MINING HIGH-LEVEL BRAIN IMAGING GENETIC ASSOCIATIONS

Xiaohui Yao

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

March 2018

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Huanmei Wu, PhD, Chair

Doctoral Committee

Li Shen, PhD

January 16, 2018

Shiaofen Fang, PhD

Jingwen Yan, PhD

DEDICATION

Dedicated to my dear family

for all their love and support along the way

ACKNOWLEDGEMENTS

First and foremost I would like to express my deepest gratitude to my advisor, Dr. Li Shen, for guiding and supporting my research during the past five years. He has provided me with great research insights and exceptional enthusiasm as well as encouragement throughout my PhD study. This work would never be materialized without him.

I also give my thanks to all the committee members for supporting this thesis work and all the constructive suggestions and feedback, that have been very helpful to keep this work on track and finally make the goal accomplished timely.

I would like to express my appreciation to all my colleagues and professors from Center for Neuroimaging: Dr. Andrew J. Saykin, Dr. Shannon L. Risacher, Dr. Kwangsik Nho, and many others, who have provided me very valuable domain expertise from neurological, biological and genetic perspectives, as well as many invaluable medical data sources. I have learned considerably from their multi-perspective insights into problems. I am very thankful to my collaborators: Prof. Casey S. Greene from University of Pennsylvania, Prof. Kim Sungeun from State University of New York at Oswego, Prof. Katy Börner, Dr. Mark H. Inlow and Michael Ginda from Indiana University, for many valuable discussions on algorithms and data visualization.

I also would like to thank my friends and roommates, for listening, offering me advice, and supporting me through this entire process. Special thanks to Shan Cong from Purdue University, Dr. Jingwen Yan from Indiana University, Dr. Hong Liang, Dr. Mo Tang and Yang Wang from Harbin Engineering University, Dr. Huan Zhou from University of Stuttgart, and my best friends Na Li, Yuli Pan and Yan Zhou from

Qingdao University. The debates, road trips, dinners, game nights, late-night gossip as well as general help and friendship were all greatly appreciated.

Finally I would like to express my special thanks and appreciation to my mom, dad, grandma and little brother. My parents have always provided unconditional love and care for my brother and myself, taught us to be honest, warm and strong, encouraged us to go for as much education as we could. My brother has been also my best friend and I love him dearly and thank for his love, support and patience. I love them so much, and I would not have made it this far without them. They are always supporting me and encouraging me with their best capabilities and wishes.

Xiaohui Yao

MINING HIGH-LEVEL BRAIN IMAGING GENETIC ASSOCIATIONS

Imaging genetics is an emerging research field in neurodegenerative diseases. It studies the influence of genetic variants on brain structure and function. Genome-wide association studies (GWAS) of brain imaging has identified a few independent risk loci for individual imaging quantitative trait (iQT), which however display only modest effect size and explain limited heritability. This thesis focuses on mining high-level imaging genetic associations and their applications on neurodegenerative diseases.

This thesis first presents a novel network-based GWAS framework for identifying functional modules, by employing a two-step strategy in a top-down manner. It first integrates tissue-specific network with GWAS of corresponding phenotype in regression models in addition to classification, to re-prioritize genome-wide associations. Then it detects densely connected and disease-relevant modules based on interactions among top reprioritizations. The discovered modules hold both phenotypical specificity and densely interaction. We applied it to an amygdala imaging genetics analysis in the study of Alzheimer's disease (AD). The proposed framework effectively detects densely interacted modules; and the reprioritizations achieve highest concordance with AD genes.

We then present an extension of the above framework, named GWAS top-neighbor-based (tnGWAS); and compare it with previous approaches. This tnGWAS extracts densely connected modules from top GWAS findings, based on the hypothesis that relevant modules consist of top GWAS findings and their close neighbors. It is applied to a hippocampus imaging genetics analysis in AD research, and yields the densest interactions among top candidate genes. Experimental results demonstrate that pre-

cise context does help explore collective effects of genes with functional interactions specific to the studied phenotype.

In the second part, a novel imaging genetic enrichment analysis (IGEA) paradigm is proposed for discovering complex associations among genetic modules and brain circuits. In addition to genetic modules, brain regions of interest also grouped to play role. We expand the scope of one-dimensional enrichment analysis into imaging genetics. This framework jointly considers meaningful gene sets (GS) and brain circuits (BC), and examines whether given GS-BC module is enriched in gene-iQT findings. We conduct the proof-of-concept study and demonstrate its performance by applying to a brain-wide imaging genetics study of AD.

Huanmei Wu, PhD, Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Imaging genetics is an emerging research field focusing on investigating influence of genetic variants on brain imaging phenotypes; and has identified a number of susceptible loci for neurodegenerative diseases. In this thesis, we present novel frameworks for mining high-level imaging genetic associations with their applications in brain disorders. In this chapter, we first briefly introduce imaging genetics with its research progresses and effects in brain degeneration disease, and discuss the advantages and limitations of present strategies and approaches employed in this research field, and then sketch the methods proposed in this thesis.

1.1 IMAGING GENETICS IN NEURODEGENERATIVE DISEASE

Recent advances in acquiring high-dimensional brain imaging and genome-wide data have provided new opportunities to assess the influence of genetic variations on neurodegenerative diseases, where the phenotype is quantitative brain imaging measurement instead of categorical disease status. Imaging genetics integrates brain imaging and molecular genetic data to improve the understanding of disease pathologies, from individual effect to complex interplay of genes and brain regions.

In the study of complex neurodegenerative diseases, evidences have shown that the pathological changes begin to develop years or even decades before the earliest clinical symptoms emerge [58]. This extended presymptomatic stage may provide essential information regarding to disease progression, as the earliest signs of disease may occur in brain and can be measured before noticeable symptoms developed. Us-

ing Alzheimer's disease as example, a number of biomarkers have been identified and measured for indicating and predicting the disease progression, including brain imaging, proteins (beta-amyloid and tau) in cerebrospinal fluid (CSF), proteins in blood, genetic risk profiling (e.g., *APP*, *APOE* allele $\epsilon 4$). Among the various biomarkers, brain imaging-including functional and structural imaging-has provided promising evidences for differentiating disease stagings from normal aging [79].

It is critical to understand the genetic architecture underlying complex neurodegenerative diseases. Over the past few decades, genetic analysis has played an increasing important role in human disease research [53]. In addition to using categorical disease status as phenotype, genetic association analysis of quantitative phenotypes has shown distinct advantages in statistical power and heritability explanation, especially the usage of iQTs due to its prominent performance on disease differentiation and prediction [79].

As a synthetic approach, imaging genetics has provided promising evidences for better understanding the underlying genetic and neurobiological mechanisms in brain disorders and their progression. In following, we review and discuss diverse strategies involved in imaging genetics from univariate association to high-level association, with their applications and significant findings in the study of AD as representative.

1.2 UNIVARIATE IMAGING GENETIC ASSOCIATION

Univariate analysis is widely employed in genetic association studies [79]. In imaging genetic association analysis, univariate strategy has been employed to evaluate the association of one or more independent genetic variants with single iQT.

GWAS, as a well-known implementation of univariate analysis, has been per-

formed in a few neurodegenerative studies to identify genetic markers such as single nucleotide polymorphisms (SNPs) that are susceptible to neuroimaging QTs [78]. For example, Potkin et al. [67] investigated the genome-wide genetic association on hippocampal volume in healthy control (HC) and AD patients, and identified 21 genes/chromosomal regions including *CAND1*, *EFNA5*, and *MAGI2*. Stein et al. [86] performed a GWAS on bilateral temporal lobe volume and identify 2 associations including rs10845840 from *GRIN2B* and rs2456930. Shen et al. [81] performed a brain-wide ROI level GWAS for investigating genome-wide associations with grey matter (GM) density, volume, and cortical thickness in HC and AD participants, and confirmed the associations of several known AD genes (e.g., *APOE*, *TOMM40*) with multiple brain regions. Stein et al. [85] proposed voxelwise GWAS (vGWAS), a massive GWAS to explore the relation between 448,293 SNPs with each of 31,622 brain voxels, and suggested several AD genes for further investigation including *CSMD2* and *CADPS2*.

In addition to the genome-wide analysis, targeted or candidate genetic association analysis of brain iQTs has also been widely investigated to increase statistical power and improve biological interpretation. For example, the associations of *APOE* with multiple magnetic resonance imaging (MRI) phenotypes have been largely examined due to it is the most risk genetic factor for AD [5,16,36,75]. Biffi et al. [8] investigated the association of AD candidate SNPs with AD related MRI measures, and confirmed the influence of four genes (*APOE*, *CLU*, *CR1* and *PICALM*) on multiple regions including hippocampal volume, amygdala volume and several others.

Most imaging genetic analyses focus on the univariate approach. This strategy can be simply implemented and is straightforward to interpret the identified single-

variant-single-iQT associations. Univariate strategy treats both genetic variants and brain iQTs as independent, which however is always not the case. Genetic variants often collectively influence susceptibility to a single or multiple iQTs, as well as one variant can affect multiple traits which is named as pleiotropy. These interlinked information may provide more significant insights into the underlying biological mechanisms of complex diseases. To address this issue and get more efficient use of the data, high-level imaging genetic association analysis has been proposed and is discussed in the following section.

1.3 HIGH-LEVEL IMAGING GENETIC ASSOCIATION

High-level association analysis has been demonstrated that can yield biologically meaningful findings by integrating prior knowledge (e.g., pathways) into a set of significant findings [72]. In the imaging domain, brain connectome studies have suggested that brain ROIs do not always have functions by each own, but functional or structural grouped to play role [10, 11, 21, 23, 44–46, 54, 80, 89–91, 93]. In the genetics domain, genes also do not perform functions individually, but always interact with others to make combined effect on complex diseases or traits. The set of functional interacted genes then forms genetic pathways or network modules.

Currently, most high-level association approaches focus on the genetics domain, where prior knowledge is from gene ontology (GO), functional annotation databases, genetic interaction networks and so on. Existing high-level imaging genetic association analysis can be classified into two categories: enrichment analysis and module identification analysis. Enrichment analysis assesses if genes from the same pathway or functional network module aggregate effects of multiple mutations to collectively con-

fer a significant susceptibility to complex diseases and traits, even when constituent genes/SNPs show only low or moderate effect sizes [70, 87]. Module identification strategy integrates GWAS data with prior knowledge (i.e., pathway, network) to construct and identify functional interacted modules that are related to specific diseases or phenotypes [3, 31, 33, 38, 96, 97].

High-level association analysis has demonstrated its efficiency for identifying and evaluating the interactive and cumulative effects of groups of genes and brain ROIs. However, most strategies either focus on only genetics domain, or ignore the context of human tissues which is essential for understanding the precise function of genes. Given the high-level imaging and genetics architectures, it is critical and challenging to understand their complex associations which may improve the understanding of underlying mechanisms of neurodegenerative diseases. Using prior knowledge from both imaging and genetics domains, functional annotation of brain circuits and genetic modules may be able to shed light on the fundamental pathology of neurodegenerative disorders.

1.4 CONTRIBUTIONS

Accordingly, the goal of this thesis is to develop and apply novel computational models for mining the high-level imaging genetic associations by integrating data-driven GWAS findings with multi-omics data and biological pathways and networks as prior knowledge. We apply proposed models in imaging genetics data from Alzheimer's Disease Neuroimaging Initiative (ADNI) as test beds to demonstrate their performances. We summarize the work in this dissertation as follows and present more details in the following chapters (Fig. 1.1).

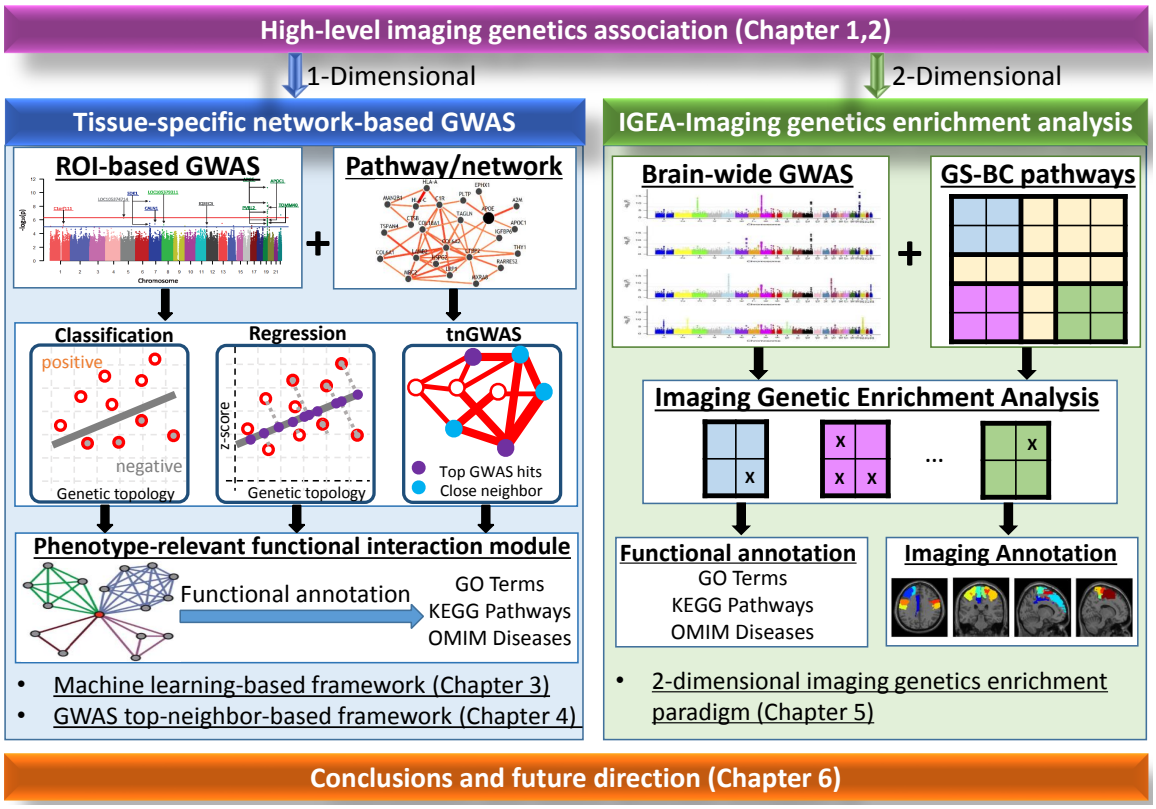

Figure 1.1: Overview and organization of the dissertation work

1.4.1 TISSUE-SPECIFIC NETWORK-BASED GWAS FRAMEWORK FOR MODULE IDENTIFICATION

In the first part of this thesis, two novel top-down network-based GWAS frameworks are proposed for constructing disease-relevant functional genetic modules that are specific to corresponding brain region. The precise functions of genes are highly related to their tissue context, and human diseases often arise from the disordered interplay of tissue-specific processes. However, current integrative analysis of GWAS always uses tissue-free interaction networks such as human protein-protein interaction (PPI) network without taking tissue specificity into account. Another limitation of existing approaches is that their efficiencies could be suboptimal when a large-scale network is present, because they employ a bottom-up strategy which needs to explore a large number of candidate modules for extracting significantly enriched ones.

To address above challenges, we develop two module identification frameworks: (i) a machine learning based approach that introduces regression models into tissue-specific network to re-prioritize GWAS results, and then construct modules from reprioritization results; and (ii) a GWAS top-neighbor-based (tnGWAS) module identification approach that extracts densely connected modules from top GWAS findings. The top-down strategy promises GWAS-enrichment and dense connection of candidate modules and increases the efficiency by limiting to explore small number of candidates. We applied these frameworks to the AD data to demonstrate their performances and help better understand mechanisms behind phenotype-specific genetic functional interactions.

1.4.2 IMAGING GENETICS ENRICHMENT ANALYSIS PARADIGM

Enrichment analysis has been widely applied in the genome-wide association findings, where genetic modules are examined for significant associations with a phenotype to help increase statistical power. Present enrichment analyses of neuroimaging phenotypes use biological pathways and networks as prior knowledge, typically ignore the interrelated structure between brain iQTs, and are insufficient to provide biological insight into the mechanisms of complex diseases that could involve multiple SNPs and multiple iQTs.

In the second part of this thesis, we expand the scope of one-dimensional genetic enrichment analysis into two-dimensional brain imaging genetic study. Given the high-dimensionality of both imaging and genetic data, we present an imaging genetic enrichment analysis (IGEA), a new enrichment analysis paradigm that jointly considers meaningful gene sets and brain circuits and examines whether any given GS-BC module is enriched in a list of gene-iQT findings. This work demonstrates its additional power for extracting biological insights on neurogenomic associations at a systems biological level.

1.5 ORGANIZATION

The rest of this dissertation is organized as follows (also see Fig. 1.1). Chapter 2 reviews existing functional annotation methods applied in system biology, including genetic module identification approaches and module enrichment analysis strategies, and discusses their applications and limitations. Chapter 3 and Chapter 4 present two novel tissue-specific network-based GWAS module identification frame-

works, the machine learning-based and the tnGWAS approaches, and evaluate their performances in Alzheimer research by integrating GWAS results of disease-relevant ROIs and prior functional interaction network knowledge. In Chapter 5, we propose a two-dimensional IGEA paradigm and conduct a proof-of-concept study. We apply the novel enrichment framework in the AD research, where the high-level imaging genetic associations are explored based on brain-wide genome-wide association study (BWGWAS) results. In Chapter 6, we summarize the work of this dissertation and discuss some future directions.

Chapter 2

RELATED WORK

Higher level genetic association analysis has been demonstrated that can yield biologically meaningful findings and help functional genetic annotation by integrating prior knowledge (e.g., pathways, networks) into a set of genetic findings. In this chapter, we review genetic functional annotation strategies from two directions: functional module identification and gene set enrichment analysis. In the first part, we start from giving a brief introduction about the concept of network-based genetic functional module identification in systems biology, and then discuss the existing methods along with their applications, advantages and limitations. We then introduce a most recently presented genome-wide interaction network resource, the tissue-specific functional network, along with a GWAS reprioritization application named NetWAS. In the second part of this section, we discuss two classes of one-dimensional gene set enrichment tests and their applications for functional annotation in complex diseases, categorized by whether using background information. In this dissertation, we write matrices and vectors as bold uppercase and lowercase letters respectively. Given a matrix $\mathbf{M} = [m_{ij}]$, we denote its i-th row as $\mathbf{m}^i$ and j-th column as $\mathbf{m}_j$. Given two column vectors $\mathbf{a}$ and $\mathbf{b}$, we use $\text{corr}(\mathbf{a}, \mathbf{b})$ to denote their correlation coefficient.

2.1 NETWORK-BASED FUNCTIONAL MODULE IDENTIFICATION

Network-based GWAS aims to identify functional modules from biological networks that are enriched by top GWAS findings. Although GWAS of brain imaging phenotypes have discovered and confirmed a number of factors susceptible for neurode-

generative disorders, a large fraction of phenotype variances still remain unexplained by these individual significant associations. The missing heritability results from a few aspects, among which genetic interactions of functional network modules and pathways account for a quite large proportion.

Accordingly, a number of efforts have been made to identify functional genetic modules, for better understanding underlying genetic architecture of complex diseases. Using ordinary genetic interaction networks as prior knowledge, recent integrative analysis of GWAS results have examined the cumulative effects of multiple variants, and shown promising performances on providing additional explanation for phenotype variances. These network-based GWAS approaches typically start from assigning GWAS statistics onto a user-specified genetic interaction network, then search for modules across the whole network to identify those can be enriched by GWAS top findings. As GWAS statistics are SNP-level p-values, they are firstly mapped to gene-level to facilitate the network node weights assigning. As such, the extracted modules would be relevant to corresponding GWAS phenotypes.

One example study is dense module GWAS (dmGWAS) [38] that applies dense module searching (DMS) strategy on human PPI to locally maximize the proportion of significant genes (i.e., genes with low p-values) in the GWAS results. In dmGWAS, the human PPI is downloaded from Protein Interaction Network Analysis platform (PINA) [98] which is constructed from six public PPI databases. SNP p-values from GWAS results are assigned as gene weights, which are then loaded into human PPI to obtain weighted PPI network. Then each gene would be selected as a seed, from which it expands to construct candidate modules by adding genes that could increase the proportion of low p-values. Therefore, the expanding process and module evaluation

step would be performed for times, same as the number of network nodes. A follow up work, EW_dmGWAS [97], boosts the power for identifying disease-relevant modules by incorporating gene differential expression information as edge weights into the dmGWAS framework. Experimental results of two approaches on breast cancer and schizophrenia demonstrate the promising of both methods for discovering disease-relevant genetic community, and illustrate the informative complement provided by gene expression data to network-based GWAS integration analysis.

