordance with documented AD risk genes than the GWAS findings. This indicates that integration of tissue-specific functional interaction network with GWAS can promote the identification of phenotype-relevant genes. *For the permuted data*, where the mean and standard deviation of AUCs together with one example ROC are shown for each model, no high concordance with AD genes was achieved by either GWAS or any NetWAS model. This suggests that the NetWAS procedure is not biased and only original data can yield meaningful findings. In addition, original GWAS and permuted GWAS obtained similar AUCs, showing the limited power of GWAS alone on the detection of disease risk markers. Ridge, although showing similar AUC with SVR and SVM, gained higher true positive rate and lower false positive rate at the beginning of the ROC. That is, Ridge gained higher concordance when taking look at top re-prioritized results.

Figure 2(C, D) shows the mean functional interaction of the top findings. We used a series of thresholds from top 50 to top 3000 (of note, ~3000 genes with *P*-value < 0.01 were identified for either left or right amygdala) to extract different scales of top genes as well as their interaction matrix. NetWAS approaches, no matter whether using original or permuted data, clearly demonstrated denser interactions among top findings than GWAS. This confirms our hypothesis that NetWAS yields more densely connected top findings.

3.3 Amygdala-relevant top predictions

We investigated top 50 re-prioritized genes obtained from three machine learning models, and compared their functional interactions in detail. Figure 3(A, B) showed heatmaps of interaction relationships among top genes and interaction networks based on different thresholds for left and right amygdalas, respectively. Taking left amygdala as example, each row shows results from different methods: Ridge, SVR, SVM and GWAS. Heatmaps show interaction matrices using the data from amygdala functional network without any filtering. Two interaction networks among top 50 genes after filtering out weak interactions using different scales (here using weights ≥ 0.1 and 0.2 as thresholds) are shown. In interaction networks, nodes are colored by their ranks in the original GWAS.

Both heatmaps and networks show much denser interactions among top 50 findings from three models than original GWAS under any scale of filtering. That facilitates the promise of our proposed method for comprehensively examining the disease-relevant genes and interactions between them. Ridge, compared with SVR and SVM, yielded much higher interactions (network density across multiple scales) and also obtained more GWAS top genes (more nodes are colored by top GWAS findings). This, combined with statistics summary from Figure 2, indicates the outstanding performance of Ridge.

3.4 Amygdala-relevant modules

The results shown above demonstrate the phenotype-relevance and dense functional interactions of the top findings obtained from integrating amygdala-specific interaction network and amygdala FDG GWAS result. We identified candidate network modules based on the interaction matrix of these top findings to make sure that they conserved high within-module connectivity. We analyzed top 50 findings from Ridge-based NetWAS given its prominent performance. In candidate module identification, only interactions with weights ≥ 0.1 were considered while weak connections were removed. We identified five modules: four from left amygdala, and one from right amygdala. All five modules were significantly enriched by top 50 GWAS findings. Supplementary Table S1 (see Supplementary Materials) shows details of these modules.

In this work, we applied our method on only top 50 predictions and used a relatively stringent selection of GWAS significant findings (top 50) to test phenotype-relevance of the candidate modules. In practice, we could include more top predictions into module identification to obtain more candidate modules and also take a larger number of GWAS top findings into enrichment test to relax phenotype-relevance.

3.5 Functional annotation of the identified modules

Functional annotation was performed to further investigate functional relevance of the identified modules. We performed pathway enrichment analysis from three aspects: (i) functional pathways, (ii) GO terms and (iii) diseases, based on KEGG pathway, GO-BP and OMIM disease databases respectively.

Supplementary Figure S2 (see Supplementary Materials) shows the KEGG pathway enrichment results mapped to 19 categories. From the results, two modules from left amygdala and one module from right amygdala have a number of significant functional enrichments, while

