

2020

Integrative multi-omic network strategies for unraveling complex disease biology and the identification of novel phenotype associated genes

<https://hdl.handle.net/2144/39630>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES
AND
COLLEGE OF ENGINEERING

Dissertation

**INTEGRATIVE MULTI-OMIC NETWORK STRATEGIES FOR
UNRAVELING COMPLEX DISEASE BIOLOGY AND THE
IDENTIFICATION OF NOVEL PHENOTYPE ASSOCIATED GENES**

by

DANIEL J. LANCOUR

B.S., Genetics, University of Wisconsin – Madison, 2013
B.S., Computer Science, University of Wisconsin – Madison, 2013

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2020

© 2020 by
Daniel J. Lancour
All rights reserved

Approved by

First Reader

Lindsay A. Farrer, Ph.D.
Distinguished Professor of Genetics
Professor of Medicine
Professor of Neurology
Professor of Ophthalmology
Boston University, School of Medicine

Professor of Biostatistics
Professor of Epidemiology
Boston University, School of Public Health

Second Reader

Simon Kasif, Ph.D.
Professor of Biomedical Engineering
Boston University, College of Engineering

Professor of Computer Science
Boston University, College of Arts and Sciences

Dedicated to the memory of my sister, Rachel

**INTEGRATIVE MULTI-OMIC NETWORK STRATEGIES FOR UNRAVELING
COMPLEX DISEASE BIOLOGY AND THE IDENTIFICATION OF NOVEL
PHENOTYPE ASSOCIATED GENES**

DANIEL LANCOUR

Boston University, Graduate School of Arts and Sciences and College of
Engineering, 2020

Major Professor: Lindsay Farrer, Professor of Medicine, Neurology,
Ophthalmology, Epidemiology, and Biostatistics

ABSTRACT

Identifying the genetic risk factors underlying a given disease is an essential step for informing effective drug targets, understanding disease architecture, and predicting at-risk individuals. A commonly applied approach for identifying novel disease-associated genes is the Genome Wide Association Study (GWAS) approach, in which a high number of individuals are sequenced and genetic variants are then tested for an association with disease status. While the GWAS approach has identified countless disease-associated genes, there remain plenty of diseases for which our genetic understanding is still incomplete. One strategy for augmenting the GWAS approach is to incorporate additional omics data in order to prioritize biologically plausible candidate genes.

In this thesis work, we integrate network-based strategies with existing genetic analysis pipelines in order to identify novel Alzheimer's disease (AD) genes. Two types of biological data inform the underlying structure of the networks: a) protein-protein interactions and b) gene expression in the human brain. Genes which interact or are co-expressed across similar conditions have been shown to have a higher probability of being functionally related. Using a set of previously known AD genes, we apply a network propagation strategy to score genes based upon their proximity to the known AD genes within these networks. Then we integrate the network score of each gene with its risk score from GWAS to identify novel candidates. To further affirm the reproducibility of findings, we further incorporate additional information in the form of knockout models in flies, bootstrap aggregation, and external genetic datasets. In addition to predicting novel genes, we are able to utilize regional co-expression networks to further understand how the known AD genes behave within the various sub-divisions of the brain. We find that regions of the brain which are known to have the earliest vulnerability to AD-induced neurodegeneration also tend to be where AD genes are highly correlated.

Table of Contents

Chapter 1: Introduction and Rationale	1
Chapter 2: Integration of Network-Based Diffusion and Studies of Genetic Interaction Leads to Consistent Biological Insights of Alzheimer’s Disease and Novel AD Candidate Genes	4
Methods	10
Results	17
Discussion.....	32
Chapter 3: Analysis of Brain Region-Specific Co-Expression Networks Reveals Clustering Properties of Genes Associated with Alzheimer Disease Risk and Identifies Novel Risk Gene Candidates	41
Methods	43
Results	48
Discussion.....	64
Chapter 4: Analysis of Median Ranking by Diffusion of General Phenotype Sets and the Effects of Cell Type and Disease on the Clustering Properties of Alzheimer Genes	70
Methods	71
Results	75
Discussion.....	83
Chapter 5: Conclusions and Future Projects	87
References	93
Curriculum Vitae.....	112

List of Tables

Table 2.1. RAD Genes and the Type of Study that Identified Them	9
Table 2.2. Proximity Between RAD Genes in PPI Network.	19
Table 2.3. Proximity of Non-RAD Hub Genes to RAD Genes.	20
Table 2.4. Top Predicted AD Genes Using Combination Approach.	29
Table 2.5. GSEA Results After Ranking Genes by Combined Z-Scores	31
Table 2.6. GSEA Results After Ranking Genes by GWAS Only Z-Scores.	31
Table 3.1. RC of RAD Genes in the Cerebrum, Cerebellum, and Brain Stem....	53
Table 3.2: RC of RAD Genes in Late and Early Stage Correlation Networks.....	56
Table 3.3. Combined Z-Scores Reveal Novel AD Gene Candidates.....	62

List of Figures

Figure 2.1: Summary of Analysis Steps.....	8
Figure 2.2: Filtering on Network Score Improves Replication Rate.	22
Figure 2.3: Comparison of GWAS and Network Z-Scores.....	25
Figure 2.4: Support Vector Machine Training to Predict GWAS and Network Z- Score Weights.	27
Figure 3.1: Principal Component Analysis Indicates Clustering by High Level Structure.....	50
Figure 3.2: RAD Genes Have Region-Specific Expression Levels.....	51
Figure 3.3: The RAD Gene Set Has Consistently High MRC in Each Individual.	58
Figure 3.4: Rankings Using RAD Genes Are Consistent in Each Individual Brain.	60
Figure 4.2: Overview of MRDs of Important AD and Survival Phenotypes.	76
Figure 4.3: Cell Type Compositions of the Early Stage Regions:	79
Figure 4.4: Cell Type Compositions of the Late Stage Regions.	80
Figure 4.5: The Effects of Age and AD on the MRC of the RAD Genes.....	82
Figure 5.1: Overview of Possible Combinations of Protein Interactions and Co- Expression.....	89