Another representative algorithm is network interface miner for multigeneic interactions (NIMMI) [3], 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. Specifically, NIMMI uses VEGAS [56] to assign GWAS SNP-level p-values as gene-level p-values, and downloads human PPI from the Biological General Repository for Interaction Datasets (BioGRID) database (`http://www.thebiogrid.org/`). It then builds biological networks weighted by connectivity, which is estimated using a modification of the Google PageRank algorithm [25]. These weights are then combined with GWAS statistics to construct network modules. The performance of NIMMI is evaluated by being applied onto three GWAS datasets, where a few of phenotype-relevant network modules could be constructed.

The integrative protein-interaction-network-based pathway analysis (iPINBPA) method is also a network-based GWAS approach [96] and is an extension of the original PINBPA [6]. First, this approach uses VEGAS [56] to calculate gene-level p-values from GWAS SNP-level results. The PPI network employed in iPINBPA is from a manually curated human protein interaction network (available from: `http:`

//www.imsgenetics.org/). The iPINBPA method 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.

There are also several other network-based GWAS analysis approaches have been implemented for improving genetic functional annotation [76]. These methods are useful for identifying genetic modules relevant to complex diseases or phenotypes, however, have limitations. First, all these approaches employ a bottom-up strategy that examine a large number of candidate modules in order to identify the GWAS enriched ones, such that their efficiencies could become suboptimal when large-scale networks are present. Second, these approaches are using tissue-free human PPI networks as prior knowledge, without taking any tissue specificity into consideration. To overcome above limitations, we design two novel frameworks that take both efficiency and tissue-specificity into account, and present them in Chapter 3 and Chapter 4.

2.2 TISSUE-SPECIFIC FUNCTIONAL INTERACTION NETWORK AND APPLICATION

The precise functions of genes are highly related to their tissue context; and heritable diseases often result from tissue-specific pathology [28]. This is because disordered genetic expressions and functions caused by germline mutations occur in only certain tissues, although these variants present across all tissues [43]. As such, understanding the tissue-specific genetic underpinnings would promote the elucidation of molecular mechanisms underlying disease pathological processes.

Recently, tissue-specific genome-wide functional interaction networks have been constructed by integrating a large number of data sources and tissue-specific knowl-

edge, in order to identify the changing functions of genes across various tissues [28]. The weights of interactions of all gene-gene pairs are calculated for each tissue; and there are total 114 tissue-specific networks constructed, of which more than twenty are brain tissues or related to neurodegenerative diseases, like hippocampus, amygdala, frontal lobe and so on.

NetWAS, a novel statistical approach, has been developed by Greene et al. [28] that integrates tissue-specific genome-scale network as prior knowledge to guide the re-prioritization of GWAS statistics. Based on the hypothesis that disease risk genes would be enriched among the nominally significant ones, NetWAS employs topological information of tissue-specific genetic network to construct support vector machine (SVM) classifier to guide the re-rank of GWAS results. Specifically, The SVM-based method has been then applied to analyze hippocampus volume in AD and demonstrated that tissue-specific network could provide helpful context for improving the understanding of complex human diseases [84]. 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 got lost during the procedures.

With the above observation, in this thesis, we develop novel top-down network-based GWAS frameworks and achieve two goals at one time: (i) introduce regression models in addition to classification model in NetWAS for re-prioritizing GWAS results with network information; (ii) expand the re-prioritization to module identification for discovering tissue-specific genetic modules. We describe the details in Chapter 3 and Chapter 4.

2.3 GENE SET ENRICHMENT ANALYSIS

Given a set of candidate genes, enrichment analysis examines if they are functional associated together with disease phenotypes, or share common biological functions, pathways, regulations and so on [57, 87]. Enrichment analysis has been widely applied in the GWAS and its expanded module identification analysis, where gene sets corresponding to biological pathways are examined for significant associations with a phenotype or complex disease. This high-level association analysis has been demonstrated that can increase statistical power and improve biological interpretation by integrating prior knowledge (e.g., pathways) into discovered gene set. Prior knowledge employed in enrichment analysis could be from existing functional annotation databases like gene ontology (GO) [1], KEGG pathway database [40] and so on.

A number of enrichment analysis methods have been proposed to functionally annotate the identified gene sets; and can be classified into two types based on different hypotheses: over-representation analysis and rank-based analysis. Below we briefly introduce both strategies with their applications in complex disease studies, and then discuss their limitations for implementing in imaging genetics study.

2.3.1 OVER-REPRESENTATION ENRICHMENT ANALYSIS

Over-representation test is to evaluate if a known class of functional gene set (e.g., genetic pathway) is over-represented in a set of candidate genes (e.g., GWAS significant findings). In this strategy, a threshold is needed to define the list of candidate genes. This strategy can be formulated as an independence test problem; and a few of statistical distributions have been applied to implement it, including hypergeometric test (Fisher's exact test), binomial test, χ^2 test and others [19, 26].

Here we formulate the over-representation enrichment analysis using a hypergeometric test (Fisher's exact test) for illustration. Assume there are a total of N genes included in the analysis, of which $n = |L|$ genes in the set L are significant ones, $m = |T|$ genes are from a given pathway T, and k out of n significant genes are from the pathway T. According to hypergeometric distribution, the over-representation p-value of having k or more genes from T in L can be calculated from the sum of the probabilities of a random set of n genes having $k, k+1, \ldots, n$ genes from T:

$$p\text{-value}_{enrich} = Pr(|L \cap T| \geq k) = \sum\nolimits_{i \geq k} \frac{\binom{m}{i} \times \binom{N-m}{n-i}}{\binom{N}{n}}. \tag{2.1}$$

Here, we use $Pr(\cdot)$ to denote the probability function.

The enrichment p-value estimated from hypergeometric test depends on the total number of involved genes, that is, the value of N. It is hard to calculate hypergeometric distribution when N is large. But this problem can be solved through approximated by a binomial distribution. Imagine that when N is large, sampling a gene set of size n without replacement has no discernible effect on the total N genes. As such the probability that a randomly selected gene will be significant is essentially constant and has the value n/N. Accordingly, the hypergeometric distribution can be approximated by a binomial distribution when N is large.

2.3.2 RANK-BASED ENRICHMENT ANALYSIS

To overcome the limitation of over-representation approach which requires threshold to define significant genes, rank-based strategy has been proposed to take all genes into account. A successful example tool is gene set enrichment analysis (GSEA) [87], which was developed for gene expression analysis and was then extended to GWAS.

The GSEA evaluates the association of studied disease or phenotypes, by examining whether genes from a pathway tend to distributed in the top (or bottom) of the ranked GWAS results. We briefly describe the implementation of GSEA in below.

Given the ranked list of genes L with association statistics (e.g., gene-level p-values from GWAS), the enrichment score (ES) of a known pathway S is calculated using a Kolmogorov-Smirnov (KS) approach with weight 1. That is, by walking down the list L, a running-sum statistic is increased when encountering a gene in S, and is decreased when encountering a gene not in S. The ES is then provided by the maximum deviation from zero of the running sum. Permutation is performed to evaluate the statistical significance of ES.

The rank-based analysis has been applied in the AD-related GWAS and successfully confirmed a few GWAS findings with a few AD-relevant pathways [71]. Besides of the advantage we mentioned above that without requiring a user-defined threshold, the phenotype-based permutation of rank-based approach can keep the correlation structure among genes, and thus provides a more reasonable assessment of significance than permuting genes. However, there are also several limitations for rank-based strategy. First, both rank-based and over-representation tests consider pathways independently, which however often overlap with one another. Because of this, a pathway may be significantly enriched due to the common genes it shares with a real enriched pathway. Second, rank-based methods take into account the ranks of genes but ignore the strength of associations. Some modifications have been proposed to improve this problem by adding weights to ranked genes based on their association strengths [57]. Third, the computational efficiency of rank-based analysis would decrease dramatically when the number of permutation largely increases.

Both types of enrichment analysis methods are applicable only to one-dimensional data, that is, genetic findings associated with each single QT. In imaging genetics, the ultimate goal is to discover high-level associations between meaningful GSs and BCs, which typically include multiple genes and multiple iQTs. It remains a major challenge to understand and interpret a set of significant genes and iQTs without any unifying biological theme. In this work, we develop a novel enrichment paradigm for mining two-dimensional imaging genetic associations and revealing complex relationships among them, by integrating multi-omics data including whole brain genomics, transcriptomics, and neuroanatomics data.

Chapter 3

TISSUE-SPECIFIC NETWORK-BASED GWAS FOR IDENTIFYING FUNCTIONAL INTERACTION MODULES: A MACHINE LEARNING BASED FRAMEWORK

Network-based GWAS methods have been implemented for identifying 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. Tissue-specific genome-wide functional interaction network has been constructed for reflecting the changing functional roles of genes across tissues. In this chapter, we present a novel framework, that integrates tissue-specific network with corresponding GWAS data to construct phenotype- or disease-relevant genetic modules, to help improve the understanding of genetic architecture of complex diseases.

3.1 BACKGROUND

GWAS has been performed to identify genetic markers such as SNPs that are associated with common human diseases. In brain imaging genetics, an emerging field that studies how genetic variation influences brain structure and function, GWAS also has discovered genes susceptible to brain iQTs [47, 78, 79]. Each identified iQT 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 effect 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 [3, 31, 33, 38, 97].

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) [38], 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 [3], 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 [96] and is an extension of the original PINBPA [6]. 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 [28]. Recently, tissue-specific genome-wide functional interaction networks have been constructed in order to identify the changing functional roles of genes across tissues [28]. 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 hippocampal volume in AD and demonstrated that tissue-specific networks could provide helpful context for understanding complex human diseases [84]. *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: (1) introduce regression models in addition to classification models for re-prioritizing GWAS results with network information; (2) 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*: (1) Our framework expands the NetWAS scope from re-prioritizing GWAS findings to module identification. (2) Our framework introduces regression models into NetWAS to embrace the complete coverage of the continuous p-value spectrum. (3) 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.

3.2 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

Table 3.1: 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±sd)	74.02±5.72	71.86±5.61	71.16±7.29	72.31±7.63	73.32±7.34
Education(mean±sd)	16.44±2.66	16.85±2.63	16.06±2.66	16.24±2.81	16.19±2.72

in the study of 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 [22, 63, 68] and has been used to assist the clinical diagnosis of AD [88]. Studies on fluorodeoxyglucose [^{18}F]FDG-PET have demonstrated different usage patterns of glucose metabolism in amygdala between AD and healthy control subjects [39].

3.2.1 IMAGING DATA, GENOTYPING DATA AND GWAS

The imaging and genotyping data used for GWAS were obtained from the ADNI 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 AD. For up-to-date information, see www.adni-info.org.

Preprocessed [^{18}F]FDG-PET scans were downloaded from the LONI website (see `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 [42]. 989 non-Hispanic Caucasian participants (Table 3.1) with complete baseline FDG amygdala measurements were studied.

Associations between amygdala measures and SNPs (allelic dosage) were examined by performing GWAS using PLINK [69], 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 $\pm$20K bp of the gene [60]. In addition, 10 GWAS permutations were performed to illustrate that only the original GWAS data yielded promising findings.

3.2.2 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 data sets covering thousands of experiments contained in more than 14,000 distinct publications [28]. 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-specific 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 was extracted after mapping to GWAS results. The weights range from 0 to 1, where larger measures represent stronger interactions.

3.2.3 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 [47], 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/`).

The following is a detailed list of genes we included: *A2M, ABCA7, ACE, AD10, AD5, AD6, AD8, ADAM10, APBB2, APOE, APP, ARID5B, BIN1, BLMH, CALHM1, CASS4, CCR2, CD2AP, CD33, CELF1, CH25H, CHRNB2, CLU, CR1, CST3, DAPK1, DSG2, ECE1, ENTPD7, EPHA1, FERMT2, GAB2, GAPDHS, GRN, HFE, HLA-DRB1, HLA-DRB5, IDE, IL1A, IL1B, IL33, INPP5D, LDLR, MEF2C, MPO, MS4A6A, MTHFR, NEDD9, NME8, NOS3, PACIP1, PGBD1, PICALM, PLAU, PRNP, PTK2B, RIN3, SLC24A4, SORCS1, SORL1, TF, TFAM, THRA, TNF, TNK1* and *ZCWPW1*.

3.2.4 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 Fig. 3.1 for the workflow.

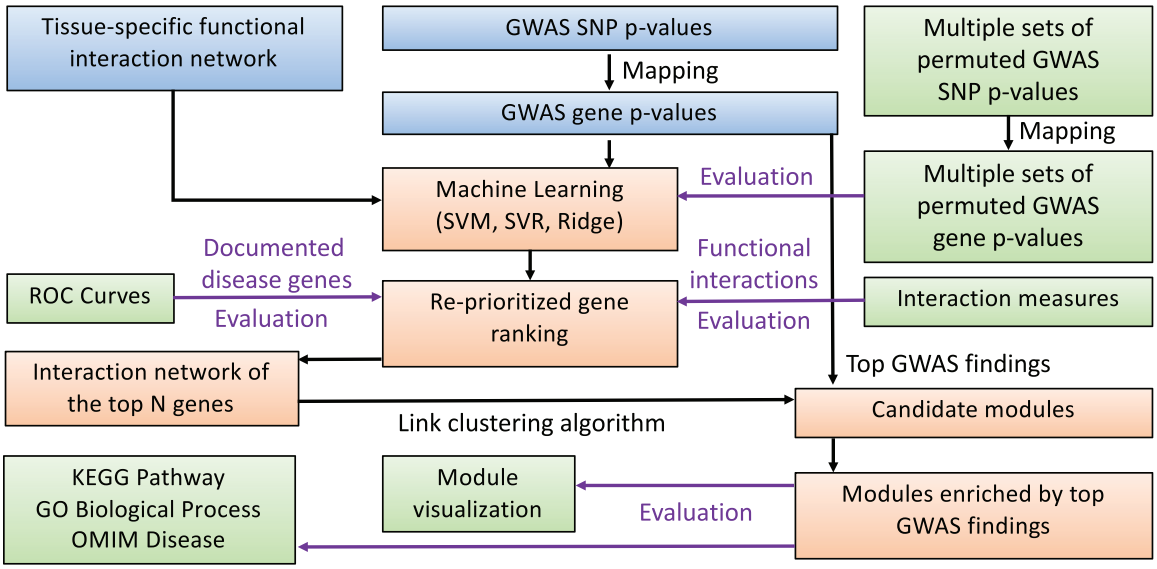

Figure 3.1: The workflow for identifying functional interaction modules from the tissue-specific network using GWAS findings.

NETWAS RE-PRIORITIZATION OF GWAS RESULTS

Following [84], 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, SVR and 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 $\mathbb{C}^{(t)}$ for model training. That is, gene-level p-values of $\mathbb{C}^{(t)}$ were used as responses (positive/negative labels for SVM), while interactions between genes from $\mathbb{C}^{(t)}$ 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 following for detailed implementation.

Now we describe the three machine learning algorithms used in this chapter for NetWAS re-prioritization of the GWAS findings: SVM, SVR and Ridge. We denote vectors as boldface lowercase letters and matrices as boldface uppercase ones. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}} \in (0, 1]^{n \times m}$ be the predictor matrix, $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ be the response vector (y_i is categorical for classification and continuous for regression), where n is the number of data samples and m is the number of features.

Given input predictors $\mathbf{X}$ and responses $\mathbf{y}$, the **Support Vector Machine (SVM)** [12] seeks to find a hyperplane with maximum margin between classes for accurate classification:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^{\mathrm{T}} \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \end{aligned} \tag{3.1}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector, C is a hyperparameter that balances the weights between the regularization term and classification error, and $\mathbf{w}^{\mathrm{T}} \mathbf{x}_i + b = 0$

is the separating plane. The dual formulation of problem (3.1) is

$$\begin{aligned} \underset{\alpha_1,\ldots,\alpha_n}{\text{maximize}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j \\ \text{subject to} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0, \text{ and } 0 \leq \alpha_i \leq C \ \forall i. \end{aligned} \tag{3.2}$$

Those with $\alpha_i > 0$ are support vectors.

Support Vector Regression (SVR) [82] uses the same principle as the SVM. The SVR tries to find a linear function of which the predicted value is deviated from the actual value by at most $\varepsilon > 0$ for all the training data (up to additional errors for outliers) and at the same time minimizes the Euclidian norm of the regression coefficients:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ \text{subject to} \quad & y_i - \mathbf{w}^{\mathrm{T}} \mathbf{x}_i - b \leq \varepsilon + \xi_i \\ & -y_i + \mathbf{w}^{\mathrm{T}} \mathbf{x}_i + b \leq \varepsilon + \xi_i^* \\ & \xi_i \geq 0, \ \xi_i^* \geq 0. \end{aligned} \tag{3.3}$$

where ε is a hyperparameter that controls the precision of prediction, and ξ_i, ξ_i^* are the slack variables that are introduced to relax the inequality constraints for outliers. The dual formulation of problem (3.3) is

$$\begin{aligned} \underset{\alpha_1,\ldots,\alpha_n}{\text{maximize}} \quad & -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \alpha_i^*) \left(\alpha_j - \alpha_j^*\right) \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j \\ & - \varepsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*) \\ \text{subject to} \quad & \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0, \text{ and } 0 \leq \alpha_i, \alpha_i^* \leq C \ \forall i. \end{aligned} \tag{3.4}$$

The SVM and SVR algorithms are implemented in R in the 'kernlab' package [4], which is used in this work.

Ridge Regression, also known as L2-regularized linear least squares regression, is designed for minimizing a weighted average of the sum of squared residuals and sum of squared regression coefficients:

$$\text{minimize} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - b - \mathbf{x}_i^{\mathrm{T}}\mathbf{w})^2 + \frac{\lambda}{2}\left\|\mathbf{w}\right\|^2. \tag{3.5}$$

In the 'glmnet' package, which is used in this work, the coordinate descent algorithm [24] is used to solve problem (3.5).

As stated in Section 3.2.2, the amygdala-specific functional network can be formulated as a symmetric matrix of interactions among all n genes, that is, $\mathbf{S} \in (0,1]^{n\times n}$. GWAS of amygdala imaging phenotype yielded a list of p-values $\mathbf{p} = (p_1, \ldots, p_n)^{\mathrm{T}} \in (0,1)^n$. In the experiments, we employed negative log transformation of p-values instead of original p-values as responses in regression models. Thus, $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}_+^n$ were used as regression responses where $y_i = -\log_{10}(p_i)$.

To train the above three models, we firstly selected a set of nominally significant genes with p-value < 0.01, denoted as $\mathbb{A} = \{a_1, \ldots, a_{|\mathbb{A}|}\}$, and then randomly partitioned the remaining genes (i.e., p-value ≥ 0.01) into five equal groups without overlap $\mathbb{B}^{(t)} = \{b_{t1}, \ldots, b_{t|\mathbb{B}^{(t)}|}\}$, $t = 1, \ldots, 5$, with $|\mathbb{B}_1| = \cdots = |\mathbb{B}_5|$. We combined nominally significant gene set $\mathbb{A}$ with each non-significant gene set $\mathbb{B}^{(t)}$ to construct set $\mathbb{C}^{(t)} = \mathbb{A} \cup \mathbb{B}^{(t)}$ for model training. Of note, this five-fold strategy was employed in our previous work [84] to balance the positive and negative samples in the training data. Gene-level association statistics of set $\mathbb{C}^{(t)}$ were applied as responses, while

functional interactions between all genes and genes in $\mathbb{C}^{(t)}$ were extracted as predictors. Thus, $\mathbf{y}^{(t)} = (y_{a_1}, \ldots, y_{a_{|\mathbb{A}|}}, y_{b_{t1}}, \ldots, y_{b_{t|\mathbb{B}^{(t)}|}})^{\mathrm{T}}$ (positive/negative labels for SVM) were responses, and functional interactions between all genes and genes in $\mathbb{C}^{(t)}$ were extracted as predictors $\mathbf{X}^{(t)} = (\mathbf{x}_{a_1}, \ldots, \mathbf{x}_{a_{|\mathbb{A}|}}, \mathbf{x}_{b_{t1}}, \ldots, \mathbf{x}_{b_{t|\mathbb{B}^{(t)}|}})^{\mathrm{T}} \in (0,1]^{(|\mathbb{A}|+|\mathbb{B}^{(t)}|)\times n}$. For each method $\in$ {SVM, SVR, Ridge}, five models $M^{(t)}$, $t = 1 \ldots 5$, were trained and then used to do predictions for all the genes. In other words, model $M^{(t)}$ was constructed based on training data $\mathbf{X}^{(t)}$ and $\mathbf{y}^{(t)}$ and then applied to the full interaction matrix $\mathbf{S}$ to obtain predictions for all the n genes, from which $\hat{\mathbf{y}}^{(t)}$ - the list of predictions (SVR and Ridge) or distances from hyperplane (SVM) - were obtained. Finally, genes were re-prioritized based on the mean predictions (SVR and Ridge) or distances from hyperplane (SVM) across five sets of results, namely, $\hat{\mathbf{y}} = \frac{1}{5}\sum_{t=1}^{5} \hat{\mathbf{y}}^{(t)}$.