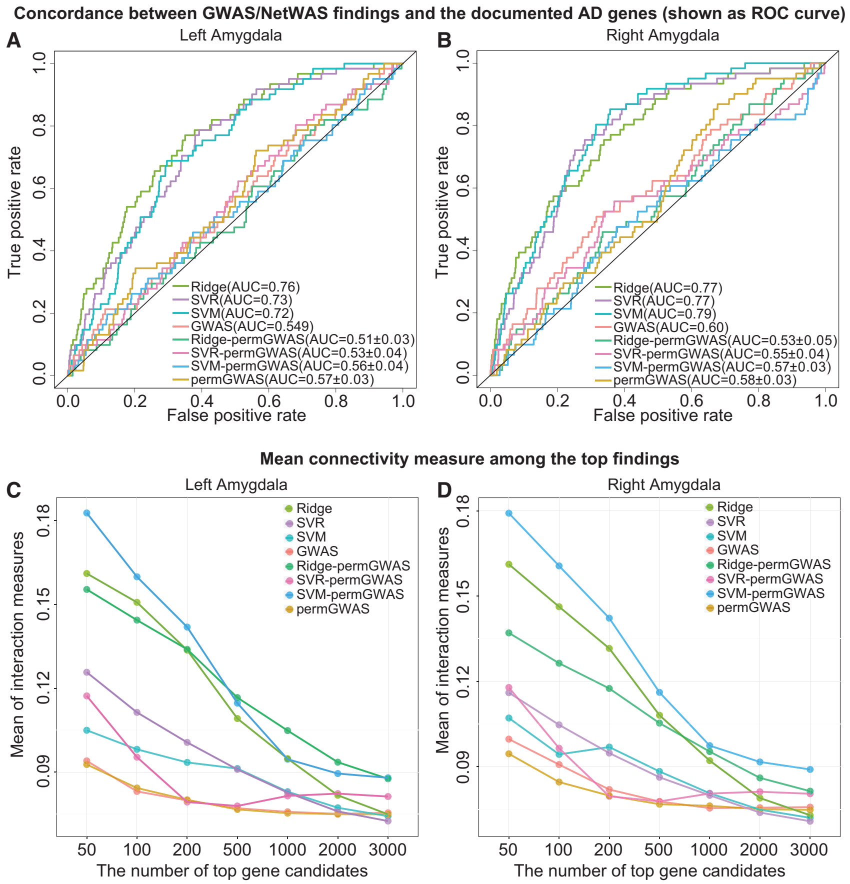

Fig. 2. Performance evaluation of re-prioritization results. (**A–B**): ROC curves with AUC results on left and right amygdalas, respectively, to measure the concordance between the GWAS/NetWAS findings and the documented AD genes. For each analysis on permuted GWAS, the mean and standard deviation of AUCs together with one example ROC are shown. (**C–D**): Mean interaction measures among top N findings (N ranging from 50 to 3000) on left and right amygdalas, respectively

the other two modules of left amygdala do not have obvious KEGG functional enrichment. Several enriched pathways are directly related to the neurodegenerative disease and its development, e.g. Alzheimer's disease enriched in Modules 03 and 04 and Huntington's disease enriched in Module 04. A number of pathways from three large categories are enriched by one or more modules, and these categories are endocrine system, nervous system and signal transduction. These major categories have been studied and shown close relation to AD. For example, the endocrine and the nervous system were highly related as hormones played a role in maintaining brain homeostasis at the senile age which might help explain the gender difference in AD (Blair *et al.*, 2015; Mielke *et al.*, 2014; Peri and Serio, 2008). Signal transduction like calcium signaling pathway (Modules 04 and 05) playing key role in short- and long-term synaptic plasticity, had shown abnormality in many neurodegenerative disorders like Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis (ALS), Huntington's disease and so on (Bezprozvanny, 2009). Neuroinflammation emerged as an important component of AD pathology recently, and immune system indicated a crucial role in the progression of AD (Heneka *et al.*, 2015). Platelet activation, enriched in Modules 04 and 05, had been studied about its involvement in neuroinflammatory diseases such as AD through enzymatic activities to generate amyloid-β peptides (Gowert *et al.*, 2014).

Supplementary Figure S3 (see Supplementary Materials) shows top GO-BP enriched terms for all five modules. As Modules 03-05 had significantly enriched a large number of GO-BP terms, only top 20 of each module were selected. Here only GO-BP terms that are significantly enriched (corrected P-value<0.05) by >1 module are listed and linked with corresponding modules. Here we observe that a large number of BP terms are related to neurological system process (e.g. cognition, learning), behavior (e.g. learning or memory), nervous system development (e.g. positive regulation of neuron projection development) and signal (e.g. regulation of synaptic transmissions). All of these have direct or indirect relationships with neurodegenerative diseases or phenotypes.