Abbreviations

Note: All *italicized* acronyms in the text are the names of genes

ABA = Allen Brain Atlas
AD = Alzheimer's disease
ADGC = Alzheimer's disease Genetic Consortium
ASP = Average Shortest Path
BP = Biological Process
BS = Brain Stem
CB = Cerebellum
CGS = Candidate Gene Study
CX = Cerebrum
DA = Defective Aging
DIOPT = DRSC Integrative Ortholog Prediction Tool
DM = Defective Memory
EOAD = Early Onset Alzheimer's disease
FDR = False Discovery Rate
FWER = Family Wise Error Rate
GSEA = Gene Set Enrichment Analysis
GTEx = Gene Tissue Expression Portal
GWAS = Genome Wide Association Study
HD = Huntington's disease
HGNC = Human Gene Nomenclature
HST = High-Level Structure
KEGG = Kyoto Encyclopedia of Genes and Genomes
LOAD = Late Onset Alzheimer's disease
LOD = Logarithm of the Odds
MST = Mid-Level Structure
MRC = Median Ranking by Correlation
MRD = Median Ranking by Diffusion
ND = Neurodegenerative disease
NIA = National Institute of Aging
OS = Oxidative Stress
PA = Premature Aging
PCA = Principal Component Analysis
PD = Parkinson's disease
PPI = Protein-Protein Interaction
QC = Quality Control
RAD = Reproducible Alzheimer's disease (genes)

RC = Ranking by Correlation
SNP = Single Nucleotide Polymorphism
SVM = Support Vector Machine
WES = Whole Exome Sequencing

Chapter 1: Introduction and Rationale

Genetic studies have achieved a strong understanding of the underlying roots of Mendelian diseases such as Huntington's, Tay-Sachs, sickle cell, and several others [1-3]. However, diseases following a multigenic pattern of inheritance have required a vast number of studies to achieve even a partial understanding of the genetic roots of the diseases [4, 5]. For example, mutations in approximately 60 genes have been reliably identified as contributing to genetic risk for Alzheimer's disease (AD) (See Table 2.1). Despite the high volume of known AD-associated genes, estimates of the missing heritability of AD and other complex diseases are sizeable [4, 6]. Datasets with higher sample sizes and greater statistical power may eventually contribute to identifying remaining unknown AD-associated genes, but alternatively other forms of readily available omics data can also contribute to addressing this issue [4].

A challenge to incorporating multiple forms of biological data into combined analyses is identifying a framework that can suitably model all the various relationships and information contained in each type of data. One flexible framework is a graph, otherwise known as a network, which is a well-studied concept in computer science for representing similarities between entities and numerous other relationships [7-11]. In a graph, each entity is represented by a node and then nodes are connected by edges if a relationship between the two

nodes exists. Weights representing the strengths of these relationships can be appended to edges, and labels can be attached to nodes, allowing for a wide variety of algorithms to be applied. Graphs have been used to optimize commonly used resources such as internet search engines, social networks, economic models, public transportation routes, and many others [12, 13]. In the context of the systems biology research, network approaches have been applied to a wide variety of biological concepts such as predicting gene function, characterizing bacterial communities, generating novel drug targets, and identifying gene clusters associated with diseases [14-18]. The data underlying the networks in these approaches is commonly predicated on either protein-protein interaction (PPI) or gene expression levels in specific biological conditions. Proteins which interact have a higher probability of having common pathways, a property referred to as functional linkage. Likewise, genes which are co-expressed in particular conditions are more likely to be co-regulated or share a functional role in those environments. These two concepts serve as the basis for the approaches described in this work.

Many studies have applied network methodology to accommodate multi-omic analysis with success across a wide variety of diseases and phenotypes [19-25]. Gene expression has been integrated with a PPI network to identify new type 2 diabetes genes [26]. Gene sets and label propagation in a PPI network have been used to determine similarities between neurodegenerative diseases [27].

The work in the following chapters outlines another form of multi-omic integration that integrates previously known disease-related genes and Genome Wide Association Studies (GWAS) with network-based propagation in order to further characterize Alzheimer's disease. As a result of this work, we identified several novel candidate AD genes and provide novel insights into how the expression of AD genes may contribute to the regional patterning of AD-based pathology in the brain.

Chapter 2: Integration of Network-Based Diffusion and Studies of Genetic Interaction Leads to Consistent Biological Insights of Alzheimer’s Disease and Novel AD Candidate Genes

The discovery of disease-associated genomic variation has numerous clinical and scientific applications, including earlier disease prognosis, improved understanding of disease pathophysiology, and development of personalized treatment therapies [28]. A commonly used technique for identifying these mutations is the genome wide association study (GWAS) approach [29].

Typically, a large sample of affected and unaffected individuals are genotyped for many single nucleotide polymorphisms (SNPs) using a high-density microarray chip and then test statistically if the allele frequency of each variant is associated with disease status [29]. Significant associations in this first step (“discovery phase”) are deemed to be robust if they replicate in an independent cohort (“replication phase”). In this study, we focused on improving the replicability of GWAS results for Alzheimer disease (AD), although our methodology is applicable to genetic data for other diseases and traits. AD is a neurodegenerative disease resulting in irreversible dementia and memory loss with elevated prevalence in older populations [30]. Recent estimates suggest that approximately 5.4 million Americans have AD, and the number of cases of AD is expected to increase dramatically in future years if medical advances continue to

improve life expectancy, thereby allowing more individuals to reach ages where AD is on the rise [30].

Genetic studies of AD have led to identifying numerous AD associated genes such as *APP* [31], *PSEN1* [32], and *PSEN2* [33] for early onset AD (EOAD), as well as *APOE* [34, 35] and *SORL1* [35, 36] for late onset AD (LOAD). Common variants in more than 20 other genes have been robustly associated with AD risk [35]. However, not all AD associated genes will reach genome wide significance in current datasets of sample sizes below 100,000 individuals. It is well recognized that incorporating other forms of biological data improves confidence in genetic findings [37-39].

Our computational framework is based on the following biological hypothesis. If a known AD variant is associated with a gene that is involved in a particular biological process (BP) (e.g. inflammation), we assume as a probabilistic prior that other AD variants might be associated with proteins involved in this BP or proteins that physically interact with this BP. This hypothesis can be tested computationally using a protein interaction network [40-42] by extending the “guilt by association” principle via propagation of probabilistic evidence in a network [43, 44]. This general idea has similarity to the Google ranking algorithm of web pages, in which a web page that has a short link distance to many “important” pages will itself be considered “important.”

In the case of protein interactions, guilt by association-based inference is typically performed by inspecting the function of direct neighbors of a predicted disease gene in a protein-interaction network. This approach has been incorporated in multiple interpretation systems as well as commercially such as Ingenuity Pathway Analysis (IPA). However, it has been shown that network propagation, diffusion or other related methods that go beyond simple neighbor-based analysis can carry functional or disease associations further in the network with improved predictive accuracies [37, 38]. This idea extends to predicting both gene function and disease phenotypes associated with genes [26, 27, 38, 45-47].

We hypothesize that this general framework, and network diffusion in particular, can be extended to aid prioritization of AD genes. Although the underlying biology of AD may be far more diverse than a single function, there are several biological pathways that are aberrantly activated in AD brains, and not surprisingly, most of the genes identified by AD GWAS contribute to these pathways [48]. For example, a primary indicator of AD is the accumulation of amyloid beta plaques in the brain, resulting from mis-processing of *APP* protein [48].