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 [84], ROC curves and AUCs were calculated using 66 documented AD candidates as gold standard positives to illustrate the concordance of gene-level results from these methods with the known AD risk genes. Specifically, the aforementioned 66 AD genes were defined as positives while all the other genes are defined as negatives for calculating AUC to see the distribution of AD risk genes in our re-prioritization results. Genes were re-ordered according to their re-prioritization values, from highest to lowest related to the studied phenotype. After labeling each gene in the re-prioritization list as positive or negative according to whether it is in the 66 AD genes, we calculated the true positive rate (TPR) and

false positive rate (FPR) by selecting each gene as a cutting point from the highest to the lowest. The ROC curve could then be created by integrating the TPR/FPR values of all the genes, and thus the AUC value could be calculated based on the ROC curve. We hypothesize that the NetWAS re-prioritization findings match the known AD genes better than the original GWAS findings, which should yield a larger AUC value. In addition, mean statistics of functional interaction measures among top genes were used to evaluate the degree of functional interactions among these re-ranked top genes.

In this work, we hypothesize that integration of functional interaction network can better identify disease- or phenotype-relevant genes. To evaluate this hypothesis, the documented AD genes are used as "ground truth" to check whether our re-prioritization results are better than the original GWAS. Of note, this evaluation step is not a part of our module identification framework. Without using the "ground truth" information, we can still identify phenotype-relevant modules using the proposed method. In this case, replication in independent cohorts is a necessary future step to confirm the identified network modules.

IDENTIFICATION OF GWAS-ENRICHED MODULES

The goal of the NetWAS re-prioritization is twofold: (1) 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; (2) 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 [2] 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 following for details.

Given a set of candidate modules which were extracted from top NetWAS re-prioritizations, enrichment analysis was performed to test the phenotype-relevance of these modules using top GWAS findings. We applied the hypergeometric test to assess whether a candidate module is significantly enriched by the top GWAS findings.

Using left amygdala as an example, we obtained all n GWAS findings from imaging genetic association analysis, and selected a set of n_t genes with the smallest p-values (denoted as $\mathbb{T}$). We also had a set of m genes (denoted as $\mathbb{D}$) from a given candidate module, of which k genes were from top GWAS findings $\mathbb{T}$. Using hypergeometric test for independence as stated in Eq. (2.1), the enrichment p-value for the given candidate module was calculated as:

$$p\text{-value}_{enrich} = Pr(|\mathbb{T} \cap \mathbb{D}| \geq k) = \sum\nolimits_{i \geq k} \frac{\binom{m}{i} \times \binom{n-m}{n_t-i}}{\binom{n}{n_t}}. \tag{3.6}$$

Enrichment p-values from Eq. (3.6) were then corrected for the number of candidate modules using the Bonferroni method.

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.

COMPARISON OF GENE-BASED ASSOCIATION APPROACHES

We employed the 2nd smallest SNP p-value as gene-level p-value to facilitate the GWAS re-prioritization. This is an efficient approach stated in [60] to summarize the information in multiple SNPs, to evade spurious associations of using the 1st smallest SNP p-value which could be a random association generated by chance.

The reason why we applied the 2nd smallest SNP p-value strategy instead of using VEGAS (the one employed in [84] and [28]) is that running VEGAS on $\sim$5 million SNPs is very time consuming and has huge memory requirement. To facilitate performance evaluation for the 2nd smallest SNP p-value approach, we created a new genotype data set consisting of 565,374 SNPs, which was imputed to the Illumina OmniExpress platform (most of ADNI-2 data collected using this platform). 989 non-Hispanic Caucasian participants (the same as those studied in the above analysis) with FDG-PET amygdala measurements were included. Using this data set, we performed additional analyses to compare the performances of four different gene-level p-value methods: 2nd smallest p-value, 1st smallest p-value, VEGAS and GATES [49].

We firstly performed GWAS to examine associations among 565,374 SNPs and FDG imaging measures in the left amygdala. We then applied four methods to map or combine SNP-level p-values to obtain gene-level p-values. To compare their performances, we performed the ROC analysis and used the AUC to illustrate the concordance of gene-level results from these methods with AD risk genes. Moreover, we employed these four lists of gene-level p-values as responses to Ridge regression-based NetWAS, and used AUC to assess their re-prioritization performances.

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 database 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.

3.3 EXPERIMENTAL 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.

3.3.1 GWAS OF AMYGDALA IQTS

GWAS were performed to examine genetic associations between 5,574,300 SNPs and FDG-PET measures in the left and right amygdalas. Using $p \leq$ 5E-8 as the threshold, nine SNPs were identified to be significantly associated with the average FDG-PET measure in the left amygdala (see Fig. 3.2 for the Manhattan plot), including two within the *APOE* gene (rs429358 with p = 1.99E-11, rs769449 with p = 3.28E-09), one within the *SDK1* gene (rs148359108 with p = 2.02E-09), one between the *APOE* and *APOC1* gene (rs10414043 with p = 8.56E-09), and five within the *APOC1* gene (rs7256200 with p = 8.56E-09, rs12721051 with p =1.11E-08, rs56131196 with p = 1.11E-08, rs4420638 with p = 1.11E-08 and rs73052335 with p = 3.50E-08). No significant findings were identified on the right side.

After mapping the 2nd smallest SNP-level p-values to genes [60] 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-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.

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

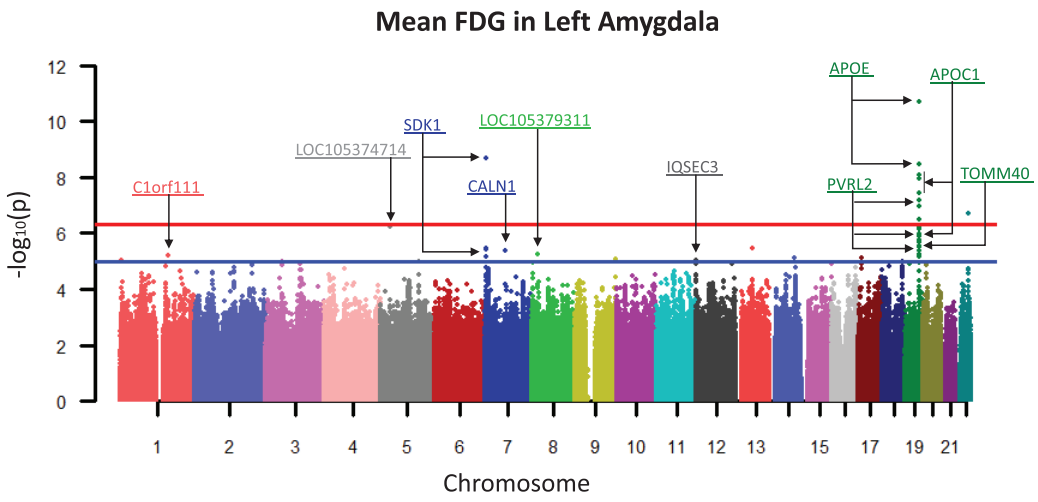

Figure 3.2: Manhattan plot of the FDG-PET imaging measure in the left amygdala. The x-axis corresponds to the genomic coordinates, and y-axis corresponds to negative logarithm of the association p-value for each SNP. Each dot on the Manhattan plot signifies a SNP.

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 $p = 2.02\text{E-}09$), and a nearly significant one at the gene level ($p = 3.35\text{E-}06$). *SDK1* was shown to specifically phosphorylate 14-3-3ζ at serine 58 [29], where the latter played an important role in amygdala cell death [37]. *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.

3.3.2 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 to-

tal 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 $20,168 \times 20,168$ and two lists of 20,168 gene-level p-values for left and right amygdala iQTs 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 can contribute useful information and yield 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-PET measures as promising AD biomarkers). We compared the re-prioritization performances of three machine learning models and GWAS using both original and permuted data.

Fig. 3.3(A,B) show 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

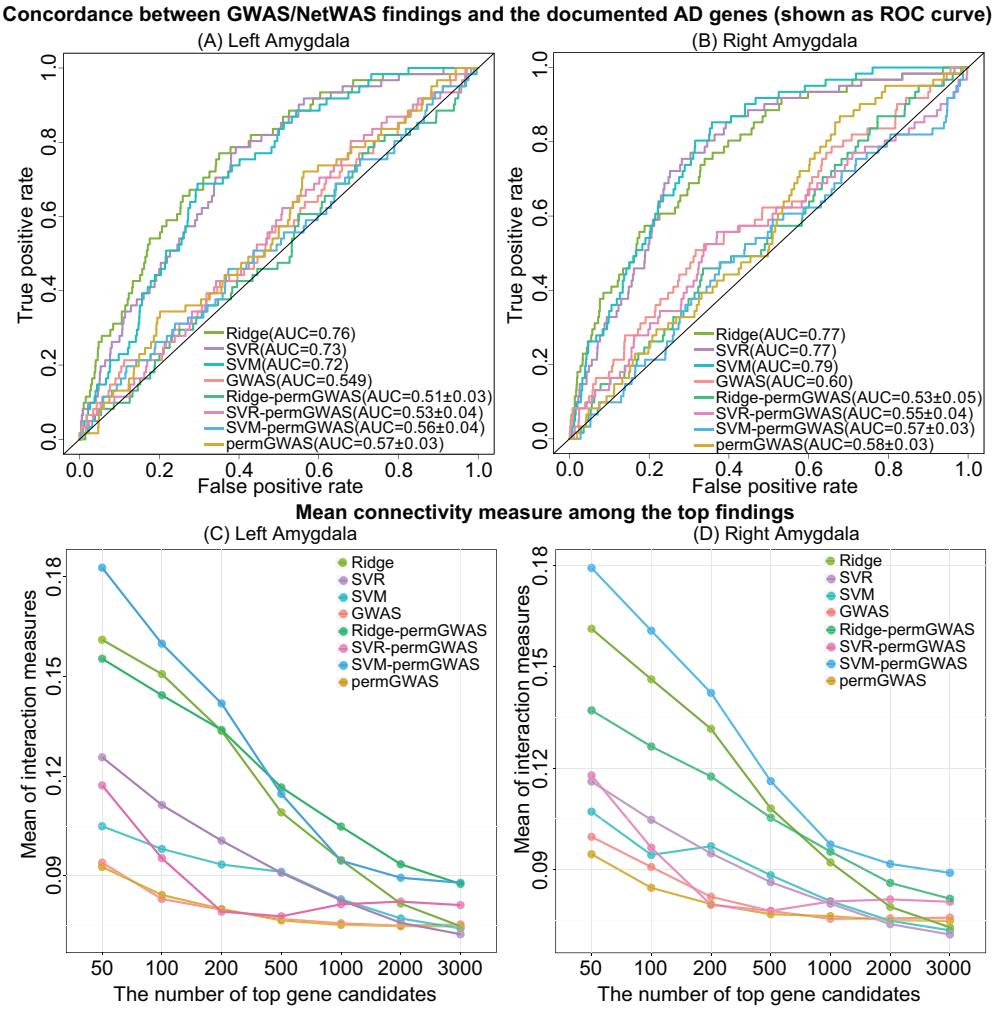

Figure 3.3: Performance evaluation of re-prioritization results. (A-B): ROC curves with AUC results on left and right amygdalas, respectively, to measure the concordance between the GWAS/NetWAS findings and the documented AD genes. For each analysis on permuted GWAS, the mean and standard deviation of 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.

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 real 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.

Fig. 3.3(C,D) show the mean functional interaction of the top findings. We used a series of thresholds from top 50 to top 3000 (of note, $\sim$ 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.

3.3.3 COMPARISON OF GENE-BASED ASSOCIATION APPROACHES

Fig. 3.4 shows the ROC curves and AUCs of four gene-based association methods on both original GWAS results and Ridge-based NetWAS re-prioritizations given its outstanding performance. The 1st and 2nd smallest SNP p-value approaches outperform the VEGAS and GATES on either GWAS results or NetWAS re-prioritizations; in particular the 2nd smallest SNP p-value approach obtains the best concordance with

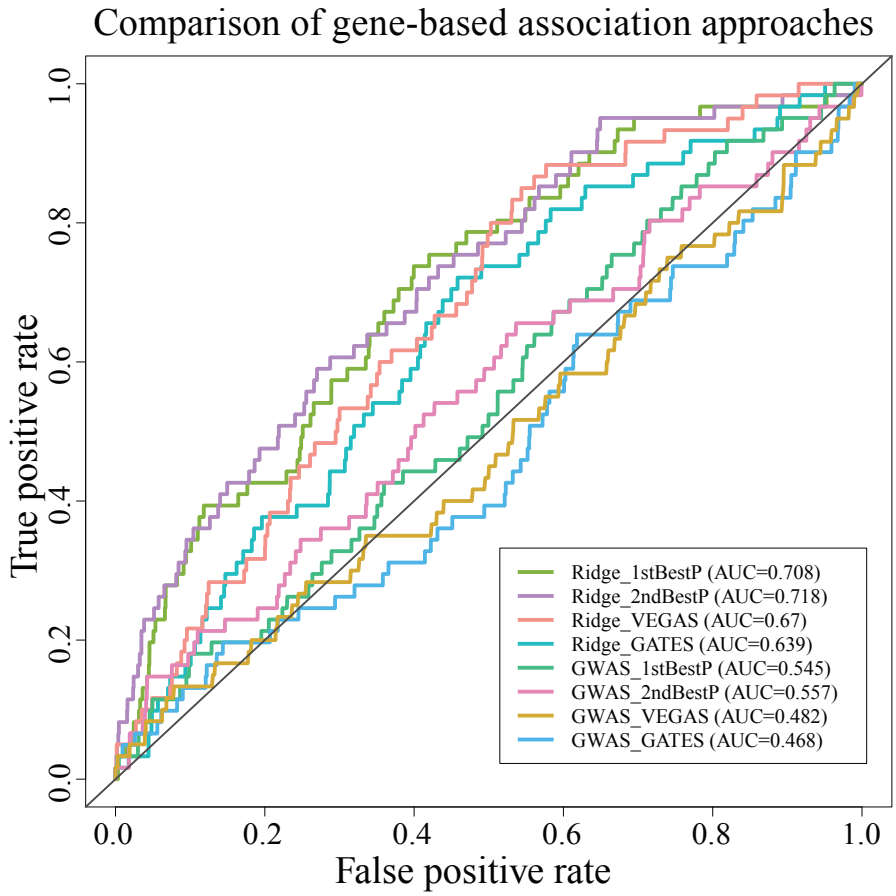

Figure 3.4: Comparison of four gene-based association approaches including 1st smallest p, 2nd smallest p, VEGAS and GATES. ROC curves with AUC results of four gene-level p-value approaches on left amygdala, to measure the concordance between the GWAS/NetWAS findings and the documented AD genes.

66 AD genes. These results might be partially attributed to that most of the documented AD genes were discovered from the GWAS analyses based on the SNP-level significance instead of the gene-level significance. In addition, Ridge-based NetWAS demonstrates much higher AUCs than the GWAS findings on all four gene-based association approaches, showing additional evidence for the power of integration of functional interaction network with GWAS.

3.3.4 AMYGDALA-RELEVANT TOP PREDICTIONS

We investigated top 50 re-prioritized genes obtained from three machine learning models, and compared their functional interactions in detail. Fig. 3.5(A,B) show 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).

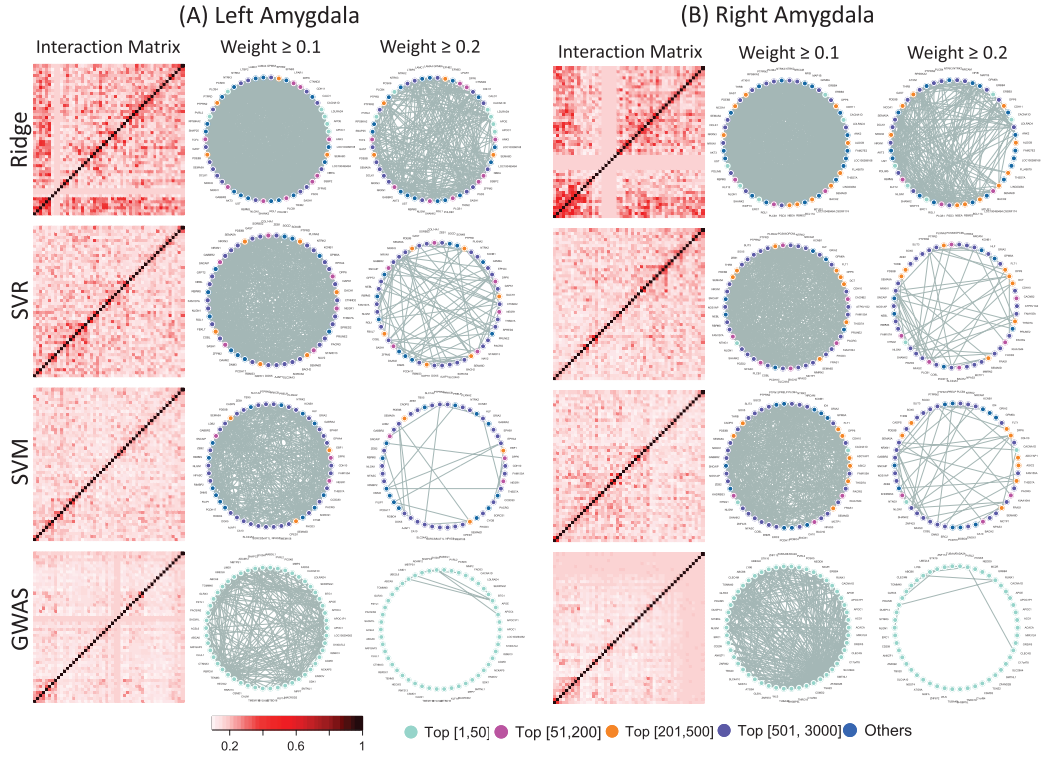

Figure 3.5: Comparison of top 50 findings by three NetWAS re-prioritization methods (Ridge, SVR and SVM) and the original GWAS. (A) and (B) represent results on left and right amygdalas, respectively. Heamaps show the complete interaction matrix of top predictions. Circular networks show interactions between genes after filtering weak connections. Nodes in circular network are colored by their ranking in the original GWAS.

Table 3.2: Modules identified by Ridge-based NetWAS.

Ridge	Module ID	# of genes	GWAS Enrichment p-value (corrected)
Left Amygdala	Module 01	18	4.58E-05
	Module 02	18	3.61E-03
	Module 03	47	2.21E-05
	Module 04	12	1.57E-03
Right Amygdala	Module 05	50	2.49E-09

This, combined with statistics summary from Fig. 3.3, indicates the outstanding performance of Ridge.

3.3.5 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. Table 3.2 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

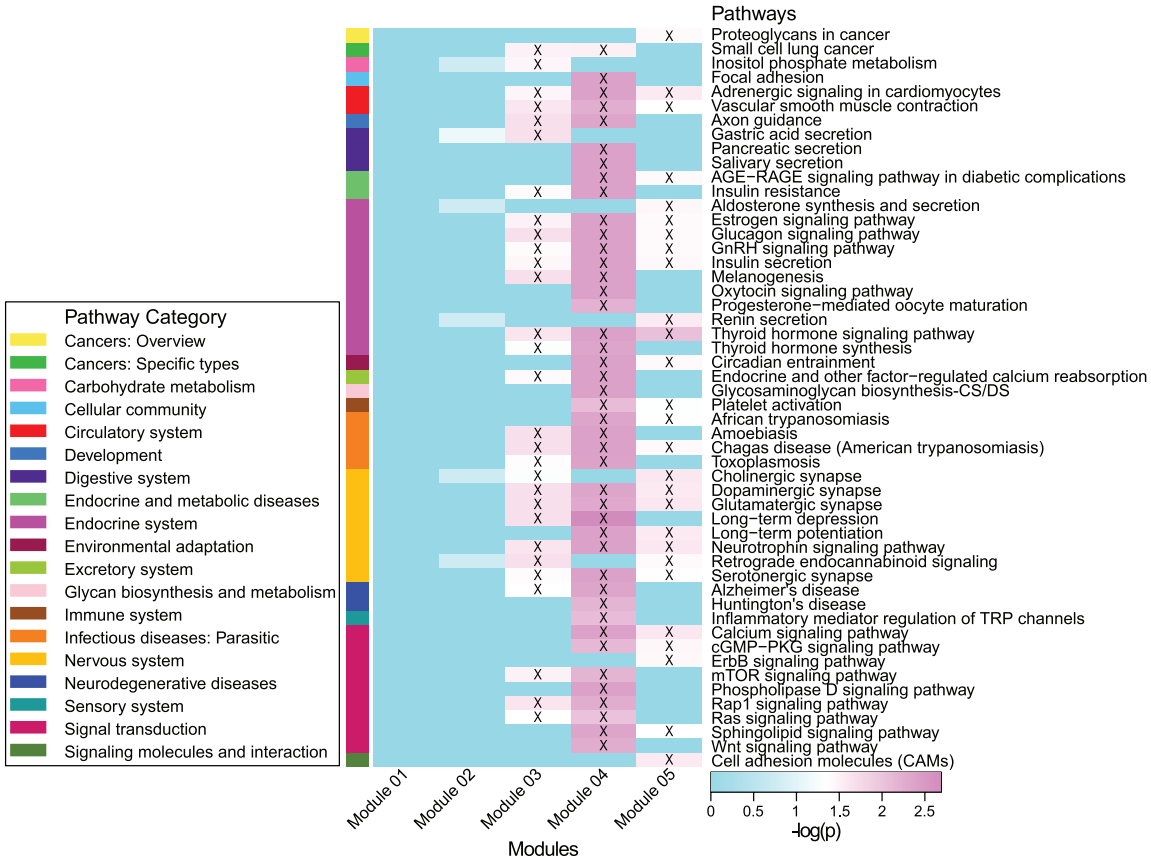

Figure 3.6: KEGG pathway enrichment of the identified modules. The x-axis corresponds to the module ID, and y-axis corresponds to the KEGG pathway. Each cell shows -log(p) of enrichment significance of a KEGG pathway by a module. A marked cell represents a significant enrichment (corrected p-value $\leq$0.05).

larger number of GWAS top findings into enrichment analysis to relax the phenotype-relevance.