OMIM disease enrichment analysis results are shown in Supplementary Table S2 (see Supplementary Materials), where three modules (Modules 01, 02 and 04) are significantly enriched by various types of diseases including heart disease (Myocardial infarction), cancer (Prostate cancer), mental disorders (Autism), eye disease (Macular degeneration) and neurodegenerative diseases (Alzheimer's disease). A number of studies suggested that there exist connections between heart diseases and dementia including AD (Heneka *et al.*, 2015; Licastro *et al.*, 2011). Epidemiological studies had shown a reciprocal inverse relationships between cancer and neurodegeneration according to abnormal cell growth and cell loss in common (Nudelman *et al.*, 2014; Realmuto *et al.*, 2012).

3.6 Module visualization and extension

Given the identified phenotype-relevant modules, we visualized functional interactions among genes as a network and extended the module by including genes having close connections with elements inside the module. We show Module 04 as an example given its

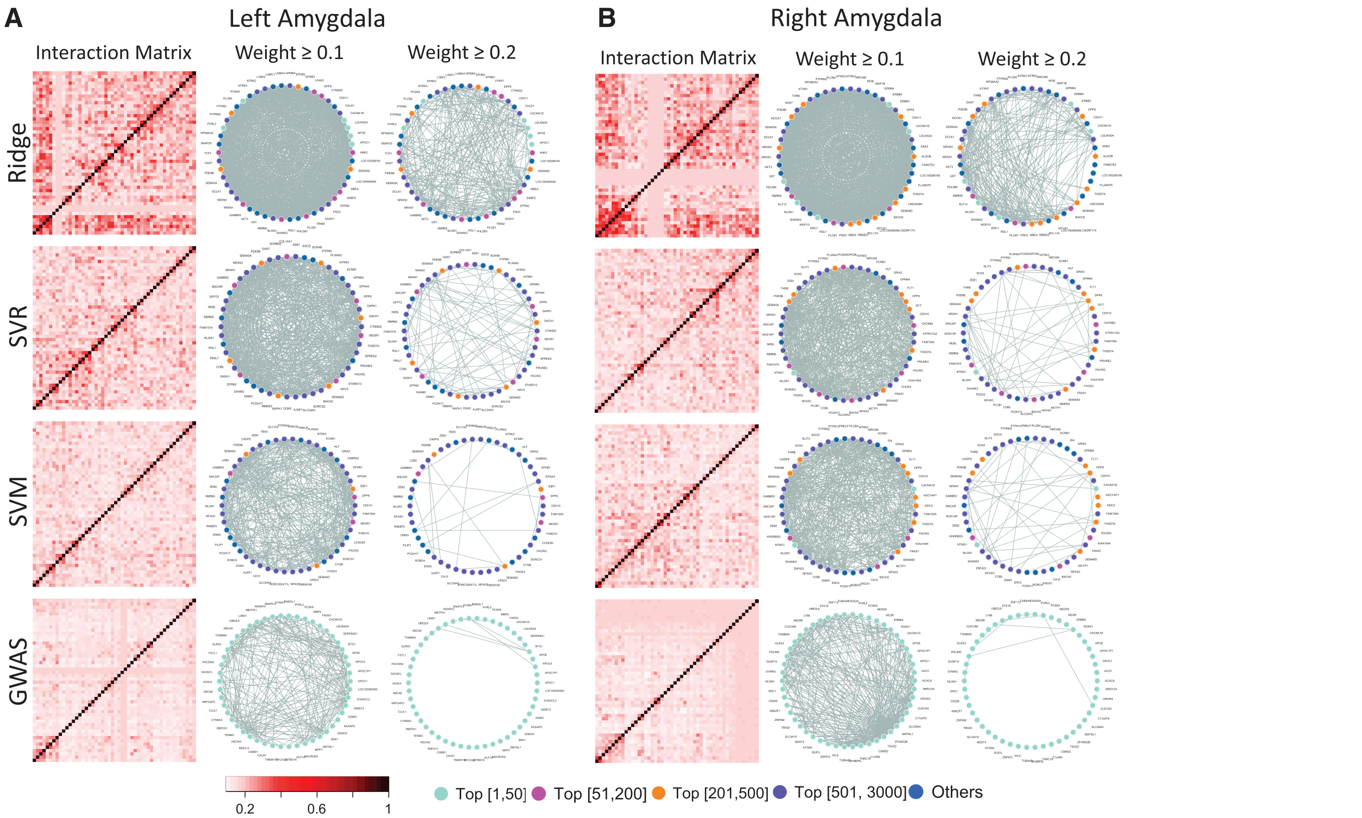

Fig. 3. Comparison of top 50 findings by three NetWAS re-prioritization methods (Ridge, SVR and SVM) and the original GWAS. (**A**) and (**B**) represent results on left and right amygdalas, respectively. Heamaps show the complete interaction matrix of top predictions. Circular networks show interactions between genes after filtering weak connections. Nodes in circular network are colored by their ranking in the original GWAS