We developed a novel re-prioritization approach that can be integrated easily into the current genetic analysis design (**Figure 2.1**). First, we curated the AD literature to produce a set of approximately 60 robust AD (RAD) genes that

includes those that have been associated with AD at the genome-wide significance level or that contain variants shown to affect AD-related processes directly (**Table 2.1**). We then constructed a network of protein-protein interactions and applied network diffusion to score and rank genes based on their proximity to the RAD genes. Network diffusion allows modeling of indirect interactions, modules and protein complexes that are not modeled if only the direct interactions of proteins are considered. Next, we combined our genetic association results with the network diffusion scores to produce a newly re-prioritized ranking of genes. Finally, we validated our methodology using a novel approach involving bootstrap aggregation on one of the largest assembled genetic datasets of AD. Network-augmented genetic results have measurably improved replication rates in this validation approach. We also show that our main results and key predictions were essentially unchanged after restricting the RAD set to 19 genes which have had been functionally validated as well as replicated in independent datasets.

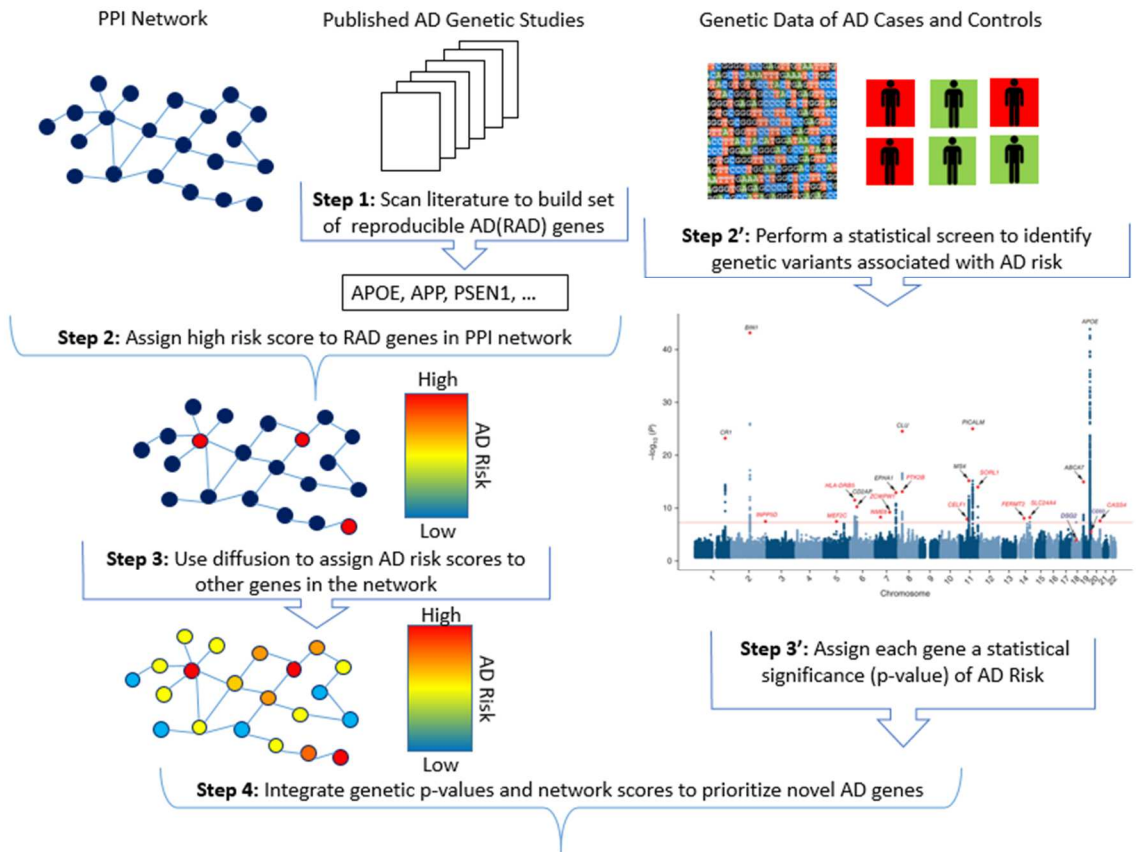


Figure 2.1: Summary of Analysis Steps. A set of AD genes that are reproducible (RAD genes) across different genetic studies was assembled through literature curation. The RAD genes were assigned a high initial risk score, and graph theoretical diffusion was employed to derive network diffusion scores for the rest of the genes in the network. Scores obtained from genetic screens and network diffusion were integrated to derive a new prioritization.

Table 2.1. RAD Genes and the Type of Study that Identified Them

Chr.	Gene	Evidence	Chr	Gene	Evidence	Chr	Gene	Evidence
1	CR1	GWAS – AD [35, 49]	7	<i>ZCWPW1</i>	GWAS – AD [35]	12	<i>SRRM4</i>	GWAS – endo [50]
1	PSEN2	Linkage [51]	7	EPHA1	GWAS – AD [52]	13	<i>SLCA10A2</i>	GWAS – AD [35, 53]
2	BIN1	GWAS – AD [35]	7	<i>PLXNA4</i>	GWAS – AD [54]	14	<i>FERMT2</i>	GWAS – endo. [35]
2	<i>INPP5D</i>	GWAS – AD [35]	8	<i>PTK2B</i>	GWAS – AD [35]	14	PSEN1	Linkage [51]
2	<i>CASP8</i>	WES [55]	8	CLU	GWAS – AD [49]	14	<i>SLC2A4A</i>	GWAS – AD [35]
3	<i>KCNMB2</i>	GWAS – endo [56]	8	<i>TP53INP1</i>	GWAS – AD [57]	14	<i>PLD4</i>	GWAS – endo. [58]
3	<i>OSTN</i>	GWAS – endo [59]	8	<i>PDGFRL</i>	GWAS – endo [60, 61]	15	<i>TRIP4</i>	GWAS – AD [62]
4	<i>UNC5C</i>	WES [63]	9	<i>LMX1B</i>	GWAS – endo [64]	16	PLCG2	GWAS – AD [65]
4	<i>GALNT7</i>	GWAS – endo [56]	9	<i>MVB12B</i>	GWAS – endo	17	MAPT	GWAS – AD [66]
5	<i>MEF2C</i>	GWAS – AD [35]	10	<i>ECHDC3</i>	GWAS – AD [61]	17	<i>KANSL1</i>	GWAS – AD [66]
5	<i>SORCS2</i>	CGS [36]	10	<i>SORCS1</i>	CGS [36]	17	ABI3	GWAS – AD [65]
5	<i>PFDN1</i>	GWAS – AD [61]	10	<i>SORCS3</i>	CGS [36]	17	<i>ACE</i>	CGS [67]
6	<i>HLA-DRB5</i>	GWAS – AD [35]	11	CELFB1	GWAS – AD [35]	19	ABCA7	GWAS – AD [35]
6	TREM2	WES [68]	11	SPI1	GWAS – AD [69]	19	<i>PLD3</i>	WES [70]
6	<i>NCR2</i>	GWAS – endo [59]	11	<i>MS4A6A</i>	GWAS – AD [35]	19	APOE	Linkage [71]
6	CD2AP	GWAS – AD [52]	11	<i>MS4A4A</i>	GWAS – AD [35]	19	<i>CD33</i>	GWAS – AD [35, 52]
6	<i>TPBG</i>	GWAS – AD [61]	11	<i>MSA6</i>	GWAS – AD [35]	20	<i>CASS4</i>	GWAS – AD [35]
7	<i>COBL</i>	GWAS – AD [53]	11	PICALM	GWAS – AD [49]	21	APP	Targeted Seq. [51]
7	AKAP9	WES [72]	11	SORL1	CGS [35, 36]	21	<i>ABCG1</i>	GWAS – endo. [56]
7	<i>PILRA</i>	GWAS – AD [73]	11	<i>C1QTNF4</i>	GWAS – endo [50]			