3.3.6 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: (1) functional pathways, (2) biological processes, and (3) diseases, based on KEGG pathway, GO-BP terms and OMIM disease databases, respectively.

Fig. 3.6 shows the KEGG pathway enrichment results mapped to 19 categories.

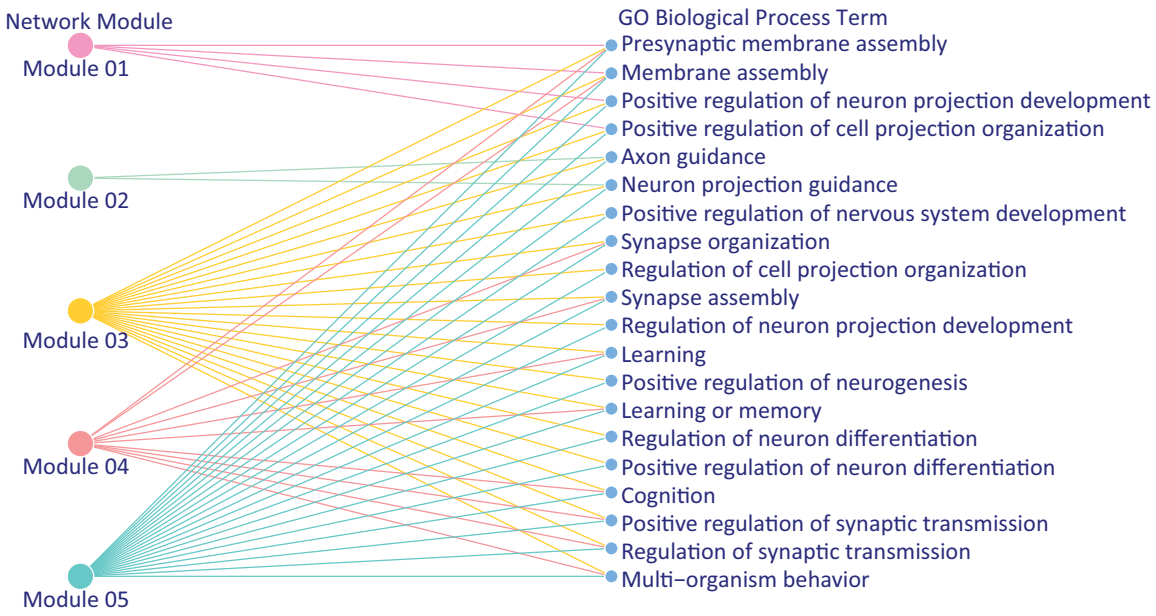

Figure 3.7: Gene Ontology Biological Process enrichment of the identified modules. Left column shows module IDs, and right column shows top enriched GO-BP terms. Links between modules and GO-BP terms represent significant enrichment findings (corrected p-value <0.05).

From the results, two modules from left amygdala and one module from right amygdala have a number of significant functional enrichments, while 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 [9, 55, 65]. Signal transduction like calcium signaling pathway (Modules 04 and 05) playing key role in short-

Table 3.3: OMIM diseases enriched by the identified modules.

Module ID	OMIM Disease	p value (corrected)
Module 01	Myocardial infarction	1.5E-02
	Macular degeneration	1.5E-02
	Alzheimer's disease	1.5E-02
Module 02	Prostate cancer	1.8E-02
Module 04	Autism	4.5E-02

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 [7]. Neuroinflammation emerged as an important component of AD pathology recently, and immune system indicated a crucial role in the progression of AD [30]. 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 [27].

Fig. 3.7 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 GO-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 Table 3.3, where three modules (Modules 01, 02, and 04) are significantly enriched by various types of dis-

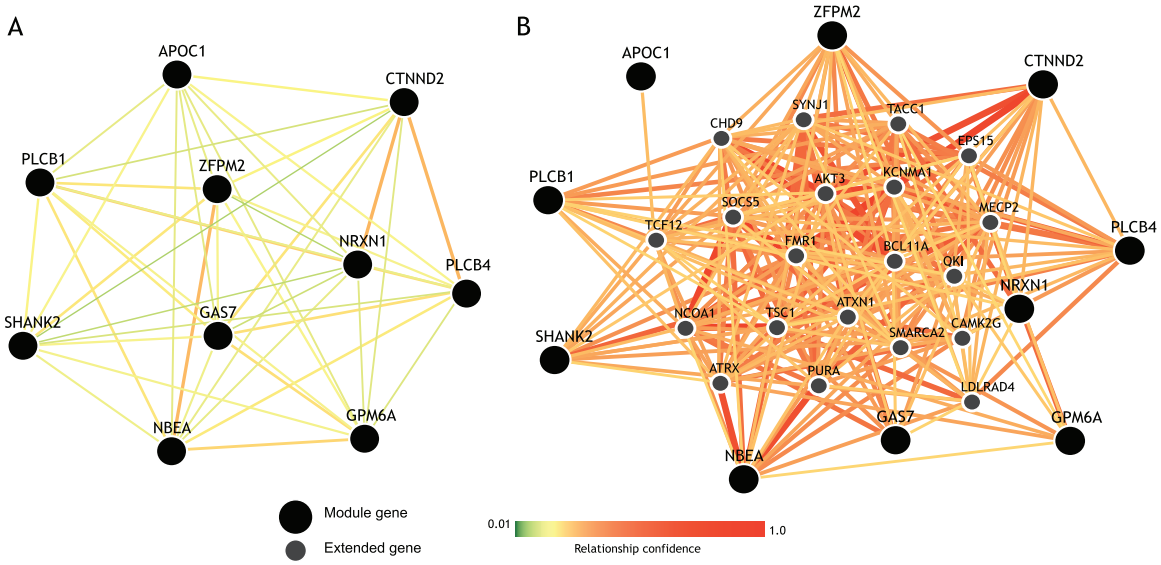

Figure 3.8: Visualization of Module 04 and its extension. (A) shows the interaction network of genes in Module 04, where color of links represents the relationship confidence from GIANT. Two genes from Module 04 are excluded as they cannot be matched to GIANT database. (B) shows the extended network using genes in Module 04 as seeds, with large nodes indicating genes from Module 04 and small nodes indicating extended nodes, where only links with interaction degree ≥ 0.2 are shown.

eases 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 [30, 50]. Epidemiological studies had shown a reciprocal inverse relationships between cancer and neurodegeneration according to abnormal cell growth and cell loss in common [61, 74].

3.3.7 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 small size as well as functional enrichment performance. Fig. 3.8(A) and

Table 3.4: Functional annotation of extended Module 04

DataBase	Pathways/Biological processes/Diseases	p-value (corrected)
GO-BP	Adult behavior	1.93E-02
	Cognition	3.14E-02
	Intraspecies interaction between organisms	1.31E-02
	Learning	1.12E-02
	Learning or memory	1.12E-02
	Multi-organism behavior	1.44E-02
	Single-organism behavior	2.84E-02
	Social behavior	1.20E-02
	Vocalization behavior	9.64E-03
KEGG	Adrenergic signaling in cardiomyocytes	1.71E-02
	Chagas disease (American trypanosomiasis)	3.86E-02
	Cholinergic synapse	1.10E-02
	Circadian entrainment	3.69E-02
	Dopaminergic synapse	1.51E-02
	Estrogen signaling pathway	3.72E-02
	Gastric acid secretion	2.47E-02
	GnRH signaling pathway	3.30E-02
	Inositol phosphate metabolism	1.63E-02
	Insulin secretion	2.77E-02
	Long-term potentiation	1.88E-02
	Melanogenesis	3.68E-02
	Pancreatic secretion	3.65E-02
	Phosphatidylinositol signaling system	2.92E-02
	Salivary secretion	3.24E-02
OMIM	Autism spectrum disorder	1.07E-02
	Autistic disorder	1.61E-02
	Developmental disorder of mental health	2.70E-02
	Pervasive developmental disorder	1.20E-02

Fig. 3.8(B) 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 Table 3.4.

3.4 DISCUSSIONS AND CONCLUSIONS

We have proposed a top-down module identification method by integrating tissue-specific functional interaction network with imaging GWAS results in machine learning models 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 top 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 suggest 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.

In conclusion, 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 AD. 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.

Chapter 4

TISSUE-SPECIFIC NETWORK-BASED GWAS FOR IDENTIFYING FUNCTIONAL INTERACTION MODULES: A GWAS TOP NEIGHBOR BASED FRAMEWORK

In this chapter, we develop a GWAS top-neighbor-based module identification framework and compared it with Ridge and SVM based approaches proposed in Chapter 3. Modules conserving both tissue specificity and GWAS discoveries are identified, showing the promise of the proposal method for providing additional insight on the molecular mechanism of neurodegenerative diseases.

4.1 BACKGROUND

As introduced in Chapter 2, most network-based GWAS of QTS are using tissue-free biological networks such as human PPI network, without considering tissue specificity. We have developed a machine learning-based module identification framework and applied to amygdala imaging genetics study in Chapter 3. The experiment results prove the benefits from tissue-specific functional network, as well as the top-down strategy.

In this chapter, we propose a new GWAS top-neighbor-based searching approach for module identification, and compare with the machine learning-based approaches. This tnGWAS strategy extracts densely connected modules from top GWAS findings, based on the hypothesis that relevant modules consist of top GWAS findings and their close neighbors. Of note, machine learning-based methods (e.g., SVM and Ridge) provide re-prioritized gene findings, while tnGWAS does not. We demonstrate the

effectiveness of the proposed framework by applying it to a hippocampal imaging genetics analysis in the study of AD. We also applied the machine learning-based framework to confirm its re-prioritization and module identification performance using a new data set.

4.2 MATERIALS AND METHODS

To demonstrate the implementation of tnGWAS and machine learning-based approaches 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 [35, 59]. We propose to identify imaging QT-relevant modules, by integrating a hippocampus-specific functional interaction network and GWAS results of hippocampal FDG-PET measures.

4.2.1 IMAGING DATA, GENOTYPING DATA AND GWAS

Imaging data were obtained from the ADNI (`adni.loni.usc.edu`). Preprocessed FDG-PET scans were downloaded from LONI, and [^{18}F]FDG-PET measurements of hippocampus were extracted based on the MarsBaR AAL atlas. Genotype data were also obtained from LONI, of which 989 non-Hispanic Caucasian participants (Table 3.1) with complete baseline FDG-PET hippocampus measurements were studied. The detail of genotype data preprocessing have been described in Section 3.2.1. Association between the average FDG-PET measure in the hippocampal region at the baseline and 5,574,300 SNPs was examined by GWAS using PLINK [69]. 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 [60], given the performance of this gene-based association approach illustrated in Chapter 3. A number of 17,881 protein-coding gene p-values were obtained. The number of genes included in this analysis is less than previous analysis (Study of Chapter 3) which is 20,168, as only protein-coding genes are considered in this study.

4.2.2 HIPPOCAMPUS FUNCTIONAL INTERACTION NETWORK

A hippocampus-specific functional interaction network was downloaded from GIANT (`http://giant.princeton.edu/`). Interactions among 17,881 protein-coding genes was extracted after mapping to GWAS results. The weights of interactions range from 0 to 1, where larger measures represent higher interactions.

4.2.3 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 [47], 15 AD-relevant genes from the Online Mendelian Inheritance in Man Disease database (OMIM), and 40 significant candidates from AlzGene database (`http://www.alzgene.org/`). The detail of documented AD genes can be found in Section 3.2.3.

4.2.4 TNGWAS MODULE IDENTIFICATION FRAMEWORK

GWAS top-neighbor-based module identification approach was proposed and compared with previously developed machine learning-based strategy. Below we describe details of tnGWAS as well as briefly recall the machine learning-based ones.

MACHINE LEARNING BASED GWAS RE-PRIORITIZATION

Following [28], we trained an SVM model using hippocampus-specific network connectivity as features and significant or nonsignificant status based on nominal $p = 0.01$ as labels to re-prioritize GWAS results. In addition to SVM, we trained a Ridge model using also the network data as features while real p-values as responses. Different from classification which required a pre-defined threshold, regression approach utilizes more information from continuous p-values.

We trained SVM and Ridge models using interactions between a subset of genes C and all genes as features, and gene-level p-values of C as responses (positive or negative labels for SVM). To balance the training data, set C was constructed from 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. In our experiment, we employed z-scores instead of p-values due to their normal distribution. 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.

To demonstrate the performance of re-prioritization, we accessed the mean interactions and the AUC of re-prioritized genes from Ridge and SVM with original GWAS using 66 documented AD candidates as gold standard positive.

TNGWAS

Starting from a set of significant GWAS findings, tnGWAS includes their immediate neighbors in the result. It 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 ≥ 1 significant genes. In the experiment, we applied gene p-value $\leq$ 1e-7 to select significant GWAS findings, and interaction weight ≥ 0.3 to define high 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 a larger number of 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 framework 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 [2] 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.

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

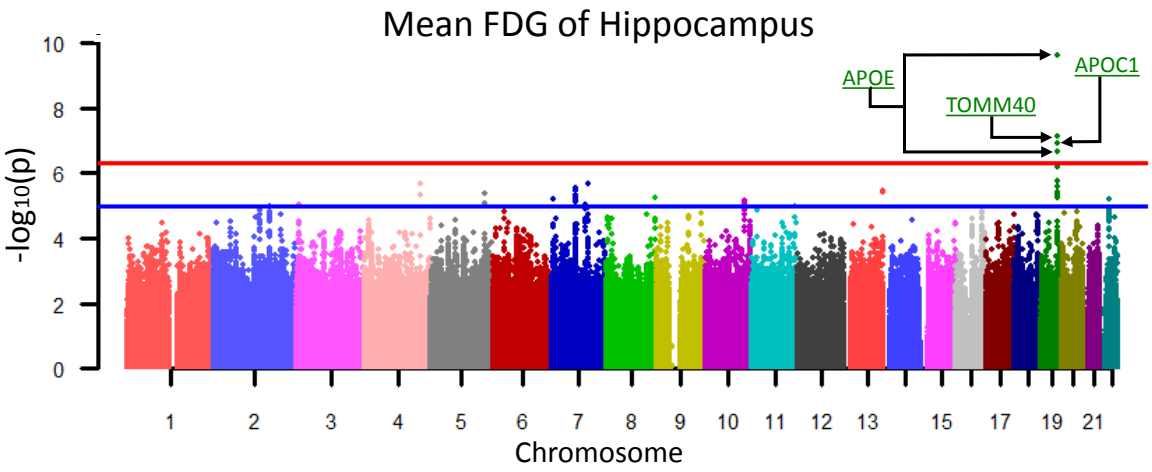

Figure 4.1: Manhattan plot of the FDG measure in the hippocampal region. Blue line indicates suggestive association threshold 5E-5 while red line indicates genome-wide significant threshold 5E-7.

FUNCTIONAL ANNOTATION

To assess functional relevance of the identified modules, we tested their over-representation on specific neurobiological functions and signalling pathways. We analyzed functional annotation using KEGG pathways and GO-BP terms.

4.3 EXPERIMENTAL RESULTS

4.3.1 GWAS OF HIPPOCAMPUS IQT

GWAS was performed to examine genetic associations between SNPs and the hippocampal FDG-PET measure. Four SNPs were identified as significant using $p \leq$ 5E-7 (see Fig. 4.1 for the Manhattan plot), including two within *APOE*, one within *TOMM40* and one within *APOC1*. After mapping to 17,881 protein coding regions, four genes were identified to be significant associations: *APOC1*, *APOE*, *PVRL2* and *TOMM40*.

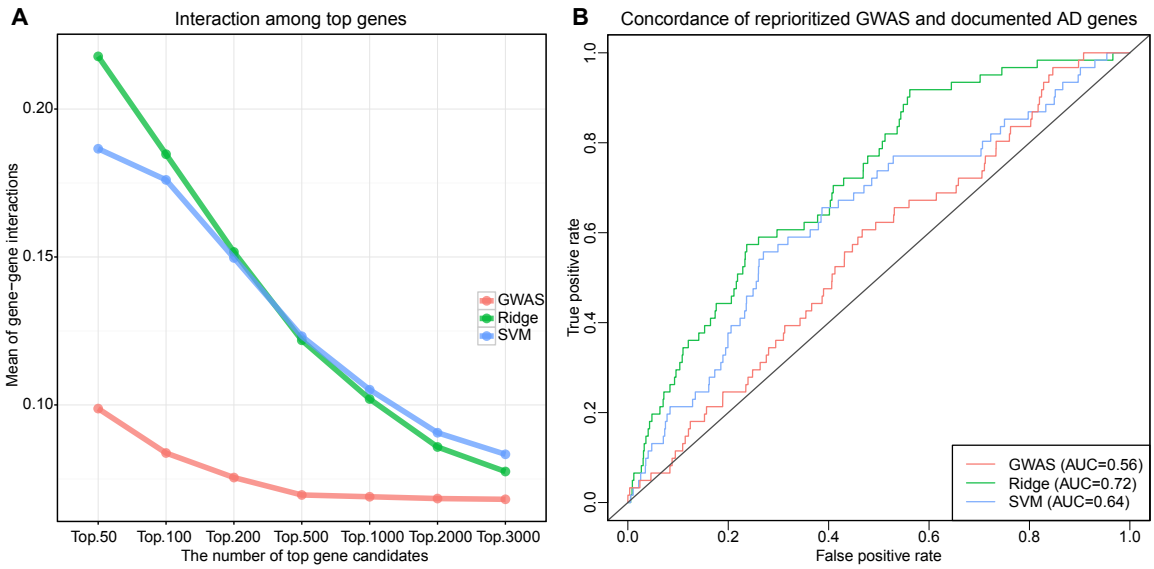

Figure 4.2: Performance evaluation of re-prioritized results. (A) Mean interaction measures among top N findings (N ranging from 50 to 3000) of three methods on hippocampus. (B) ROC curves with AUC results on hippocampus, to measure the concordance between the GWAS/NetWAS findings and the documented AD genes.

4.3.2 MACHINE LEARNING BASED RE-PRIORITIZATION

As mentioned earlier, top predictions from machine learning based re-prioritization would conserve both densely functional interaction and strong phenotype-relevance. Since tnGWAS did not assign ranks to top neighbors, we compared top predictions from Ridge, SVM with original GWAS to access their re-prioritization performance. Mean statistics of functional interactions and AUC were assessed on different scales of top predictions and shown in Fig. 4.2.

From Fig. 4.2(A), both Ridge and SVM yielded much stronger connectivity than GWAS. Dense interaction among top predictions demonstrated the advantage of network-based integration. From Fig. 4.2(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 inter-

acted [14, 20, 64]. Ridge performed better than SVM in both evaluations, suggesting that continuous p-values do provide more valuable information than significance status. Combined with results from Chapter 3, we confirmed the outstanding re-prioritization performance of Ridge-based NetWAS.

4.3.3 HIPPOCAMPUS-RELEVANT TOP PREDICTIONS

We compared the functional connectivity of top findings among tnGWAS, two machine learning-based methods (Ridge and SVM), 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 4.2.4). Fig. 4.3 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 including immediate neighbors, gained 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-based approaches seem to perform better as a whole as they integrate GWAS results and the tissue-specific network in a better fashion.