small size as well as functional enrichment performance. Supplementary Figure S4(A) and (B) in Supplementary Materials respectively show Module 04 and an expanded version of Module 04 by including additional genes with minimum relationship confidence 0.2 using GIANT. Functional annotation of the expanded version of Module 04 has been tested and shown in Supplementary Table S3 in Supplementary Materials.

4 Discussion

We have proposed a top-down module identification method by integrating tissue-specific functional interaction network with imaging GWAS results to detect phenotype-relevant modules for better mechanistic understanding of complex diseases. At the global level, machine learning models were applied to re-prioritize genes which facilitates the detection of genes with both phenotype-relevance and dense interactions. After that, candidate modules were extracted using link community clustering algorithm. At the local level, each candidate module was tested for enrichment significance using GWAS findings. This study is among the first to incorporate tissue-specific context with GWAS data to understand underlying functional relevance in a precise way.

Our strategy is different from previous network module identification methods that define and examine candidate modules by forming sub-networks based on individual genes (e.g. genes with promising *P* values or high scores). We start from the whole interaction network to re-rank genes so that the top findings are not only densely connected and but also enriched by highly scored genes. Machine learning methods can facilitate the re-prioritization using network data as features. This step makes use of both the functional network information and GWAS discoveries to ensure the phenotype-relevance and dense connection of the top re-prioritized genes. The second step is designed simply for assigning an enrichment score to each candidate module so that modules not enriched by GWAS findings can be filtered out. We treat the whole process as a single discovery step. In order to validate the findings, replication analysis in independent cohorts should be performed.

As to the NetWAS comparison among three machine learning based models on our data, Ridge performed better than SVR, and SVR generally outperformed SVM. This suggests that continuous GWAS *P*-values supply more valuable information than binary significant/non-significant labels. Re-prioritization results show the strength of the NetWAS framework from another perspective that top predictions hold denser interactions and are matched to more disease risk genes than GWAS findings. Our experimental results on permuted data also suggests that the NetWAS procedure is not biased and only original data can yield meaningful findings.

Given that we only have one tissue-specific network available for the studied phenotype, we are limited on validating the stability of the findings. In the future, if multiple tissue-specific interaction networks can be obtained independently for a studied tissue, stability study can be performed to check whether similar network modules can be identified from multiple networks.

5 Conclusions

We have proposed a top-down module identification method by integrating tissue-specific functional network with imaging GWAS results. We have demonstrated its effectiveness using real data from an imaging genetics study in Alzheimer's disease. Modules identified from our method conserve both dense interactions and high phenotype-relevance, showing the promise of the proposed method. This work can be further expanded towards several future directions. For example, one direction is to compare the proposed method with other existing module identification strategies to further evaluate its performance. Another direction is to apply this method to other tissues and brain regions for revealing tissue-specific genetic mechanisms for complex brain disorders.

Acknowledgements

See ADNI Acknowledgements in Supplementary Materials.

Funding

At Indiana University, this work was supported by National Institute of Health R01 EB022574, R01 LM011360, U19 AG024904, U54 AG054345, R01 AG19771, P30 AG10133, UL1 TR001108, R01 AG 042437, R01 AG046171, R03 AG050856 and R00 LM011384; NSF IIS-1117335; United States Department of Defense W81XWH-14-2-0151, W81XWH-13-1-0259 and W81XWH-12-2-0012; National Collegiate Athletic Association 14132004; Indiana Clinical and Translational Sciences Institute Strategic Pharma-Academic Research Consortium for Translational Medicine Program; and Alzheimer's Association, Michael J Fox Foundation and Alzheimer's Research UK BAND project. At University of Pennsylvania, the work was supported by National Institute of Health R01 LM011360, R01 LM009012 and R01 LM010098.

Conflict of Interest: none declared.

References

Ahn,Y.Y. *et al.* (2010) Link communities reveal multiscale complexity in networks. *Nature*, **466**, 761–764.