Methods

Assembling an AD Gene List

A set of genes ascribed to AD with a high degree of certainty was assembled through curation of published findings ascertained through PubMed searches that emerged from studies using a variety of approaches including GWAS of AD risk and AD-related endophenotypes, family-based linkage analysis, positional cloning, whole exome sequencing (WES), and candidate gene testing (CGS) (Table 2.1). Criteria for inclusion in this set included (1) genome-wide significance for GWAS and WES studies ($p < 5 \times 10^{-8}$) and LOD score > 3 for linkage studies and (2) replication of association signals in independent datasets; or (3) biological evidence that demonstrate functional relevance to AD of associated variants or the encoded protein.

Harmonizing Protein-Protein Interaction Databases

A set of interacting gene-gene pairs (in HGNC symbol format) is required as input for this software. To compile this set, three databases (RefIndex v14 [41], ConsensusPathDB v31 [40], and Human Interactome Y2H DB vHI-II-14 [42]) were selected based on their demonstrated utility in recent work [27]. iREFINDEX and ConsensusPathDB interactions were filtered to remove self and complex (more than two proteins) interactions. The ConsensusPathDB interactions are given in uniProt ID format, which were converted to HGNC symbols using the official website (<http://www.genenames.org>). iREFINDEX

provides a HGNC symbol for each interactor of an interaction when possible, and so only interactions which had a HGNC for both interactors were kept. The Human Interactome DB already provides a set of binary gene-gene interactions in HGNC format, so no processing was required. The union of the processed sets from each database was used as the final interaction set. The unified set contains 19,972 unique gene symbols and 236,642 interactions. These databases are curated collections of experimentally determined interactions (typically binding or affinity) reported in the literature, such as from co-immunoprecipitation, as well as predicted interactions in a small number of databases.

Assigning Network Scores to Genes Through Diffusion

Network diffusion is a very well-studied spectral approach to graph clustering and annotation [7, 44, 74, 75]. It attempts to mimic node-to-node distance in the graph that in turn aims to capture functional relevance. The first step of the diffusion method is to model the protein interactions as a network. A network is comprised of a set of nodes, V , and a set of edges between nodes, E . For this work, nodes represent genes, and edges represent an interaction present in the unified set. Although we use unweighted edges in this work, our network methods and software are able to receive weighted input as well, such as protein interactions with confidence measures taken from STRING [76]. The construction of diffusion kernels using weighted edges has been well studied and is equally

valid [74]. n is the number of nodes in the network, which is 19,972 (yielding 236,642 edges). All network methods were implemented in R. The regularized Laplacian kernel [74] is constructed by:

$$K = (I + \alpha L)^{-1}$$

where K is the resulting kernel, I is the identity matrix, L is the graph Laplacian, and alpha is a constant ([74] for additional details). For this study, an alpha value of 0.1 was used, consistent with other work in this field [44]. Next, a network diffusion score was computed for each gene. To do this, the diffusion score vector, y , was initialized to be a length n vector that contains 1's in the indices of the RAD genes, and 0's otherwise. Risk scores for all genes in the graph were then derived by multiplication of K by the diffusion score vector y : $\tilde{y} = Ky$

Validation of Diffusion Using a Leave-One-Out Approach

To test if RAD genes had closer than random diffusion proximity to other RAD genes in a network, leave-one-out cross validation [77] was applied to the RAD gene set. First, a single RAD gene from the RAD set was set to 0 in the initial diffusion score vector, y . Then, diffusion scores were computed based upon this new initialization of y . The diffusion scores were sorted and the sorted rank of the removed RAD gene's diffusion score was determined in comparison to all other non-RAD genes. This process was repeated for each gene in the RAD set, resulting in a list of ranks. If diffusion proximity is informative and potentially predictive, the average rank of the RAD genes should be significantly lower than

the average rank of all genes, $(n+1) / 2$, which was verified using a one-tailed t-test.

ADGC GWAS Dataset

The Alzheimer's Disease Genetics Consortium (ADGC) is an NIA-funded project whose goal is to identify genes associated with an increased risk of developing late-onset Alzheimer disease (LOAD) by assembling and analyzing genetic and phenotypic data from large cohorts containing rigorously evaluated AD cases and cognitively normal controls of various ethnic ancestries. Details of ascertainment, collection, quality control (QC), and analysis of genotype and phenotype data in the individual datasets of the ADGC are provided elsewhere [35, 78]. Here we examined genotype data that were generated using high-density SNP microarrays from 32 prospective, case-control, and family-based studies of LOAD comprising 16,175 case and 17,176 controls of European ancestry. After QC steps to filter low-quality SNPs and individuals with low genotype call rates, principal components (PCs) of ancestry were computed within each dataset using EIGENSTRAT [79] and a set of 21,109 SNPs common to all genotyping platforms and datasets in order to account for population substructure in genetic association analysis. Samples with outlier PC values >six standard deviations from the mean were excluded from subsequent analyses. Genotypes for a much larger set of SNPs were imputed using the Haplotype Reference Consortium panel release 1.1 [80, 81], which includes 64,976

haplotypes derived from 39,235,157 SNPs, and the Michigan Imputation Server (<https://imputationserver.sph.umich.edu/>) running MiniMac3 [82, 83].