4.3.4 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

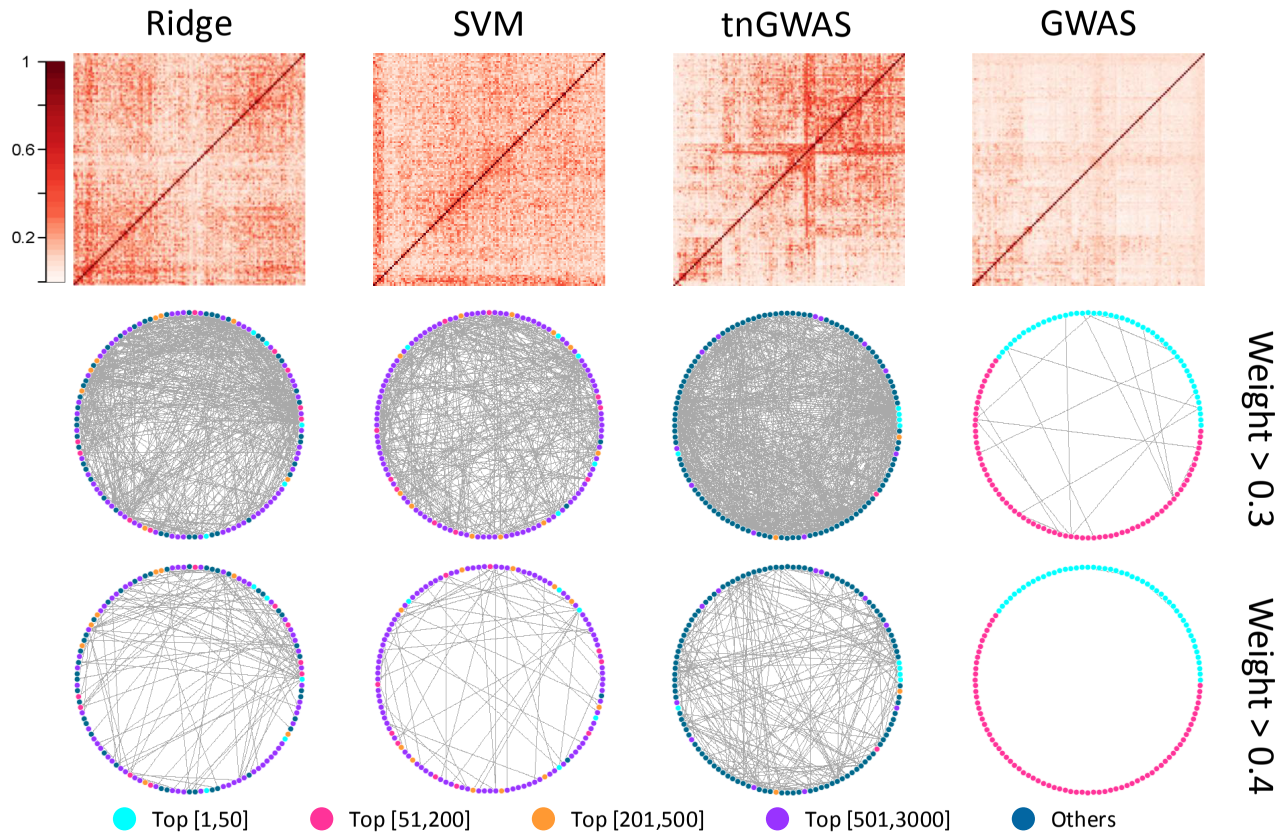

Figure 4.3: Comparison of top 124 findings from tnGWAS, Ridge, SVM and original GWAS. Heatmaps show the complete interaction matrix of top predictions. Circular networks show interactions after filtering weak connections. Nodes in circular network are colored based on their ranks in original GWAS result.

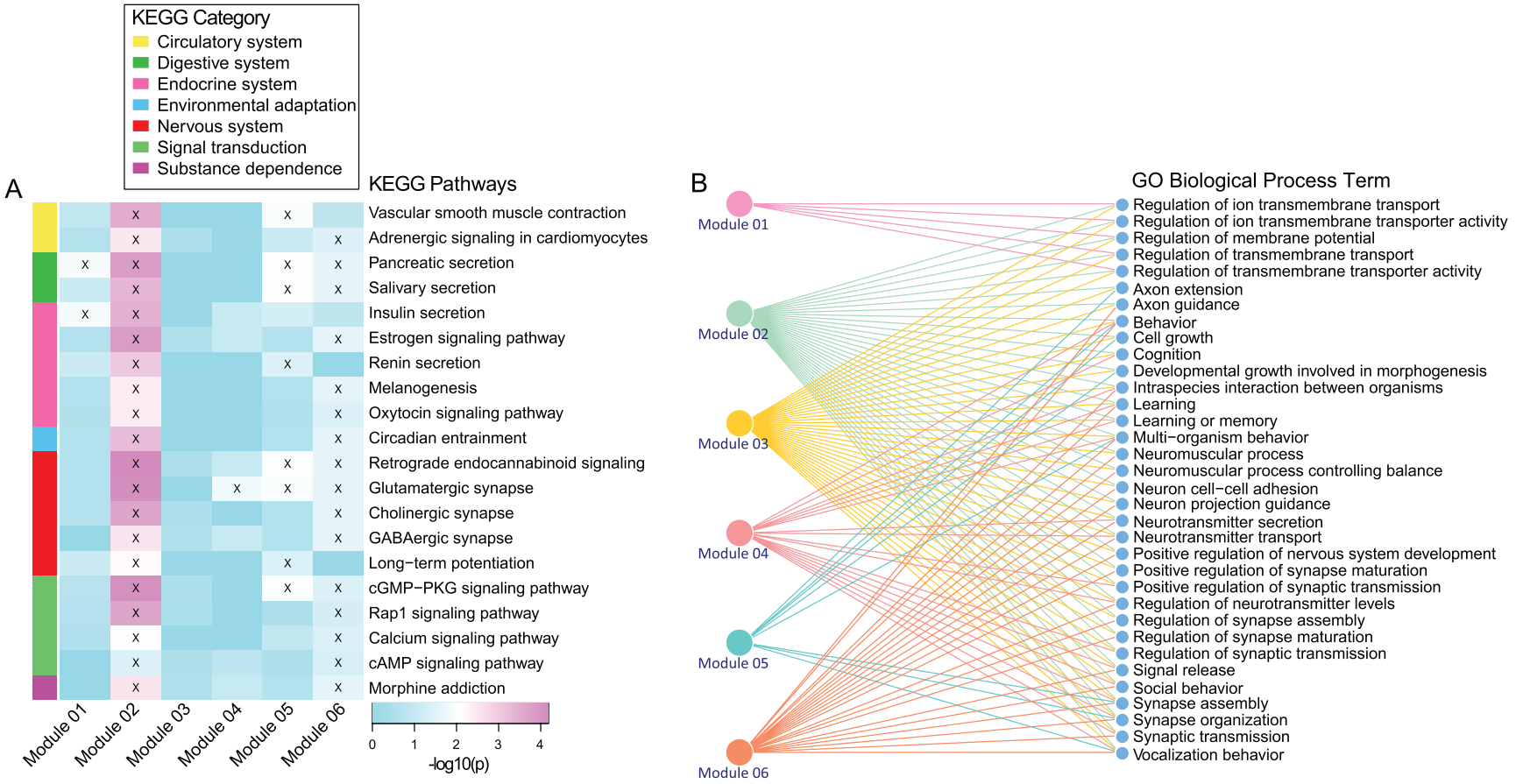

Figure 4.4: Functional annotation of modules from Ridge.

Table 4.1: 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

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. 6 out of 21 were significantly enriched by top 50 GWAS findings; see Table 4.1. Functional annotation was applied to further examine functional relevance of identified modules. Fig. 4.4 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 [7]). Fig. 4.4(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.

4.4 DISCUSSIONS AND CONCLUSIONS

We have proposed two top-down module identification frameworks: machine learning-based and GWAS top-neighbor-based. 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 GWAS enrichment assessment implies both tissue and GWAS specificity 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.

Chapter 5

TWO-DIMENSIONAL ENRICHMENT ANALYSIS PARADIGM FOR MINING HIGH-LEVEL IMAGING GENETIC ASSOCIATIONS

Enrichment analysis has been widely applied in the GWAS, where gene sets corresponding to biological pathways are examined for significant associations with a phenotype to help increase statistical power and improve biological interpretation. In this work, we expand the scope of enrichment analysis into brain imaging genetics, an emerging field that studies how genetic variation influences brain structure and function measured by neuroimaging QT. Given the high dimensionality of both imaging and genetic data, we propose to study Imaging Genetic Enrichment Analysis (IGEA), a new enrichment analysis paradigm that jointly considers meaningful GS and BC and examines whether any given GS-BC pair is enriched in a list of gene-iQT findings. Using gene expression data from Allen Human Brain Atlas and imaging genetics data from Alzheimer's Disease Neuroimaging Initiative as test beds, in this chapter, we present an IGEA framework and conduct a proof-of-concept study. This empirical study identifies 25 significant high-level two-dimensional imaging genetics modules. Many of these modules are relevant to a variety of neurobiological pathways or neurodegenerative diseases, showing the promise of the proposal framework for providing insight into the mechanism of complex diseases.

5.1 BACKGROUND

Brain imaging genetics is an emerging field that studies how genetic variation influences brain structure and function. GWAS has been performed to identify genetic

markers such as SNPs that are associated with brain iQTs [77, 79]. Using biological pathways and networks as prior knowledge, enrichment analysis has also been performed to discover pathways or network modules enriched by GWAS findings to enhance statistical power and help biological interpretation [31]. For example, numerous studies on complex diseases have demonstrated that genes functioning in the same pathway can influence iQTs collectively even when constituent SNPs do not show significant association individually [72]. Enrichment analysis can also help identify relevant pathways and improve mechanistic understanding of underlying neurobiology [32, 48, 62, 71].

In the genetic domain, enrichment analysis has been widely studied in gene expression data analysis to test the functional relevance of differential expressed genes; and has recently been modified to analyze GWAS data to assess the collective effects of a set of significant GWAS findings. GWAS-based enrichment analysis first maps SNP-level scores to gene-based scores, and then test whether a pre-defined gene set S (e.g., a pathway) is enriched in a set of significant genes L (e.g., GWAS findings). As we introduced in Chapter 2, two strategies are often used in genetic enrichment analysis to compute the enrichment significance: over-representation test [17, 18, 41, 92] and rank-based test [87]. Over-representation approaches aim to solve an independence test problem (e.g., χ^2 test, hypergeometric test, or binomial z-test) by treating genes as significant if their scores exceed a threshold. Rank-based methods take into account the score of each gene to determine if the members of S are randomly distributed throughout L.

In brain imaging genetics, the above enrichment analysis methods are applicable only to genetic findings associated with each single iQT. Our ultimate goal is to

discover high-level associations between meaningful gene sets and brain circuits, which typically include multiple genes and multiple iQTs. To achieve this goal, we propose to study Imaging Genetic Enrichment Analysis (IGEA), a new enrichment analysis paradigm that jointly considers sets of interest (i.e., GS and BC) in both genetic and imaging domains and examines whether any given GS-BC pair is enriched in a list of gene-iQT findings.

Using whole brain whole genome gene expression data from Allen Human Brain Atlas (AHBA) and imaging genetics data from ADNI as test beds, we present a novel IGEA framework and conduct a proof-of-concept study to explore high-level imaging genetic associations based on brain-wide genome-wide association study (BWGWAS) results. For consistency purpose, in this study, we use GS to indicate a set of genes and BC to indicate a set of ROIs in the brain. The proposed framework consists of the following steps (see also Figure 5.1): (1) conduct BWGWAS on ADNI amyloid imaging genetics data to identify SNP-iQT and gene-iQT associations, (2) use brain-wide-genome-wide expression data from AHBA to construct meaningful GS-BC modules, (3) perform IGEA to identify GS-BC modules significantly enriched by gene-iQT associations using an over-representative strategy, and (4) visualize and interpret the identified GS-BC modules.

5.2 MATERIALS AND DATA SOURCES

To demonstrate the proposed IGEA framework for identifying two-dimensional imaging genetic modules, we apply it to the brain-wide amyloid imaging genetic analysis in the study of AD. "Amyloid cascade hypothesis" has been considered the leading pathogenesis of AD for decades where brain amyloid deposition is thought to be hap-

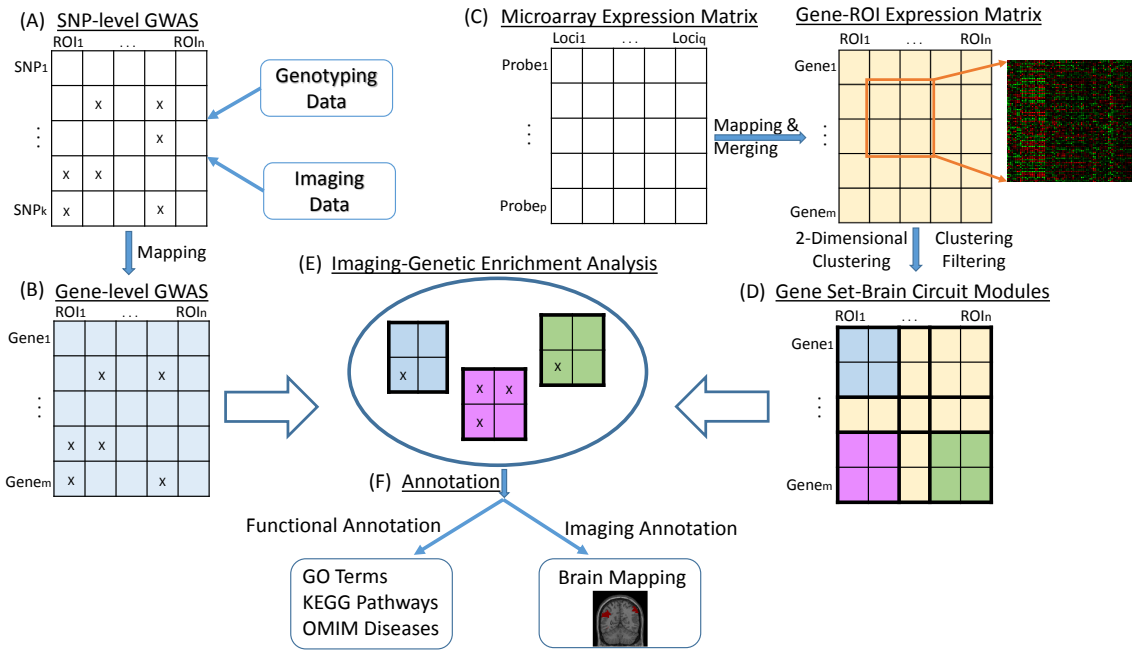

Figure 5.1: Overview of the proposed Imaging Genetic Enrichment Analysis framework. (A) Perform SNP-level GWAS of brain wide imaging measures. (B) Map SNP-level GWAS p-values to gene-based p-values. (C) Construct gene-ROI expression matrix from AHBA data. (D) Construct GS-BC modules by performing two-dimensional hierarchical clustering, and then filter out biclusters with an average correlation below a user-given threshold. (E) Perform IGEA by mapping gene-based p-values to the identified GS-BC modules. (F) For each enriched GS-BC module, examine the GS using GO terms, KEGG pathways, and OMIM disease databases, and visualize the identified BC by mapping to brain.

Table 5.1: 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	231	90	288	196	175
Gender (M/F)	119/112	36/54	163/125	114/82	105/70
Age(mean±sd)	76.18±6.64	72.49±5.72	71.67±7.25	73.85±8.49	75.26±7.76
Education(mean±sd)	16.43±2.67	16.80±2.61	16.12±2.63	16.35±2.80	15.86±2.73

pened over years before the early symptom of AD [52, 73], and can be measured by brain imaging methods.

5.2.1 BRAIN WIDE GENOME WIDE ASSOCIATION STUDY (BWGWAS)

The imaging and genotyping data used for BWGWAS were obtained from the ADNI 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 MCI and early AD. For up-to-date information, see www.adni-info.org.

Preprocessed [^{18}F]Florbetapir PET scans (i.e., amyloid imaging data) were downloaded from `adni.loni.usc.edu`, then aligned to the corresponding MRI scans and normalized to the MNI space as $2 \times 2 \times 2$ mm voxels. ROI level amyloid measurements were further extracted based on the MarsBaR AAL atlas. Genotype data of both ADNI-1 and ADNI-GO/2 phases were also downloaded, and then quality controlled, imputed and combined as described in [42]. A total of 980 non-Hispanic Caucasian participants (Table 5.1) with both complete amyloid measurements and

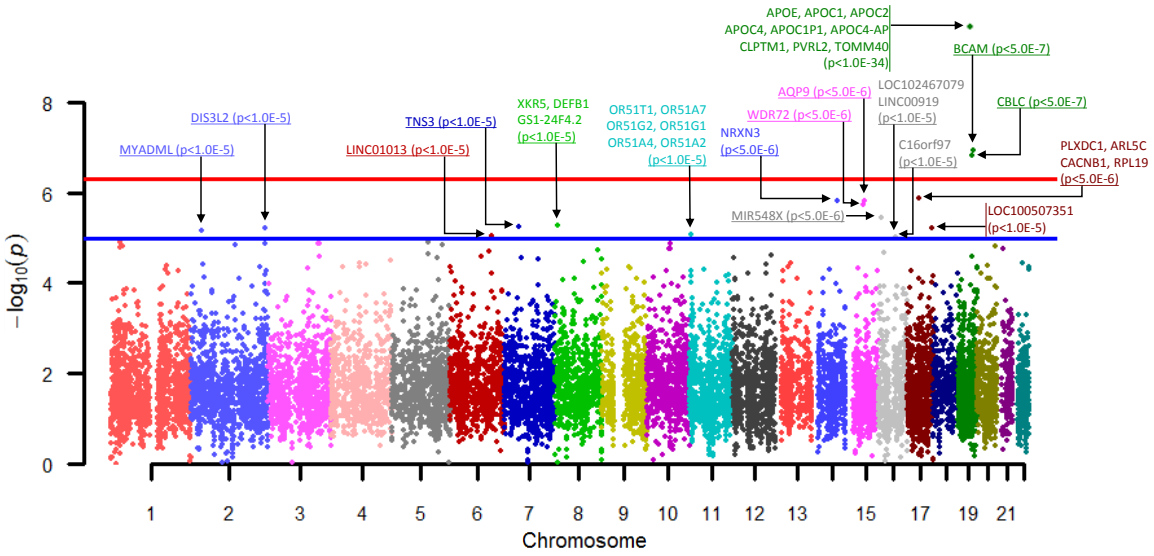

Figure 5.2: Manhattan plot of imaging quantitative genome wide association for AD individuals based on precuneus (right) measurement from amyloid imaging data. The x-axis represents the chromosomes and the y-axis represents $-\log_{10}(p)$, where p is the gene-based significance.

genome-wide data were studied. Associations between 105 (out of a total 116) baseline amyloid measures and 5,574,300 SNPs were examined by performing SNP-based GWAS using PLINK [69] with sex, age and education as covariates. To facilitate the subsequent enrichment analysis, a gene-based p-value was determined as the smallest p-value of all SNPs located in $\pm$20K bp of the gene [60].

5.2.2 CONSTRUCTING GS-BC MODULES USING AHBA

There are many types of prior knowledge that can be used to define meaningful GS and BC entities. In the genomic domain, the prior knowledge could be based on Gene Ontology or functional annotation databases; in the imaging domain, the prior knowledge could be neuroanatomic ontology or brain databases. In this work, to demonstrate the proposed IGEA framework, we use gene expression data from the Allen Human Brain Atlas (AHBA, Allen Institute for Brain Science, Seattle, WA;

available from http://www.brain-map.org/) to extract GS and BC modules such that genes within a GS share similar expression profiles and so do ROIs within a BC. We hypothesize that, given these similar co-expression patterns across genes and ROIs, each GS-BC pair forms an interesting high level imaging genetic entity that may be related to certain biological function and can serve as a valuable candidate for two-dimensional IGEA.

The AHBA includes genome-wide microarray-based expression covering the entire brain through systematic sampling of regional tissue. Expression profiles for eight health human brains have been released, including two full brains and six right hemispheres. One goal of AHBA is to combine genomics with the neuroanatomy to better understand the connections between genes and brain functioning. As an early report indicated that individuals share as much as 95% gene expression profile [100], in this study, we only included one full brain (H0351.2001) to construct GS-BC modules. First all the brain samples (~ 900) were mapped to MarsBaR AAL atlas, which included 116 brain ROIs. Due to many-to-one mapping from brain samples to AAL ROIs, there are > 1 samples for each ROI. Following [99], samples located in the same ROI were merged using the mean statistics. Probes were then merged to genes using the same strategy. Finally the preprocessed gene-ROI profiles were normalized for each ROI. As a result, the expression matrix contained 16,076 genes over 105 ROIs.