Akula,N. *et al.* (2011) A network-based approach to prioritize results from genome-wide association studies. *PloS One*, **6**, e24220.

Baranzini,S.E. *et al.* (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Gen.*, **18**, 2078–2090.

Bezprozvanny,I. (2009) Calcium signaling and neurodegenerative diseases. *Trends Mol. Med.*, **15**, 89–100.

Blair,J.A. *et al.* (2015) Hypothalamic-pituitary-gonadal axis involvement in learning and memory and Alzheimer's disease: more than just estrogen. *Front. Endocrinol.*, **6**, 45.

Fjell,A.M. *et al.* (2010) CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease. *J. Neurosci.*, **30**, 2088–2101.

Gowert,N.S. *et al.* (2014) Blood platelets in the progression of Alzheimer's disease. *PloS One*, **9**, e90523.

Greene,C.S. *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.*, **47**, 569–576.

Hamaguchi,A. *et al.* (2003) Sphingosine-dependent protein kinase-1, directed to 14-3-3, is identified as the kinase domain of protein kinase C delta. *J. Biol. Chem.*, **278**, 41557–41565.

Heneka,M.T. *et al.* (2015) Innate immunity in Alzheimer's disease. *Nat. Immunol.*, **16**, 229–236.

Hirschhorn,J. (2009) Genomewide association studies-illuminating biologic pathways. *N. Engl. J. Med.*, **360**, 1699–1701.

Ideker,T. and Krogan,N.J. (2012) Differential network biology. *Mol. Syst. Biol.*, **8**, 565.

Jeong,E.A. *et al.* (2010) Phosphorylation of 14-3-3zeta at serine 58 and neurodegeneration following kainic acid-induced excitotoxicity. *Anat. Cell Biol.*, **43**, 150–156.

Jia,P.L. *et al.* (2011) dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, **27**, 95–102.

Johnson,K.A. *et al.* (2012) Brain imaging in Alzheimer disease. *Cold Spring Harb. Perspect. Med.*, **2**, a006213.

Kim,S. *et al.* (2013) Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. *PLoS One*, **8**, e70269.

Lambert,J.C. *et al.* (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.*, **45**, 1452–1458.

Licastro,F. *et al.* (2011) Sharing pathogenetic mechanisms between acute myocardial infarction and Alzheimer's disease as shown by partially overlapping of gene variant profiles. *J. Alzheimers Dis.*, **23**, 421–431.

Mielke,M. *et al.* (2014) Clinical epidemiology of Alzheimer's disease: assessing sex and gender differences. *Clin. Epidemiol.*, **6**, 37–48.

Nam,D. *et al.* (2010) GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Res.*, **38**, W749–W754.

Nudelman,K.N.H. *et al.* (2014) Association of cancer history with Alzheimer's disease onset and structural brain changes. *Front. Physiol.*, **5**, 423.

Palmer,K. *et al.* (2007) Predictors of progression from mild cognitive impairment to Alzheimer disease. *Neurology*, **68**, 1596–1602.

Peri,A. and Serio,M. (2008) Neuroprotective effects of the Alzheimer's disease-related gene seladin-1. *J. Mol. Endocrinol.*, **41**, 251–261.

Poulin,S.P. *et al.* (2011) Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Res. Neuroimag.*, **194**, 7–13.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Realmuto,S. *et al.* (2012) Tumor diagnosis preceding Alzheimer's disease onset: is there a link between cancer and Alzheimer's disease?. *J. Alzheimers Dis.*, **31**, 177–182.

Saykin,A.J. *et al.* (2015) Genetic studies of quantitative MCI and AD phenotypes in ADNI: progress, opportunities, and plans. *Alzheimers Dement.*, **11**, 792–814.

Shen,L. *et al.* (2014) Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. *Brain Imaging Behav.*, **8**, 183–207.

Song,A. *et al.* (2016) Network-based analysis of genetic variants associated with hippocampal volume in Alzheimer's disease: a study of ADNI cohorts. *BioData Min.*, **9**, 3.

Tang,X.Y. *et al.* (2014) Shape abnormalities of subcortical and ventricular structures in mild cognitive impairment and Alzheimer's disease: detecting, quantifying, and predicting. *Hum. Brain Mapp.*, **35**, 3701–3725.

Wang,L.L. *et al.* (2015) PINBPA: Cytoscape app for network analysis of GWAS data. *Bioinformatics*, **31**, 262–264.