Genome-wide Association Analysis

Association of AD with the imputed dosage of the minor allele for each SNP (a quantitative estimate between 0 and 2) genome-wide was conducted using logistic regression models implemented in PLINK [84] that included covariates for age-at-onset/age-at-exam, sex, the first three PCs, and an indicator variable for each dataset. Joint analysis was chosen in favor of meta-analysis to avoid problems that could be introduced if bootstrap aggregation under-sampled small cohorts, resulting in unreliable association estimates for those cohorts. To account for relatedness in family datasets, subsets of maximally-unrelated affected and unaffected individuals were sampled from each pedigree. Each variant was annotated to a gene region according to RefSeq release 69 [85] using the program ANNOVAR [86]. Then, each gene was assigned the minimum p-value of all variants annotated to it, after applying the following formula:

$$P_g^{Gene'} = 1 - (1 - P_g^{BestSNP})^{\frac{N+1}{2}}$$

where N is the number of variants analyzed that were annotated to the gene.

Previously, this correction [87] has been shown to perform comparably to more complex adjustments based upon gene length, recombination hotspots, and similar gene features [88].

Validation of Genetic Re-Prioritization Through Bootstrap Aggregation

Since the availability of large AD genetic datasets is limited, bootstrap aggregation [89] was used to generate a high number of datasets for method validation. First, the full ADGC dataset was equally separated into discovery and replication halves. Then, 25 iterations of bootstrap aggregation were applied to the discovery half and then the replication half. The resultant 25 discovery and 25 replication datasets were then matched (D1 and R1, D2 and R2...D25 and R25). To further ensure robustness, the splitting procedure was repeated a total of 5 times, with 25 iterations of bootstrap aggregation applied each time, resulting in 125 total pairings (D1 and R1, D2 and R2....D125 and R125). Each pairing represents a discovery dataset as well as an independent replication dataset.

For each pairing, the previously described genetic analysis was conducted on the discovery half. Then all genes that passed a designated significance threshold (the number of passing genes is denoted as r) were selected to be tested again in the replication half using a significance threshold of $(0.05 / r)$. The replication rate was computed by determining the percentage of passing genes in the discovery half that also passed in the replication half. A replication rate was estimated for each pairing, and the mean replication rate was then determined. Next, the replication rate was re-determined for each pairing, with the added criterion that selected genes must also have a top percentile network diffusion score (top 10th, 20th, 30th, 40th, and 50th were tested). The average replication

rate for each filtering threshold was compared to the average replication rate without filtering.

Integrating GWAS and Network Diffusion Scores

The p-values from genetic analysis of the ADGC dataset were converted to Z-scores using the `qnorm` function in R. Then, the network diffusion scores were converted into percentiles. The percentiles are transformed into Z-scores using the `qnorm` function, with the additional specification of `lower.tail=F`. The weighting scheme from METAL was applied to combine the GWAS and network Z-scores:

$$Z_{combined} = \frac{w_1 * Z_{gwas} + w_2 * Z_{network}}{\sqrt{w_1^2 + w_2^2}}$$

Although any weight selection can be used, the weights were “learned” using an SVM [90] due to the observation that the GWAS and network scores did not contribute equally to predicting replication rate. First, a replication rate was determined for each gene. If a gene had a p-value of <0.05 in d discovery datasets and a replication p-value of <0.05 in r of the paired replication datasets, it was assigned a replication rate of r/d . To reduce model overfitting, create sufficient separation between the classes, and achieve a balance of high and low replicating genes, only high replication genes (≥ 0.7 , $n = 676$) and low replication genes (< 0.1 , $n = 475$) representing approximately 8.4% of the total genes with both a network and GWAS scores were extracted. By comparison, using a threshold of 0.8 or 0.9 would result in an imbalanced training set with very few

high replication genes because highly replicating genes are uncommon. A linear SVM [90] was trained using the network Z-scores and the genetic association Z-scores as features, and “high” and “low” as the classes. The resulting slope of decision boundary was then used to determine appropriate weights ($w_1 = 0.703$, $w_2 = 0.297$).

Pathway Analysis Using the Re-Prioritized Ordering of Genes

Pathway enrichment was performed using the Gene Set Enrichment Analysis (GSEA) software [91]. GSEA’s pre-ranked analysis tool requires that the user provide a numeric measure for ordering genes. To establish a baseline, enrichment was done using our internal GWAS Z-scores to order genes. Then, enrichment was done using the alternative ordering genes based upon their combined Z-scores (see above for combination method). The gene sets tested for enrichment were the GSEA C2 pathways in MSigDb, which are the “curated gene sets” compiled from multiple sources including KEGG [92], Reactome [93], and domain experts. The significance threshold was set at $FDR < 0.25$, as suggested previously for this hypothesis generating approach [91].

Results

RAD Genes Are Proximal in a PPI Network

We assembled a PPI network using interactions pooled from multiple PPI databases (ConsensusPathDB [40], iRefIndex [41], and Human Interactome Y2H

[42]) inspired by recent work [27] . Pooling interactions from these three databases resulted in a connected network that includes a large percentage of the genes in our GWAS dataset. We then determined if the RAD genes are proximal within this network. The first proximity measure tested was the average shortest path (ASP) distance [94]. The ASP distance between RAD genes, determined by a leave one out strategy (See Methods), is much smaller than would be expected by random chance (**Table 2.2**). One problem is that ASP distance between RAD genes and genes with many interactions (the number of interactions a gene has corresponds to its “degree” and high degree genes are considered to be hubs) tends to be small (**Table 2.3**). In this situation, all hub genes will be falsely predicted to be AD-related. Thus, we incorporated instead the Regularized Laplacian diffusion kernel [74] which penalizes paths going through hubs. The diffusion distance between RAD genes is smaller than would be expected by chance ($p = 0.00054$) (**Table 2.2**). Simultaneously, the problematic hub genes in the network have discounted scores as demonstrated by the notable drop in ranking of the 10 genes with the highest number of overall interactions (**Table 2.3**).

Table 2.2. Proximity Between RAD Genes in PPI Network. Each RAD gene was ranked (in comparison to the other 19,972 genes in the network) based upon its degree (number of interactions in network), its ASP distance to the RAD genes, and total diffusion distance from the RAD genes. The average ranking of the RAD genes was 7,949 using ASP (60th percentile, t-test $p = 0.015$) and 6,959 for diffusion (65th percentile, t-test $p = 0.00054$).