We use $\mathbf{E}$ to denote this expression matrix, where $\mathbf{e}^i$ is the expression level of gene i across all the 105 ROIs in $\mathbf{E}$, and $\mathbf{e}_j$ is the expression profile of ROI j across all the 16,076 genes in $\mathbf{E}$. Given two genes i_1 and i_2, we use the Pearson correlation coefficient to define their dissimilarity $d_{\text{gene}}(i_1, i_2)$ as follows:

$$d_{\text{gene}}(i_1, i_2) = 1/2 \times (1 - \text{corr}((\mathbf{e}^{i_1})^T, (\mathbf{e}^{i_2})^T)). \tag{5.1}$$

Similarly, given two ROIs j_1 and j_2, we define their dissimilarity $d_{\text{roi}}(j_1, j_2)$ as follows:

$$d_{\text{roi}}(j_1, j_2) = 1/2 \times (1 - \text{corr}(\mathbf{e}_{j_1}, \mathbf{e}_{j_2})). \tag{5.2}$$

We performed a two-dimensional clustering analysis on $\mathbf{E}$ to identify interesting GS-BC modules. First, we calculated the distance matrices for both genes and ROIs, using Eq. (5.1) and Eq. (5.2), respectively. Next, two dendrograms were constructed by applying hierarchical clustering to two distance matrices separately, using the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algorithm [83]. After that, in the genomic domain, as most enrichment analyses placed constraints on genetic pathways of sizes from 10 to 200 [72], we cut the dendrogram at half of its height to build genetic clusters (i.e., GSs) whose sizes are mostly within the above range. Finally, in the imaging domain, we also employed the same parameter to construct ROI clusters (i.e., BCs).

Let $\mathbf{X}$ be a GS-BC module with n genes and m ROIs, where $\mathbf{x}^i$ is the expression level of gene i across all the m ROIs in $\mathbf{X}$, and $\mathbf{x}_j$ is the expression profile of ROI j across all the n genes in $\mathbf{X}$. For each pair of genes in $\mathbf{X}$, i.e., $((\mathbf{x}^{i_1})^T, (\mathbf{x}^{i_2})^T)$, we calculate its correlation coefficient. For each pair of ROIs in $\mathbf{X}$, i.e., $(\mathbf{x}_{j_1}, \mathbf{x}_{j_2})$, we also calculate its correlation coefficient. After that, we transform each of these correlation coefficients, say c, to Fisher's z-statistic $z(c)$ using the following Eq. (5.3):

$$z(c) = \frac{1}{2}\log\left(\frac{1+c}{1-c}\right). \tag{5.3}$$

We then define $\overline{z}_{\text{gene}}(\mathbf{X})$, the *gene-based* average Fisher's z-statistics of correlation coefficient of $\mathbf{X}$, as follows:

$$\overline{z}_{\text{gene}}(\mathbf{X}) = \frac{2}{n(n-1)} \sum_{0<i_1<i_2\leq n} z(\text{corr}((\mathbf{x}^{i_1})^T, (\mathbf{x}^{i_2})^T)). \tag{5.4}$$

Similarly, we define $\overline{z}_{\text{roi}}(\mathbf{X})$, the *ROI-based* average Fisher's z-statistics of correlation coefficient of $\mathbf{X}$, as follows:

$$\overline{z}_{\text{roi}}(\mathbf{X}) = \frac{2}{m(m-1)} \sum_{0<j_1<j_2\leq m} z(\text{corr}(\mathbf{x}_{j_1}, \mathbf{x}_{j_2})). \tag{5.5}$$

Based on these average gene-based and ROI-based z-statistics, respectively, we select the top 20% of all the GS-BC modules and include those in our subsequent analyses, to ensure our studied modules have comparatively high co-expression profiles. Thus, in this work, we focus on the analysis of the following three types of GS-BC modules with top z-statistics:

1. *Gene-based*: These are the modules with relatively high co-expression profiles between genes, i.e., $\overline{z}_{\text{gene}}(\mathbf{X})$ is ranked in the top 20% of all the $\overline{z}_{\text{gene}}$ scores.

2. *ROI-based*: These are the modules with relatively high co-expression profiles between ROIs, i.e., $\overline{z}_{\text{roi}}(\mathbf{X})$ is ranked in the top 20% of all the $\overline{z}_{\text{roi}}$ scores.

3. *Gene&ROI-based*: Both (1) and (2) hold.

5.2.3 IMAGING GENETIC ENRICHMENT ANALYSIS (IGEA)

Pathway enrichment analysis has been extensively employed to genomic domain to analyze the genetic findings associated with a specific iQT. In this study, our goal is to

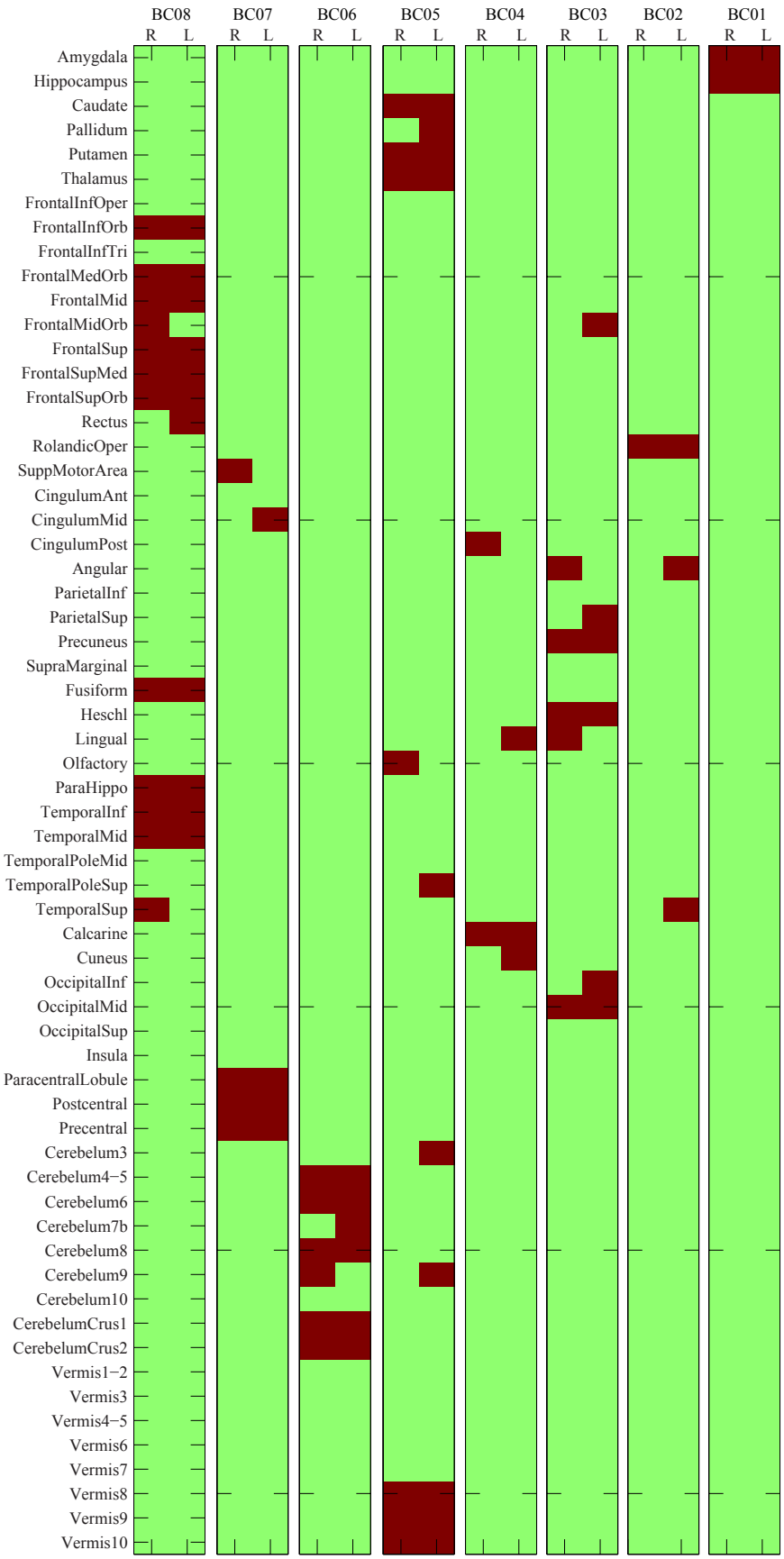

Figure 5.3: Eight unique brain circuits (BCs) identified from IGEA. ROIs belonging to each BC are colored in red.

identify high level associations between gene sets and brain circuits, which typically include multiple genes and multiple iQTs.

In this study, we propose the over-representative-based IGEA by extending the existing threshold-based enrichment analysis. SNP-level findings have been mapped to gene-level findings in Section 5.2.1. The GWAS findings are a list L of $N = N_G \times N_B$ gene-iQT associations, where we have a set G_d of $N_G = |G_d|$ genes and a set B_d of $N_B = |B_d|$ iQTs in our analysis. From Section 5.2.2, GS-BC modules have been constructed, where either relevant genes share similar expression profiles across relevant ROIs, or relevant ROIs share similar expression profiles across relevant genes, or both. Given an interesting GS-BC module with gene set G_k and iQT set B_k, IGEA aims to determine whether the target GS-BC module $T = \{(g, b)|g \in G_d \cap G_k, b \in B_d \cap B_k\}$ is enriched in L.

Now we describe our threshold-based IGEA method. We have N gene-iQT pairs from GWAS. Out of these, $n = |A|$ pairs (the set A) are significant ones with GWAS p-value passed a certain threshold. We also have $m = |P|$ (the set P) gene-iQT pairs from a given GS-BC module, and k significant pairs are from P. Using Fisher's exact test for independence, the enrichment p-value for the given GS-BC module is calculated as:

$$p\text{-value} = Pr(|A \cap P| \geq k) = \sum\nolimits_{i \geq k} \frac{\binom{m}{i} \times \binom{N-m}{n-i}}{\binom{N}{n}}. \tag{5.6}$$

Here, $Pr(\cdot)$ is the probability function.

5.2.4 EVALUATION OF THE IDENTIFIED GS-BC MODULES

For evaluation purpose, we tested the statistical significance of the IGEA results. We hypothesize that the gene-iQT associations from BWGWAS of the original data should be overrepresented in certain GS-BC modules, and the BWGWAS results on permuted data should not be enriched in a similar number of GS-BC modules. We performed the IGEA analysis on $n = 50$ permuted BWGWAS results, and estimated the p-value for the number of significant GS-BC modules discovered from the original data using a t-distribution with $n - 1$ degrees of freedom.

Given a BWGWAS result R, let $Prop(R)$ be the proportion of modules which are significantly enriched by R. Let R_{orig} be the original BWGWAS result, and $R_{\text{perm}(i)}$ be the i-th permuted BWGWAS result. Let $S = \{Prop(R_{\text{perm}(i)}) \mid 1 \leq i \leq n\}$ be the set of these proportion values for all the permuted results. Then the p-value is estimated using Eq. (5.7).

$$p\text{-value} = Pr\left(T_{n-1} \geq \frac{Prop(R_{\text{orig}}) - \mu_{\text{perm}}}{\sqrt{1 + 1/n} \times \sigma_{\text{perm}}}\right). \tag{5.7}$$

where T_{n-1} is the t-distribution with $n - 1$ degrees of freedom, μ_{perm} is the sample mean of S and σ_{perm} is the sample standard deviation of S.

To determine the functional relevance of the enriched GS-BC modules, we also tested whether genes from each module are over-represented for specific neurobiological functions, signaling pathways or complex neurodegenerative diseases. We performed pathway enrichment tests using GO terms, KEGG pathways, and OMIM disease database.

Table 5.2: Twenty-five significantly enriched GS-BC modules from IGEA. See also Section 5.3.2 and Fig. 5.3 for details about relevant GSs and BCs respectively.

Module ID	Top 20% CoExp[a]	BC ID	# of ROIs	GS ID	# of genes	p-value (G[b])	p-value (R[c])	p-value (G&R[d])
01	R[c]	BC07	8	GS01	81	-	2.61E-06	-
02	G, R, G&R[d]	BC02	4	GS02	168	9.06E-06	9.06E-06	9.06E-06
03	G[b]	BC03	11	GS02	168	2.54E-11	-	-
04	G, R, G&R	BC04	5	GS02	168	1.44E-06	1.44E-06	1.44E-06
05	G	BC05	14	GS02	168	6.42E-06	-	-
06	R	BC06	13	GS02	168	-	5.91E-07	-
07	R	BC08	23	GS02	168	-	5.65E-22	-
08	G, R, G&R	BC01	4	GS03	55	1.38E-06	1.38E-06	1.38E-06
09	G	BC02	4	GS03	55	4.39E-13	-	-
10	R	BC04	5	GS03	55	-	1.41E-15	-
11	G	BC05	14	GS03	55	1.01E-14	-	-
12	R	BC06	13	GS03	55	-	1.72E-08	-
13	R	BC07	8	GS03	55	-	2.40E-21	-
14	R	BC07	8	GS04	66	-	4.00E-07	-
15	G, R, G&R	BC01	4	GS05	19	3.83E-05	3.83E-05	3.83E-05
16	G, R, G&R	BC02	4	GS05	19	6.88E-09	6.88E-09	6.88E-09
17	G, R, G&R	BC04	5	GS05	19	2.64E-10	2.64E-10	2.64E-10
18	R	BC06	13	GS05	19	-	2.26E-11	-
19	G, R, G&R	BC07	8	GS05	19	1.54E-14	1.54E-14	1.54E-14
20	G, R, G&R	BC02	4	GS06	28	4.87E-08	4.87E-08	4.87E-08
21	G	BC02	4	GS07	24	7.69E-05	-	-
22	G&R	BC01	4	GS08	33	-	-	1.97E-04
23	G	BC02	4	GS08	33	1.11E-07	-	-
24	R	BC04	5	GS08	33	-	7.39E-09	-
25	G	BC02	4	GS09	111	4.07E-05	-	-

[a]To indicate whether the top 20% modules are selected based on the gene-based, ROI-based or gene&ROI-based strategy.

[b]G: Gene-based.

[c]R: ROI-based.

[d]G&R: Gene&ROI-based.

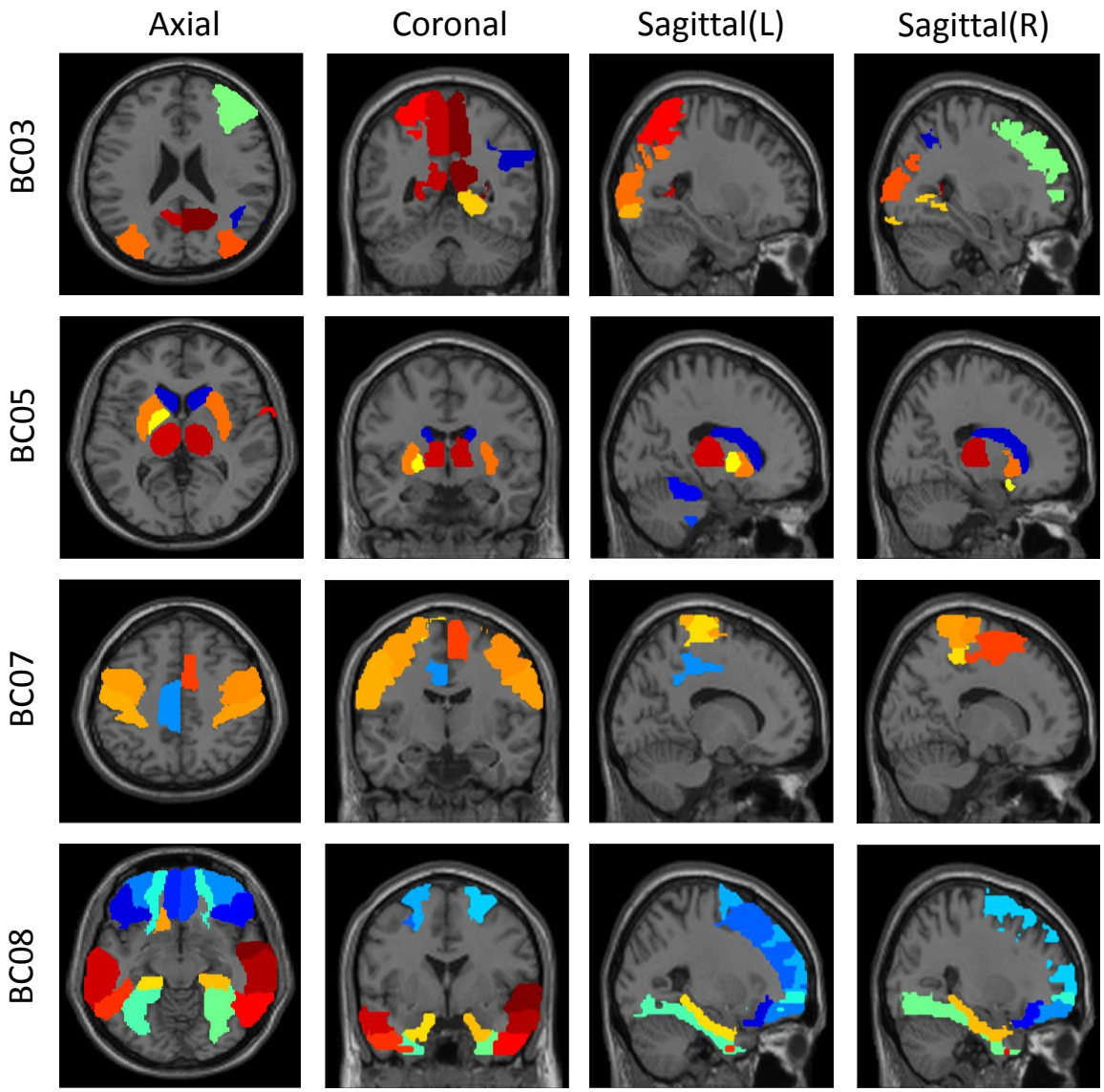

Figure 5.4: Brain maps of four brain circuits (BCs) identified from IGEA.

5.3 EXPERIMENTAL RESULTS

5.3.1 SIGNIFICANT GS-BC MODULES

By performing hierarchical clustering on both genetic and imaging domains, 171 out of 216 genetic clusters (only those with size ranging from 10 to 200) and 9 imaging clusters (with size ranging from 4 to 23, no clusters are excluded) were identified. 1,539 GS-BC modules were generated by combining each pair of genetic and ROI clusters. Two sets of 308 (20% of 1,539) modules were selected according to gene-based and ROI-based z-statistics, respectively. Among them, 90 modules were among top 20% in both gene-based and ROI-based ranking results. We used a moderate size thresholds for the selection, to avoid the exclusion of potentially interesting candidates.

For the BWGWAS results, we obtained $16,076 \times 105 = 1,687,980$ gene-iQT associations after mapping SNP-based p-values to genes. Out of these, 1,402 gene-iQT associations passed the BWGWAS p-value of 1.0E-5. Fig. 5.2 shows the gene-based GWAS result of an example iQT (i.e., the average amyloid deposition in the right precuneus). Precuneus amyloid concentration has been demonstrated to be associated with disordered activity in AD [34].

Three sets of constructed GS-BC modules (308, 308, and 90 with top z-statistics using gene-based, ROI-based and gene&ROI-based strategies respectively, see Section 5.2.2) were tested separately for whether they could be enriched by BWGWAS results using IGEA. Across three sets, totally 25 modules turned out to be significant after Bonferroni correction (see Table 5.2), of which 15, 17, and 9 are from gene-based, ROI-based, and both gene&ROI-based categories, respectively. We also tested the significance of the number of identified GS-BC modules. Compared to the

Table 5.3: Top enriched OMIM diseases of identified GSs.

GS ID	# of gene	OMIM Disease	p-value
GS01	81	Encephalopathy	4.2E-2*
		Dementia	3.6E-2*
GS02	168	Encephalopathy	5.0E-2
		Breast cancer	9.5E-2
GS03	55	Leukemia	2.7E-2*
		Alzheimer's disease	8.9E-2
GS04	66	Hypertension	5.0E-2
GS05	19	Anomalies	2.4E-2*
		Alzheimer's disease	4.5E-2*
GS06	28	Ectodermal dysplasia	2.0E-2*
GS07	24	Hypertension	3.4E-2*
		Spinocerebellar ataxia	4.3E-2*
GS08	33	Glycogen storage disease	1.6E-2*
GS09	111	Immunodeficiency	1.4E-2*

*Significantly enriched.

permuted BWGWAS results, the analysis on the original data yielded a significantly larger number of enriched GS-BC modules with estimated p-values of 7.6E-25, 1.2E-9, and 1.8E-25, corresponding to gene-based, ROI-based, and gene&ROI-based strategies respectively, indicating that imaging genetic associations existed in these enriched GS-BC modules.

Across all 25 identified modules, there are 9 and 8 unique GS and BC entities respectively. Fig. 5.3 shows the 8 unique identified BCs with corresponding ROI names, and Fig. 5.4 maps four of those onto the brain. For example, BC03 and BC04 include structures that are major spots for amyloid accumulation in AD (e.g., cingulum, precuneus). BC05 involves structures responsible for motivated behaviors (e.g., caudate, pallidum, putamen) and sensory information processing (e.g., thalamus). BC08 involves various frontal regions responsible for executive functions. Details of all 25 modules are listed in Table 5.2. We can find that some modules share common gene sets with different brain circuits, and some share the same brain circuits with

different gene sets. This illustrates the complex associations among multiple genes and multiple brain ROIs.

5.3.2 PATHWAY ANALYSIS OF IDENTIFIED GS-BC MODULES

To explore and analyze functional relevance of our identified GS-BC modules, we performed pathway enrichment analysis from three aspects including GO terms, functional pathways and diseases using GO terms, KEGG pathways and OMIM diseases databases, respectively.

Fig. 5.5 shows the KEGG pathway enrichment results which were mapped to 15 categories. From the results, most identified GSs had a number of significant functional enrichments. Several of them were directly related to the neurodegenerative disease and its development, e.g., Alzheimer's disease enriched in GS05 and Parkinson's diease enriched in GS01. Another major part of them were also related to the neurodegenerative diseases and their development. For instance, caffeine as the most widely used psychoactive substance, its metabolism (from GS09 located in Module 25) can affect brain metabolism and has potential benefits on Parkinson's Disease treatment [66]. There are also several enriched pathways related to oxidative stress, which is a critical factor for a range of neurodegenerative disorders. For example, glycolysis and gluconeogenesis (from GS02 located in Modules 02-07) are associated with hypoxia, ischemia, and AD [13]. Gap junctions (from GS03 located in Modules 08-13) can couple various kinds of cells in the central nervous system (CNS) which play an important role in maintaining normal function. Signaling transduction like calcium signaling pathway (from GS03 located in Modules 08-13) playing key role in short- and long-term synaptic plasticity, has shown abnormality in many neurode-

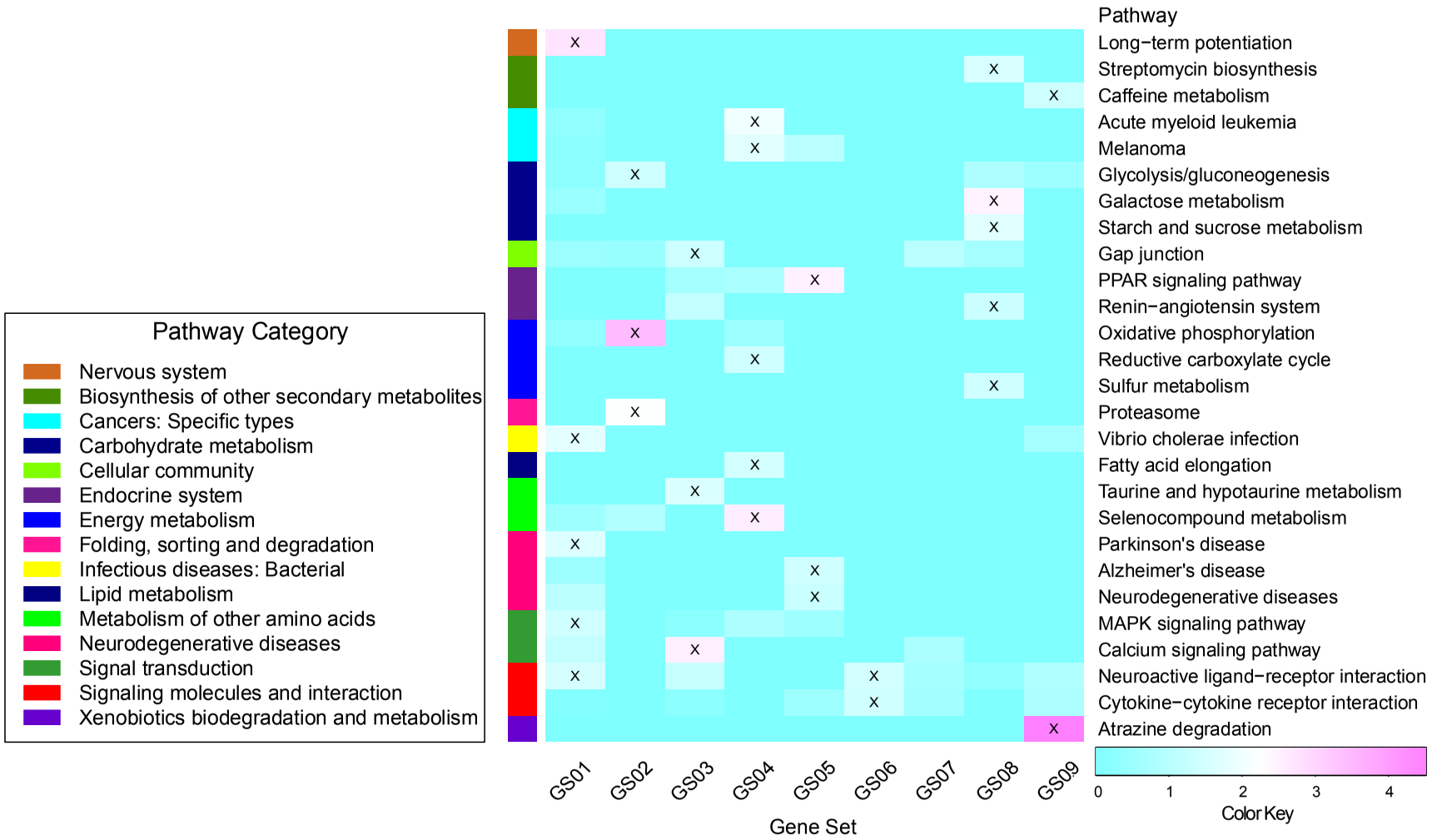

Figure 5.5: Results of KEGG pathway enrichment for identified GSs. The x-axis represents unique GS ID, and y-axis represents $-log_{10}(p)$ of enrichment significance of KEGG pathways. Marked cell represents significant enrichment (p-value <0.05).

generative disorders including Alzheimer's Disease, Parkinson's disease, amyotrophic lateral sclerosis (ALS), Huntington's disease, spinocerebellar ataxias (SCA) and so on [7].

Table 5.3 shows the OMIM disease enrichment results. Several neurodegeneration-related and age-related diseases and complex disorders were enriched in various gene sets, such as Alzheimer's disease from GS03 and GS05, Encephalopathy from GS01 and GS02, and Anomalies from GS05. Besides neurodegeneration diseases and disorders, several cancer-related entities be detected including breast cancer from GS02 and leukemia from GS03. These findings provided potential evidence for the studies that focused on investigating the relationship between cancer and neurodegeneration, with abnormal cell growth and cell loss in common.

GO enrichment result indicates the relationship between identified GSs and GO terms from three categories including biological process (BP), cellular component (CC), and molecular function (MF) (`http://geneontology.org/`). For the GO enrichment of all 9 gene sets, 163 various GO terms were significantly enriched. Top enriched terms were selected and grouped to 7 categories including behavior, cell communication, mitochondrion, metabolic process, neurological system process, response to stimulus, and signal transduction, as shown in Table 5.4. A large number of these terms have direct or indirect relationships with neurodegenerative diseases or phenotypes.

5.4 DISCUSSIONS AND CONCLUSIONS

We have presented a two-dimensional imaging genetic enrichment analysis (IGEA) framework to explore the high level imaging genetic associations by integrating whole

Table 5.4: Top enriched GO terms of GSs from identified GS-BC modules.

Group	GS ID	# of genes	GO Category	Corrected p-value
Behavior	GS03	55	Behavior	2.2E-2
			Learning or memory	4.4E-2
Cell Communication	GS01	81	Regulation of synaptic transmission	2.7E-6
			Neuron-neuron Synaptic transmission	2.9E-3
	GS03	55	Synaptic transmission	1.7E-4
Metabolic Process	GS05	19	Fat-soluble vitamin metabolic process	4.3E-2
			Organic hydroxy compound biosynthetic process	4.8E-2
	GS06	28	Regulation of translational termination	2.8E-2
Mitochondrion	GS02	168	Mitochondrial membrane part	2.5E-3
			Mitochondrial respiratory chain complex I	4.9E-3
Neurological System Process	GS03	55	Associative learning	1.1E-2
			Learning	4.5E-6
	GS09	111	Detection of chemical stimulus involved in sensory perception	1.1E-4
			Olfactory receptor activity	1.9E-5
Response To Stimulus	GS03	55	Response to amphetamine	2.0E-3
			Visual behavior	4.5E-3
	GS05	19	Response to cholesterol	3.6E-2
			Response to sterol	3.7E-2
	GS09	111	Detection of chemical stimulus	1.6E-4
Signal Transduction	GS01	81	Glutamate receptor signaling pathway	7.3E-4
	GS03	55	Adenylate cyclase-activating dopamine receptor signaling pathway	3.1E-3
			Dopamine receptor signaling pathway	1.4E-2
	GS05	19	Transmembrane receptor protein kinase activity	4.4E-2
	GS09	111	Olfactory receptor activity	1.9E-5

brain genomic, transcriptomic and neuroanatomic data. Traditional pathway enrichment analysis focused on investigating genetic findings of a single phenotype one at a time, and relationships among iQTs could be ignored. Such approach could be inadequate to provide insights into the mechanisms of complex diseases that involve multiple genes and multiple iQTs. In this chapter, we have proposed a novel enrichment analysis paradigm IGEA to detect high level associations between gene sets and brain circuits. By jointly considering the complex relationships between interlinked genetic markers and correlated brain imaging phenotypes, IGEA provides additional power for extracting biological insights on neurogenomic associations at a systems biology level and new insights into the complex associations among multiple genes and multiple ROIs, which can be treated as candidates to examine mechanisms of AD more specifically. Take module GS03-BC05 for instance which is significantly enriched in GWAS findings, several ROIs (e.g. caudate, pallidum, and putamen) from BC05 have been indicated responsible for motivated behaviors [15], meanwhile both KEGG and GO functional enrichment results of GS03 show high relevance to behavior and normal function maintaining (see Figure 5.5 and Table 5.4).

The real power of IGEA, however, can be affected by several aspects. First, the constructed GS-BC modules should reflect the real relationships among genes as well as brain ROIs. Thus it is crucial to define meaningful gene sets and brain circuits. In our analysis, GSs and BCs were separately extracted from AHBA brain-wide expression data based on hierarchical clustering, which were then combined to provide GS-BC modules. This strategy was based on the idea that interlinked genetic markers (or brain ROIs) would conserve similar expression pattern, that is, would be highly co-expressed. Second, the statistical measure of enrichment evaluation can be

based on different strategies. We adopted hypergeometric test in our experiment to estimate the over-representation of our defined GS-BC modules to the list of gene-iQT pair.

Based on these two considerations, our proposed paradigma can be further improved. From our GS-BC module construction, GSs (or BCs) are clustered together based on their co-expression pattern across all the ROIs in the whole brain (or across all the genes in the genome). Although statistical measures was calculated using Fisher's z-transformation to restrict our analyses on only highly co-expressed modules from our bi-clustering results, we could be missing other highly co-expressed GSs (or BCs) if they only had similar expression patterns on a small set of ROIs (or genes). In other words, our module construction strategy considered the global expression pattern but ignored the local ones. It is worth further investigation to try other reasonable strategies by applying prior knowledges such as pre-defined genetic pathways/networks or brain circuits, or by using different co-clustering algorithms (e.g., [95]) to take into consideration of relevant local expression patterns.

Hypergeometric test requires a pre-defined threshold to determine the list of gene-iQT pairs. Another limitation is that it considers only the count of significant gene-iQT pairs, but ignores the strength of gene-iQT associations. There are a number of rank-based enrichment analysis methods (e.g. GSEA [87]) that can be employed in our two-dimensional enrichment analysis to overcome these disadvantages. Another issue is that we used the smallest SNP-level p-value within the gene to represent the gene-based p-value. Therefore, another possible future direction is to explore other set-based methods for calculating gene-based p-values such as VEGAS [51], GATES [49] and so on. Besides, from mathematical perspective, associating GS-BC

modules and gene-iQT findings can be seen as a similarity discovery over two matrices that would be addressed from machine learning perspectives like the study proposed by Wang et al. [94].

Chapter 6

CONCLUSIONS

In this final chapter, we summarize the contributions of this thesis and discuss directions for future work.

6.1 SUMMARY

In this thesis, we investigate high-level imaging genetic association strategies and their applications in neurodegenerative disease for discovering disease-relevant modules. Existing approaches focus on only genetic domains, as well as overlook the tissue-context of genetic functional interactions. The main contributions of this thesis involve designing novel models for understanding high-level imaging genetic associations, and are summarized as follows:

Module identification: Network analysis of genomics data has been applied as complementary approach to GWAS and has promoted the understanding of molecular mechanisms for complex diseases and phenotypes. In this work, we propose NetWAS-based module identification framework with following threefold novelties: (i) expands the NetWAS scope from GWAS re-prioritization to module identification; (ii) introduces regression models into NetWAS to embrace the complete coverage of the continuous p-value spectrum; and (iii) 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. This proposed module identification strategy is among the first to incorporate tissue context with GWAS data to understand underpinning genetic functional interaction in a precise

way. It is applied to a real amygdala imaging genetics analysis in the study of AD. The constructed modules from this approach yield both strong tissue-specific interactions and disease-relevances, confirm the hypothesis that GWAS significant findings are enriched among nominally significant and functional interacted genes.

We further extend the above hypothesis and propose the tnGWAS to include immediate neighbors of top GWAS findings as disease-relevant candidates. The tnGWAS extracts densely connected modules from top GWAS findings, based on the hypothesis that relevant modules consist of top GWAS findings and their close neighbors. It is applied to a real hippocampal imaging genetics analysis in the study of AD, and yield the densest interaction among top candidate genes. Experimental results from both NetWAS-based and tnGWAS approaches demonstrate that precise context helps explore the collective effects of genes with biologically meaningful interactions specific to the studied diseases and phenotypes.

Imaging genetics enrichment analysis: In the second part of this thesis, we focus on the functional annotation of interested modules. Traditional enrichment analysis annotates biological functions for only gene sets. It is inadequate for imaging genetic analysis, where functional or structural interactions present among both genes and brain iQTs. We propose IGEA, a new enrichment paradigm that expands the scope of one-dimensional genetic enrichment analysis into brain imaging genetics. This integrative framework jointly investigates multi-omics data to form meaningful GSs and BCs, and examine whether any given GS-BC module is enriched in a list of gene-iQT findings. We demonstrate the power of proposed IGEA for providing additional insights on neurogenomic associations by applying it to BWGWAS of AD, where whole brain genomic, transcriptomic, and neuroanatomic data are integrated.

6.2 FUTURE DIRECTIONS OF RESEARCH

This thesis provides us a basis to continue to pursue the research in the area of high-level imaging genetic association analysis, which, we believe, has a host of fundamental problems yet to be solved, especially for large scale and heterogeneous multi-omics data covering both imaging and genetics domains. In this section, we discuss a few promising future research directions as follows.

Kernel-based module identification: In our proposed module identification frameworks, genetic interactions from functional network have been used as features to construct regression models. However, the interaction network can be directly used as kernel matrix to construct machine learning models. Instead of using the interaction network as features, the kernel function maps input data (i.e. features) into a high-dimensional space, which essentially represents the topological distance of inputs. Currently, the challenge is the raw interaction matrix is not or even not approximate positive semi-definite, which is the sufficient condition for a matrix to be a kernel. Given above observations, future efforts will be made to further explore the topological information embedded in tissue-specific networks. It is also of great interest to compare the performance of these two types of usage of network information.

Multiple-network analysis: We have used individual tissue-specific network as prior knowledge to discover disease- or phenotype-relevant genetic modules and have shown their promising performances. There are a number of tissue-specific networks have been constructed, among which quite a few are related to brain tissues. As introduced in Chapter 1, brain ROIs do not always have functions by each own, but

are always functional or structural grouped to play role. Through taking brain region functional relationships or brain region-specific networks similarity into account, the integration analysis of multiple networks with multiple corresponding GWAS results can be expected to shed more light on discovering functional modules. In future, we propose to expand our one-dimensional tissue-specific module identification framework to two-dimensional, for extracting multi-gene-multi-ROI modules which conserve significances across multiple ROI-specific networks analysis.

In addition, there are multi-omics data are available and multi-layered networks have been constructed including genetic, transcriptomic, proteomic, metabolic and so on. These networks can help reflect the conserving or complementary functional roles of biological components. These give us numerous opportunities to integrate different types of networks to construct modules and compare the modules obtained from multi-layer networks, to gain more comprehensive understanding of human complex diseases.

Efficient rank-based IGEA: In our proposed two-dimensional enrichment analysis, we employ hypergeometric test to evaluate the enrichment significance of constructed GS-BC modules. Over-representation strategy requires a pre-defined threshold to determine the list of gene–iQT pairs. Another limitation is that it considers only the count of significant gene–iQT pairs, but ignores the strength of gene–iQT associations. Rank-based approaches have been proposed and applied in genetic enrichment analysis, and can be expanded in our two-dimensional framework to overcome these limitations. However, due to the high dimensionality of both brain wide and genome wide scales as well as the computational efficiency of rank-based enrichment, it cannot be directly applied into two-dimensional enrichment framework. Given the

advantages of high performance computational frameworks and computing resources such as Map/Reduce, GPU programming, it will be appreciated to give a more scalable and efficient parameterization framework to facilitate the advanced enrichment analysis.

Bi-clustering for IGEA: In IGEA framework, we have applied hierarchical clustering to construct candidate two-dimensional GS-BC modules which have no overlaps with each other. In practice, both genetic and imaging modules are not independent but always overlap with one another. Bi-clustering approach and its extended approaches have been proposed for gene expression analysis, and have efficiently detected overlapped bi-dimensional clusters. However, these approaches cannot be directly applied in our brain-wide-gene-wide expression data, as we hypothesize that the modules hold both high-expression and local pattern. In the following work, we will develop novel bi-clustering method by adding additional constrains to detect candidate GS-BC modules which satisfy above conditions. The candidate modules would hold more biological meanings, which is essential for the next-step enrichment analysis.

The availability of large-scale biological data has greatly promoted the development of data-driven research, increased our knowledge on system biology and benefited the prediction of underlying biological processes. In imaging genetics, multi-omics data have been collected and provide us more opportunities as well as new challenges for understanding neurodegenerative diseases in a more comprehensive manner. Overall, our ultimate goal is to develop advanced computational biological methods to integrate the multi-omics and high-dimensional data, for helping provide more insights on the prediagnosis, preclinic and prevention of complex diseases.

REFERENCES

[1] The Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Research*, 38:D331–D335, 2010.

[2] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466:761, 2010.

[3] N. Akula, A. Baranova, D. Seto, J. Solka, M. A. Nalls, A. Singleton, et al. A network-based approach to prioritize results from genome-wide association studies. *PloS One*, 6(9), 2011.

[4] K. Alexandros, S. Alexandros, H. Kurt, and Z. Achim. kernlab–An S4 package for kernel methods in R. *J. Stat. Softw.*, 11(9):1–20, 2004.

[5] J. P. Andrawis, K. S. Hwang, A. E. Green, J. Kotlerman, D. Elashoff, J. H. Morra, et al. Effects of ApoE4 and maternal history of dementia on hippocampal atrophy. *Neurobiology of Aging*, 33(5):856 – 866, 2012.

[6] S. E. Baranzini, N. W. Galwey, J. Wang, P. Khankhanian, R. Lindberg, D. Pelletier, et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Human Molecular Genetics*, 18(11):2078–2090, 2009.

[7] I. Bezprozvanny. Calcium signaling and neurodegenerative diseases. *Trends Mol. Med.*, 15(3):89–100, 2009.

[8] A. Biffi, C. Anderson, R. Desikan, et al. Genetic variation and neuroimaging measures in Alzheimer disease. *Archives of Neurology*, 67(6):677–685, 2010.

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

[10] M. Bota and L. W. Swanson. BAMS Neuroanatomical Ontology: design and ismplementation. *Front Neuroinform*, 2:2, 2008.

[11] M. Bota and L. W. Swanson. Collating and curating neuroanatomical nomenclatures: principles and use of the brain architecture knowledge management system (BAMS). *Front Neuroinform*, 4:3, 2010.

[12] C. J. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov*, 2(2):121–67, 1998.

[13] D. Butterfield and M. Lange. Multifunctional roles of enolase in Alzheimer's disease brain: beyond altered glucose metabolism. *J Neurochem*, 111(4):915–33, 2009.