Gene	Rank			Gene	Rank			Gene	Rank		
	Degree	ASP	Diffusion		Degree	ASP	Diffusion		Degree	ASP	Diffusion
APP	2	2	1248	MEF2C	3012.5	3072.5	2619	SORCS2	12984	14902.5	1081
CASP8	238.5	76	754	ABI3	3012.5	10739	3228	SORCS3	14153	16106.5	1170
PSEN1	558.5	119.5	441	SORL1	4372.5	9964	2675	ABCG1	14153	7689.5	16627
MAPT	600.5	9	342	TPBG	4516.5	4551.5	5100	TP53INP1	14153	11727	10975
PTK2B	800	175	670	PDGFRL	4862	13192.5	7434	PLXNA4	14153	15296.5	14933
CLU	883	785	1935	LMX1B	5236.5	10441.5	7905	KCNMB2	15703.5	11038.5	12216
PFDN1	930.5	2268	4465	HLA-DRB5	5666.5	4554	7104	SORCS1	15703.5	17153	9425
CD2AP	1043.5	2275.5	585	CD33	5666.5	2281.5	1682	MS4A6A	15703.5	19883.5	19955
PSEN2	1188	454	642	PLD3	5891.5	4554.5	4320	ABCA7	15703.5	7689.5	17609
AKAP9	1230	4547.5	2996	CELF1	5891.5	789	3793	SRRM4	18290	18462.5	18934.5
PLCG2	1255	281	868	PILRA	6640.5	13274.5	8762	CASS4	18290	14847.5	16647.5
APOE	1517	283	626	CR1	7296.5	15652	12460	ECHDC3	18290	19700.5	19390
INPP5D	1582	455	795	GALNT7	7296.5	7688	8782	PLD4	18290	7689.5	17433
BIN1	1691	457	977	MVB12B	7995.5	7688.5	4498	TREM2	18290	19587	1566
TRIP4	2509	4548.5	5679	ACE	8878	4555	9212	SLC10A2	18290	7689.5	17128
PICALM	2640	3070.5	1207	EPHA1	9380.5	7689	8437	ZNF804B	18290	18465	18406
KANSL1	2780	3069.5	3734	COBL	9928.5	13930	9416	NCR2	18290	19587	1566
FERMT2	2857.5	1496.5	3313	UNC5C	12984	14796.5	15064				

Table 2.3. Proximity of Non-RAD Hub Genes to RAD Genes. The resulting ranking each hub (high degree) gene received when propagating from the RAD genes using both ASP and then Diffusion was determined. Hub genes tended to have top 50 rankings under ASP, whereas diffusion more appropriately penalized the high degree of the hubs to allow lower degree genes to rank higher.

Gene	Degree	Rank	
		ASP	Diffusion
UBC	1	1	1433
SUMO2	2	20.5	1570
CUL3	3	51	2515
SUMO1	4	20.5	1502
EGFR	5.5	3	937
TP53	5.5	7	983
GRB2	7	2	905
SUMO3	8	181	2433
HSP90AA1	9	10	978
MDM2	10	51	1096

Filtering by Network Diffusion Score Improves Replication Rate

We next tested if genes with high diffusion scores replicate more frequently in order to demonstrate that diffusion scores are informative when used in conjunction with genetic data. Bootstrap aggregation [89] was applied to our genetic dataset to produce a large number of pairs of discovery and replication datasets (See Methods). In each discovery + replication pair, we conducted a standard genetic workflow, beginning with a screen in the discovery dataset followed by validating top findings in the replication dataset. For each pair, a replication rate was calculated by determining the percentage of genes that surpass a given significance threshold also replicated. To test if network diffusion scores improved replication, we altered the standard discover + replication approach. We ranked genes by their network diffusion score and then iteratively dropped genes that had ranking diffusion scores below a given stringency threshold. At first we retained only genes in the 50th percentile of network scores, then gradually increased the threshold to only include genes in the 60th, 70th, 80th, and 90th percentiles. For each threshold, we computed the replication rate and compared to the baseline. As shown in **Figure 2.2**, filtering based upon network score percentile noticeably increased replication rate. Genes with a $-\log(p\text{-value})$ of > 6 replicated at a rate of approximately 16% in simulations (farthest right purple point), while additional strict network filtering improved the replication rate to nearly 34% (farthest right red point).

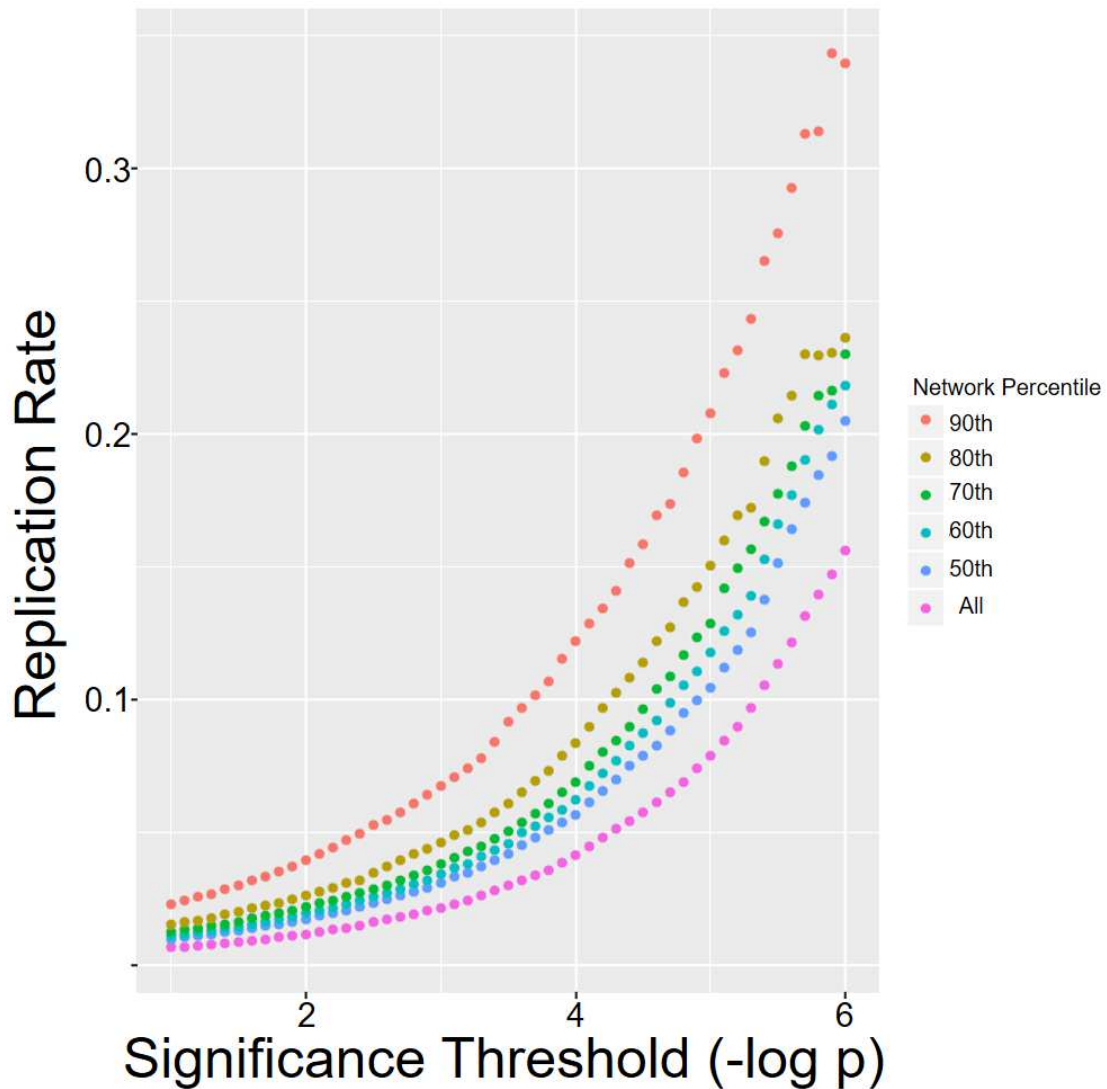


Figure 2.2: Filtering on Network Score Improves Replication Rate. The replication rate was computed for all genes surpassing the significance threshold for each GWAS. This procedure was repeated in each bootstrapped dataset and the average replication rate was determined (purple). This process was repeated using increasingly strict filters on the network diffusion scores. The baseline

replication rate without utilizing network scores (naïve method) is represented by the purple points. The strictest network filter (red) has a consistently higher replication rate than the naïve method.

Combined Z-Scores Predict Novel AD Genes

Since filtering on network diffusion score improved replication rate, we next sought to integrate the network diffusion scores and genetic results into a single score. First, we converted the p-value of each gene from genetic analysis into a Z-score (“GWAS Z-Scores”) and then converted the network diffusion percentile of each gene into a Z-score (“Network Z-scores”). Linear regression analysis showed that the Network and GWAS Z-scores are independent (**Figure 2.3A**). Next, we assigned each gene a replication rate based upon how frequently the gene replicated in our bootstrapped validation datasets (See Methods). We observed that replication rates were higher for genes with higher network Z-scores compared to genes with lower network Z-scores (**Figure 2.3B**).

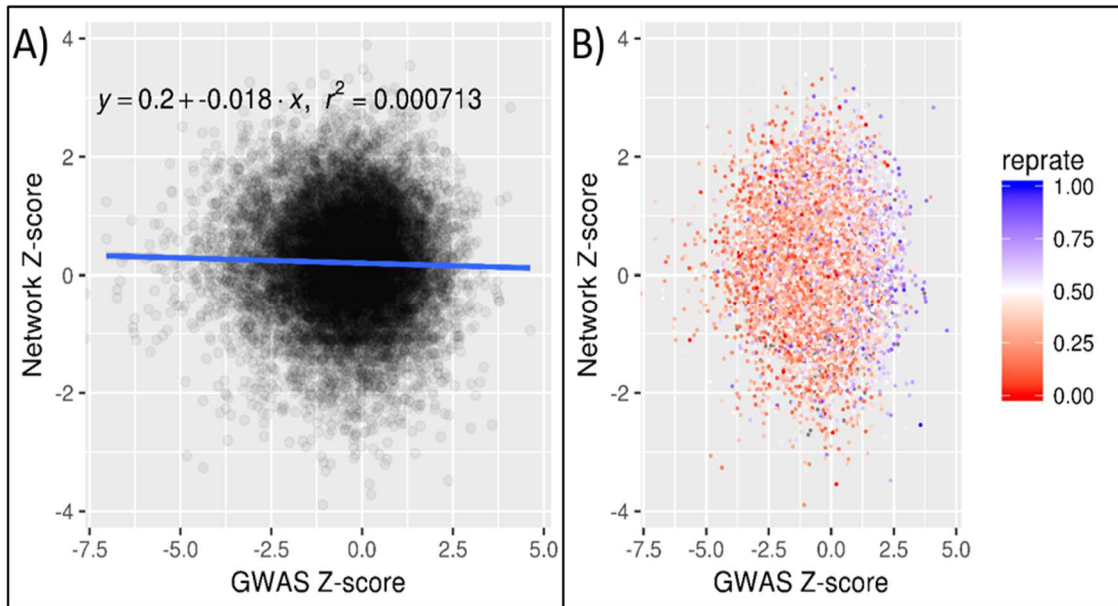


Figure 2.3: Comparison of GWAS and Network Z-Scores. A. Transformed Z-scores are uncorrelated. **B.** Genes with high network scores had higher replication rates compared to those with low network scores, as further visualized and confirmed statistically as shown in **Figure 2.4**. Replate = replication rate.

To combine the Network and GWAS Z-scores, we developed an approach that uses a linear support vector machine (SVM) [90] to determine how heavily each type of score should be weighted in order to maximize replication rate (See Methods). These weights were then used in conjunction with the meta-analysis method for combining summary results implemented in METAL [95]. The weights predicted by the SVM (**Figure 2.4**) were 0.703 (GWAS) and 0.297 (Network). As further confirmation, we conducted binomial (logit family) logistic regression using network and GWAS Z-scores as predictors and the replication class (high/low) as the outcome. Both network and GWAS score were significant, (GWAS: coefficient = -0.659, $p < 2.0 \times 10^{-16}$) (Network: coefficient = -0.229, $p = 0.0016$). The coefficients derived from logistic regression are very similar to the SVM-derived weights (GWAS weight = 0.742, Network weight = 0.258).