[14] H. J. Cordell. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

[15] M. R. Delgado, V. A. Stenger, and J. A. Fiez. Motivation-dependent responses in the human caudate nucleus. *Cereb Cortex*, 14(9):1022–30, 2004.

[16] R. S. Desikan, L. K. McEvoy, D. Holland, W. K. Thompson, J. B. Brewer, P. S. Aisen, et al. APOE E4 does not modulate amyloid-beta associated neurodegeneration in preclinical Alzheimer's disease. *AJNR Am J Neuroradiol*, 34(3):505–10, 2013.

[17] S. Draghici, P. Khatri, et al. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, 2003.

[18] S. Draghici, P. Khatri, et al. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*, 31(13):3775–81, 2003.

[19] S. Drăghici, P. Khatri, R. P. Martins, G. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, 2003.

[20] M. Emily et al. Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, 17(10):1231–1240, 2009.

[21] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, et al. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

[22] A. M. Fjell et al. CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease. *J. Neurosci.*, 30(6):2088–2101, 2010.

[23] P. T. Fox, A. R. Laird, S. P. Fox, P. M. Fox, A. M. Uecker, M. Crank, et al. BrainMap taxonomy of experimental design: description and evaluation. *Hum Brain Mapp*, 25(1):185–98, 2005.

[24] J. Friedman et al. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

[25] H.-H. Fu, D. K. J. Lin, and H.-T. Tsai. Damping factor in Google page ranking. *Applied Stochastic Models in Business and Industry*, 22(5-6):431–444, 2006.

[26] J. J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[27] N. S. Gowert et al. Blood platelets in the progression of Alzheimer's disease. *PloS One*, 9(2), 2014.

[28] C. S. Greene et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.*, 47(6):569–576, 2015.

[29] A. Hamaguchi et al. Sphingosine-dependent protein kinase-1, directed to 14-3-3, is identified as the kinase domain of protein kinase Cδ. *J. Biol. Chem.*, 278(42):41557–41565, 2003.

[30] M. T. Heneka et al. Innate immunity in Alzheimer's disease. *Nat. Immunol.*, 16(3):229–236, 2015.

[31] J. Hirschhorn. Genomewide association studies-illuminating biologic pathways. *N. Engl. J. Med.*, 360(17):1699–1701, 2009.

[32] M. G. Hong, A. Alexeyenko, et al. Genome-wide pathway analysis implicates intracellular transmembrane protein transport in Alzheimer disease. *J Hum Genet*, 55(10):707–9, 2010.

[33] T. Ideker and N. J. Krogan. Differential network biology. *Mol. Syst. Biol.*, 8:565, 2012.

[34] M. D. Ikonomovic, W. E. Klunk, E. E. Abrahamson, J. Wuu, C. A. Mathis, S. W. Scheff, et al. Precuneus amyloid burden is associated with reduced cholinergic activity in alzheimer disease. *Neurology*, 77(1):39–47, 2011.

[35] K. Ishii 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*, 48(5):704–711, 2007.

[36] C. J. Jack, P. Vemuri, H. Wiste, et al. Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. *Archives of Neurology*, 69(7):856–867, 2012.

[37] E. A. Jeong et al. Phosphorylation of 14-3-3ζ at serine 58 and neurodegeneration following kainic acid-induced excitotoxicity. *Anat. Cell Biol.*, 43(2):150–156, 2010.

[38] P. L. Jia et al. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, 27(1):95–102, 2011.

[39] K. A. Johnson et al. Brain imaging in Alzheimer disease. *Cold Spring Harb. Perspect. Med.*, 2(4), 2012.

[40] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.

[41] P. Khatri and S. Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–95, 2005.

[42] S. Kim et al. Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. *PLoS ONE*, 8(7):e70269, 2013.

[43] K. Lage, N. T. Hansen, E. O. Karlberg, A. C. Eklund, F. S. Roque, P. K. Donahoe, et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences*, 105(52):20870–20875, 2008.

[44] A. R. Laird, S. B. Eickhoff, P. M. Fox, A. M. Uecker, K. L. Ray, J. Saenz, J. J., et al. The BrainMap strategy for standardization, sharing, and meta-analysis of neuroimaging data. *BMC Res Notes*, 4:349, 2011.

[45] A. R. Laird, S. B. Eickhoff, F. Kurth, P. M. Fox, A. M. Uecker, J. A. Turner, et al. ALE meta-analysis workflows via the Brainmap Database: Progress towards a probabilistic functional brain atlas. *Front Neuroinform*, 3:23, 2009.

[46] A. R. Laird, J. L. Lancaster, and P. T. Fox. BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics*, 3(1):65–78, 2005.

[47] J. C. Lambert et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.*, 45(12):1452–1458, 2013.

[48] J. C. Lambert, B. Grenier-Boley, et al. Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis. *J Alzheimers Dis*, 20(4):1107–18, 2010.

[49] M. X. Li et al. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am. J. Hum. Genet.*, 88(3):283–293, 2011.

[50] F. Licastro et al. Sharing pathogenetic mechanisms between acute myocardial infarction and Alzheimer's disease as shown by partially overlapping of gene variant profiles. *J. Alzheimers Dis.*, 23(3):421–431, 2011.

[51] J. Z. Liu et al. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet*, 87(1):139–45, 2010.

[52] I. López González, P. Garcia-Esparcia, F. Llorens, and I. Ferrer. Genetic and transcriptomic profiles of inflammation in neurodegenerative diseases: Alzheimer, Parkinson, Creutzfeldt-Jakob and Tauopathies. *International Journal of Molecular Sciences*, 17(2):206, 2016.

[53] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, et al. Finding the missing heritability of complex diseases. *Nature*, 461:747, 2009.

[54] D. S. Marcus, M. P. Harms, A. Z. Snyder, M. Jenkinson, J. A. Wilson, M. F. Glasser, et al. Human connectome project informatics: Quality control, database services, and data visualization. *Neuroimage*, 80:202–219, 2013.

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

[56] A. Mishra and S. Macgregor. VEGAS2: Software for more flexible gene-based testing. *Twin Res Hum Genet*, 18(1):86–91, 2015.

[57] M. A. Mooney and B. Wilmot. Gene set analysis: A step-by-step guide. *American journal of medical genetics: Part B.*, 168(7):517–527, 2015.

[58] J. C. Morris. Early-stage and preclinical Alzheimer disease. *Alzheimer Disease & Associated Disorders*, 19(3):163–165, 2005.

[59] 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*, 49(3):390–398, 2008.

[60] D. Nam et al. GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Res.*, 38:W749–54, 2010.

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

[62] C. O'Dushlaine, E. Kenny, et al. Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Mol Psychiatry*, 16(3):286–92, 2011.

[63] K. Palmer et al. Predictors of progression from mild cognitive impairment to Alzheimer disease. *Neurology*, 68(19):1596–1602, 2007.

[64] 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*, 124(1):19–29, 2008.

[65] A. Peri and M. Serio. Neuroprotective effects of the Alzheimer's disease-related gene seladin-1. *J. Mol. Endocrinol.*, 41(5-6):251–261, 2008.

[66] R. B. Postuma, A. E. Lang, R. P. Munhoz, K. Charland, A. Pelletier, M. Moscovich, et al. Caffeine for treatment of parkinson disease: a randomized controlled trial. *Neurology*, 79(7):651–8, 2012.

[67] S. G. Potkin, G. Guffanti, A. Lakatos, J. A. Turner, F. Kruggel, J. H. Fallon, et al. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One*, 4(8), 08 2009.

[68] S. P. Poulin et al. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Res. Neuroimaging*, 194(1):7–13, 2011.

[69] S. Purcell et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–75, 2007.

[70] E. Pérez-Palma, B. I. Bustos, C. F. Villamán, M. A. Alarcón, M. E. Avila, G. D. Ugarte, et al. Overrepresentation of glutamate signaling in Alzheimer's disease: Network-based pathway enrichment using meta-analysis of genome-wide association studies. *PLoS ONE*, 9(4):1–16, 04 2014.

[71] V. Ramanan, S. Kim, et al. Genome-wide pathway analysis of memory impairment in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort implicates gene candidates, canonical pathways, and networks. *Brain Imaging Behav*, 6(4):634–48, 2012.

[72] V. Ramanan, L. Shen, et al. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet*, 28(7):323–32, 2012.

[73] V. K. Ramanan, S. L. Risacher, K. Nho, S. Kim, L. Shen, B. C. McDonald, et al. GWAS of longitudinal amyloid accumulation on 18F-florbetapir PET in Alzheimer's disease implicates microglial activation gene IL1RAP. *Brain*, 138(10):3076–3088, 2015.

[74] S. Realmuto et al. Tumor diagnosis preceding Alzheimer's disease onset: is there a link between cancer and Alzheimer's disease? *J. Alzheimers Dis.*, 31(1):177–182, 2012.

[75] S. Risacher, L. Shen, J. West, S. Kim, B. McDonald, L. Beckett, et al. Longitudinal MRI atrophy biomarkers: Relationship to conversion in the ADNI cohort. *Neurobiol Aging*, 31(8):1401–18, 2010.

[76] E. J. Rossin, K. Lage, S. Raychaudhuri, R. J. Xavier, D. Tatar, Y. Benita, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLOS Genetics*, 7:1–13, 01 2011.

[77] A. J. Saykin, L. Shen, et al. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. *Alzheimers Dement*, 6(3):265–73, 2010.

[78] A. J. Saykin, L. Shen, X. Yao, S. Kim, K. Nho, S. L. Risacher, et al. Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans. *Alzheimers Dement.*, 11(7):792–814, 2015.

[79] L. Shen et al. Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. *Brain Imaging Behav.*, 8(2):183–207, 2014.

[80] L. Shen, S. Kim, Y. Qi, M. Inlow, S. Swaminathan, K. Nho, et al. Identifying neuroimaging and proteomic biomarkers for MCI and AD via the elastic net. *Multimodal Brain Image Analysis*, 7012:27–34, 2011.

[81] L. Shen, S. Kim, S. L. Risacher, K. Nho, S. Swaminathan, J. D. West, et al. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *NeuroImage*, 53(3):1051 – 1063, 2010.

[82] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Stat. Comput.*, 14(3):199–222, 2004.

[83] R. Sokal and C. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.

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

[85] J. L. Stein, X. Hua, S. Lee, A. J. Ho, A. D. Leow, A. W. Toga, et al. Voxelwise genome-wide association study (vGWAS). *NeuroImage*, 53(3):1160–1174, 2010.

[86] J. L. Stein, X. Hua, J. H. Morra, S. Lee, D. P. Hibar, A. J. Ho, et al. Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease. *NeuroImage*, 51(2):542–554, 2010.

[87] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[88] X. Y. Tang et al. Shape abnormalities of subcortical and ventricular structures in mild cognitive impairment and Alzheimer's disease: detecting, quantifying, and predicting. *Hum. Brain Mapp.*, 35(8):3701–3725, 2014.

[89] J. A. Turner and A. R. Laird. The cognitive paradigm ontology: design and application. *Neuroinformatics*, 10(1):57–66, 2012.

[90] J. A. Turner, J. L. Mejino, J. F. Brinkley, L. T. Detwiler, H. J. Lee, M. E. Martone, et al. Application of neuroanatomical ontologies for neuroimaging data annotation. *Front Neuroinform*, 4, 2010.

[91] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[92] I. Ulitsky, A. Maron-Katz, et al. Expander: from expression microarrays to networks and functions. *Nature Protocols*, 5(2):303–322, 2010.

[93] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. J. Behrens, R. Bucholz, et al. The Human Connectome Project: A data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

[94] H. Wang, H. Huang, and C. Ding. Correlated protein function prediction via maximization of data-knowledge consistency. *RECOMB'14*.

[95] H. Wang, F. Nie, H. Huang, and F. Makedon. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. *Proceedings of 22rd International Joint Conference on Artificial Intelligence (IJCAI'11)*.

[96] L. L. Wang et al. PINBPA: Cytoscape app for network analysis of GWAS data. *Bioinformatics*, 31(2):262–264, 2015.

[97] Q. Wang, H. Yu, Z. Zhao, and P. Jia. EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression profiles. *Bioinformatics*, 31(15):2591–2594, 2015.

[98] J. Wu, T. Vallenius, K. Ovaska, J. Westermarck, T. P. Makela, and S. Hautaniemi. Integrated network analysis platform for protein-protein interactions. *Nat Meth*, 6(1):75–77, 2009.

[99] J. Yan, L. Du, et al. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. *Bioinformatics*, 30(17):i564–71, 2014.

[100] H. Zeng, E. H. Shen, et al. Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell*, 149(2):483–96, 2012.

CURRICULUM VITAE

Xiaohui Yao

Education

2013-2018 Ph.D in Bioinformatics, minor in Computer Science Indiana University

2009-2012 M.S. in Computer Science University of Science & Technology of China

2005-2009 B.S. in Computer Science Qingdao University

Honors and Awards

2017 1st Place Poster Award, IIGC, Irvine

2016 Travel Fellowship, ICIBM, Houston

2016 Travel Fellowship for AAIC, Indiana Univeristy

2015 Travel Fellowship, AAIC, Washington, DC

2015 Travel Fellowship, AIC, Washington, DC

2009-2012 Two Fellowships, University of Science and Technology of China

2005-2009 Three Fellowships, Qingdao University

Research and Training Experience

2013-2018 Advisor: Dr. Li Shen Indiana University School of Medicine

1. Network-based GWAS for identifying tissue-specific functional module
2. New paradigm of imaging genetic enrichment analysis (IGEA)
3. ADNI genetic database
4. Graphic mining of high-order drug interactions and directional effects using EMR
5. Scientific mapping and visualization of ADNI
6. Multivariate GWAS on amyloid imaging phenotypes

Peer-Reviewed Publications

1. **Yao X**, Yan J, Liu K, Kim S, Nho K, Risacher SL, Greene CS, Moore JH, Saykin AJ, Shen L. Tissue-specific network-based genome wide study of amygdala imaging phenotypes to identify functional interaction modules. *Bioinformatics*. 2017. 33(20):3250-3257.

2. **Yao X**, Yan J, Ginda M, Börner K, Saykin AJ, Shen L. Mapping longitudinal scientific progress, collaboration and impact of the Alzheimer's disease neuroimaging initiative. *PLoS ONE*. 2017. 12(11):e0186095.

3. **Yao X**, Yan J, Kim S, Nho K, Risacher SL, Inlow M, Moore JH, Saykin AJ, Shen L. Two-dimensional enrichment analysis for mining high-level imaging genetic associations. *Brain Informatics*. 2017. 4(1):27–37.

4. **Yao X**, Yan J, Risacher SL, Moore JH, Saykin AJ, Shen L. Network-based genome wide study of hippocampal phenotypes in Alzheimer's disease to identify functional interaction modules. *ICASSP'17: The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing*. Mar 5-9, 2017. New Orleans, LA.

5. Wang X, Yan J, **Yao X**, Kim S, Nho K, Risacher SL, Saykin AJ, Shen L, Huang H. Longitudinal genotype-phenotype association study via temporal structure auto-learning predictive model. *RECOMB'17: The 21st Annual International Conference on Research in Computational Molecular Biology*. May 3-7, 2017. Hong Kong.

6. Liu K, **Yao X**, Yan J, Chasioti D, Risacher SL, Nho K, Saykin AJ, Shen L. Transcriptome-guided imaging genetic analysis via a novel sparse CCA algorithm. *MICGen'17: MICCAI Workshop on Imaging Genetics*. September 10, 2017, Quebec

City, Canada.

7. Du L, Liu K, **Yao, X**, Yan J, Risacher SL, Han J, Guo L, Saykin AJ, Shen, L. Pattern discovery in brain imaging genetics via SCCA modeling with a generic non-convex penalty. *Scientific Reports*. 2017. 7:14052.

8. Hao X, Li C, Du L, **Yao X**, Yan J, Risacher SL, Saykin AJ, Shen L, Zhang D. Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease. *Scientific Reports*. 2017. 7:44272.

9. Hao X, Li C, Yan J, **Yao X**, Risacher SL, Saykin AJ, Shen L, Zhang D. Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis. *Bioinformatics (ISMB/ECCB 2017 Issue)*. 2017. 33(14):i341–i349.

10. Du L, Zhang T, Liu K, Yan J, **Yao X**, Risacher SL, Saykin AJ, Han J, Guo L, Shen L. Identifying associations between brain imaging phenotypes and genetic factors via a novel structured SCCA approach. *IPMI'17: Information Processing in Medical Imaging*. June 25-30, 2017. Boone, NC.

11. Hao X, **Yao X**, Yan J, Risacher SL, Saykin AJ, Zhang D, Shen L. Identifying multimodal intermediate phenotypes between genetic risk factors and disease status in Alzheimer's disease. *Neuroinformatics*. 2016. 14(4):439–52.

12. Hao X, Yan J, **Yao X**, Risacher SL, Saykin AJ, Zhang D, Shen L. Diagnosis-Guided Method for Identifying Multi-Modality Neuroimaging Biomarkers Associated with Genetic Risk Factors in Alzheimer's Disease. *PSB'16: Pacific Symposium on Biocomputing*. Jan 4-8, 2016. Big Island, Hawaii.

13. Saykin AJ, Shen L, **Yao X**, Kim S, Nho K, Risacher SL, Ramanan VK, Foroud

TM, Faber KM, Sarwar N, Munsie LM, Hu X, Soares HD, Potkin SG, Thompson PM, Kauwe JS, Kaddurah-Daouk R, Green RC, Toga AW, Weiner MW. Genetic Studies of Quantitative MCI and AD Phenotypes in ADNI: Progress, Opportunities, and Plans. *Alzheimers Dement.* 2015. 11(7):792–814.

14. **Yao X**, Yan J, Kim S, Nho K, Risacher SL, Inlow M, et al. Two-dimensional enrichment analysis for mining high-level imaging genetic associations. *BIH'15: International Conference on Brain Informatics & Health.* Aug 30-Sep 2, 2015. London, UK.

15. Liang H, Meng X, Chen F, Zhang Q, Yan J, **Yao X**, et al. A network-based framework for mining high-level imaging genetic associations. *MICGen'15: MICCAI Workshop on Imaging Genetics.* Oct 9, 2015. Munich, Germany.

Book Chapter

1. Yan J*, Du L*, **Yao X***, Shen L. Machine learning in brain imaging genomics. Wu G, Sabuncu M, editors. Elsevier Inc. 2016. [Equal contribution].

Conference Abstracts

1. **Yao X**, Yan J, Ginda M, Börner K, Kim S, Nho K et al. Genetic findings using ADNI multimodal quantitative phenotypes: A 2016 update. *AAIC'17: Alzheimer's Association Int. Conf. on Alzheimer's Disease*, London, UK.

2. **Yao X**, Yan J, Nho K, Risacher SH, Greene CS, Moore JH, et al. Identifying phenotype-relevant modules from a tissue-specific biological network: Application to an amygdala imaging genetics study. *NetSci'17: Int. School and Conf. on Network Science*, Indianapolis, IN. [**Platform talk**]

3. **Yao X**, Yan J, Risacher SL, Greene CS, Moore JH, Saykin AJ, Shen L. Network-based genome wide study of hippocampal imaging phenotype in Alzheimer's disease to identify functional interaction modules. *IIGC'17: 13th International Imaging Genetics Conference*. Irvine, CA. [**Best Poster Award, 1st Place**].

4. Chasioti D, **Yao X**, Zhang P, Ning X, Li L, Shen L. Mining directional drug interaction effects on myopathy using the FAERS database. *PSB'17: Pac Symp Biocomput.*, Big Island of Hawaii.

5. **Yao X**, Yan J, Ginda M, Börner K, Kim S, Nho K et al. The Growth and Impact of ADNI Genetics Publications as Measured by Science Mapping. *AAIC'16: Alzheimer's Association Int. Conf. on Alzheimer's Disease*, Toronto, CA.

6. **Yao X**, Kim S, Yan J, Risacher SL, Inlow M, Moore JH et al. Joint Analysis of Multiple Amyloid Imaging Phenotypes in a Genome Wide Association Study of the ADNI Cohort. *IIGC'15: 11th Internationasl Imaging Genetic Conference*, UC Irvin, CA.

7. **Yao X**, Yan J, Kim S, Nho K, Risacher SL, Börner K, et al. Genetic findings using ADNI multimodal quantitative phenotypes: A 2014 update. *AAIC'15: Alzheimer's Association Int. Conf. on Alzheimer's Disease*, Washington DC, USA.

8. **Yao X**, Chen R, Kim S, Yan J, Du L, Nho K et al. Genetic Findings using ADNI Multimodal Quantitative Phenotypes: A Review of Papers Published in 2013. *AAIC'14: Alzheimer's Association Int. Conf. on Alzheimer's Disease*, Copenhagen, Denmark.