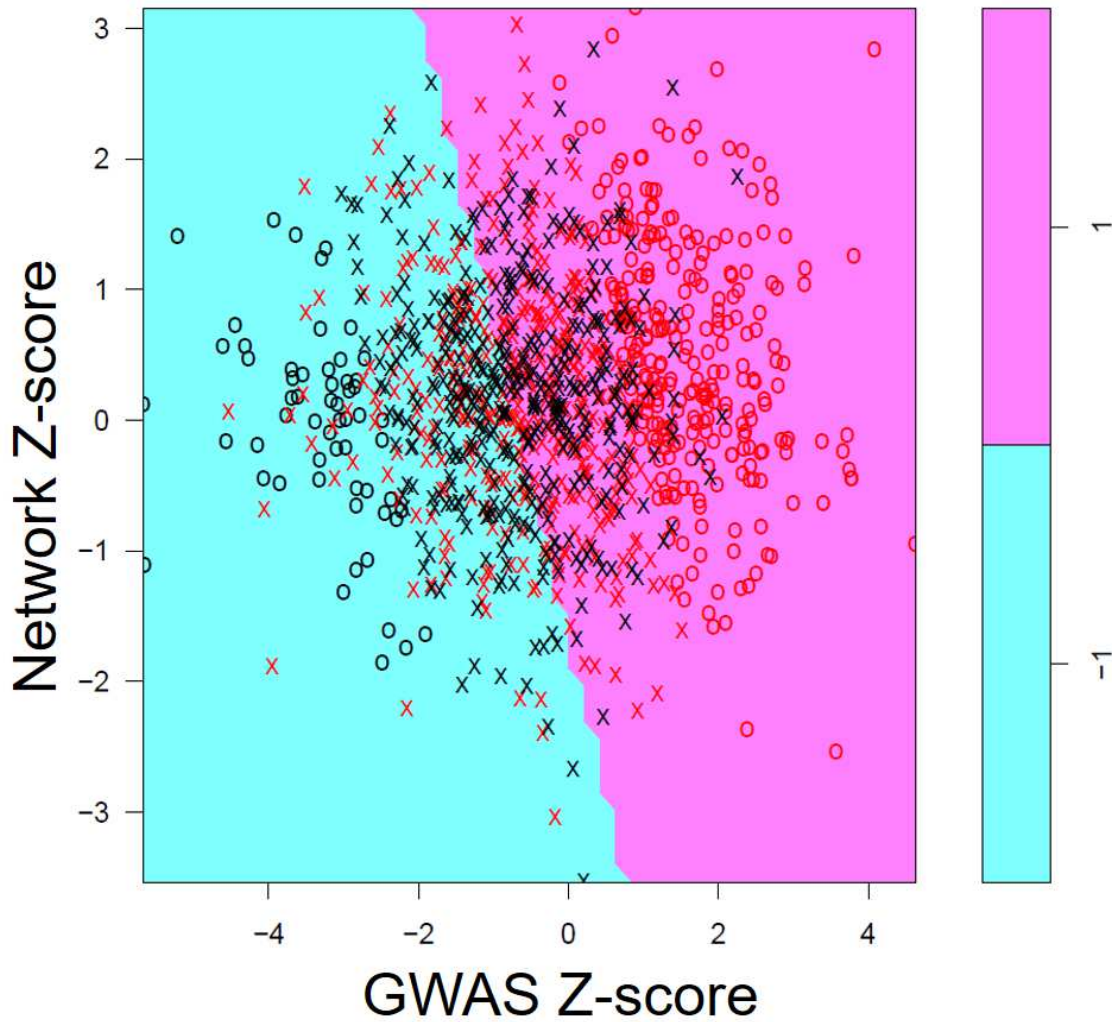


Figure 2.4: Support Vector Machine Training to Predict GWAS and Network Z-Score Weights. Selection of genes with a high replication rate (> 0.7 , blue points) and low replication rate (< 0.1 , red points) yielded a balanced number of genes in each replication class (high/low). A linear SVM model was trained to predict replication class using the GWAS and network Z-scores of each gene. Both network and GWAS Z-scores contributed to the decision boundary, as

demonstrated by the significance of their predicted coefficients using logistic regression (GWAS: $p < 2.0 \times 10^{-16}$, Network: $p = 0.0016$).

Next, we applied our combined approach genome-wide, excluding the RAD genes and genes containing significantly associated variants ($p < 1.0 \times 10^{-7}$) to focus on novel candidates. Among the genes with largest combined Z-scores (**Table 2.4**), several have important roles in inflammation. *CR2* ($p = 5.95 \times 10^{-7}$) is a receptor protein involved in immune response (genecards.com [96]). *SHARPIN* ($p = 1.43 \times 10^{-5}$) is a component of the LUBAC complex that plays a regulatory role in inflammation [96]. *PTPN2* ($p = 3.21 \times 10^{-5}$) is a phosphatase that also serves an important role in regulation of inflammation and glucose homeostasis [96]. The Bonferroni-corrected significance threshold when considering only genes in the 75th percentile of network scores is $p = 1.46 \times 10^{-5}$, although this is likely to be overly strict since proximally located genes are not inherited independently.

Table 2.4. Top Predicted AD Genes Using Combination Approach. The GWAS and Network Z-scores were combined into a single Z-Score using the weights determined by the SVM. The top 10 ranking genes are depicted in the table.

Gene	Z-Score		
	GWAS	Network	Combined
CR2	4.084	2.832	4.857
SHARPIN	3.983	1.320	4.185
PTPN2	3.805	1.259	3.997
C4B	2.846	2.928	3.750
TUBB2B	3.166	1.314	3.428
EPS8	3.156	1.156	3.358
PSMC3	3.145	1.036	3.302
STRAP	3.051	1.157	3.262
HSPA2	2.977	1.325	3.258
STUB1	2.895	1.407	3.213

We performed pathway analysis using Gene Set Enrichment Analysis (GSEA) [91] to determine if AD-related pathways are more enriched when genes are ranked by their combined Z-scores versus GWAS-only Z-scores (See Methods). Notably, ranking genes based upon combined Z-scores resulted in several significantly enriched AD-related pathways including immune response, FOXO3 targeting (indicates enrichment for aging), and hippocampal development (**Table 2.5**). By comparison, ranking genes based only upon their GWAS Z-scores resulted in virtually no significant pathways entirely (**Table 2.6**).

Table 2.5. GSEA Results After Ranking Genes by Combined Z-Scores.

NAME	SIZE	ES	NES	FWER p-val
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	64	0.487289	2.230758	0.042
DELPUECH_FOXO3_TARGETS_DN	37	0.527147	2.180592	0.07
BIOCARTA_PGC1A_PATHWAY	20	0.612627	2.179728	0.071
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	85	0.436425	2.171402	0.073
MURAKAMI_UV_RESPONSE_6HR_DN	20	0.592367	2.123952	0.117
GOLUB_ALL_VS_AML_DN	18	0.628548	2.118205	0.127
REACTOME_RNA_POL_I_PROMOTER_OPENING	28	0.551692	2.099766	0.149
MODY_HIPPOCAMPUS_PRENATAL	36	0.519005	2.097922	0.153
FARMER_BREAST_CANCER_CLUSTER_5	17	0.632376	2.089917	0.161
ZUCCHI_METASTASIS_DN	35	0.516274	2.066772	0.197
NAKAYAMA_SOFT_TISSUE_TUMORS_PCA1_UP	61	0.456125	2.058099	0.205
INGA_TP53_TARGETS	15	0.635151	2.049176	0.222

Table 2.6. GSEA Results After Ranking Genes by GWAS Only Z-Scores.

NAME	SIZE	ES	NES	FWER p-val
NAKAYAMA_SOFT_TISSUE_TUMORS_PCA1_UP	61	0.4396	2.108	0.134
FARMER_BREAST_CANCER_CLUSTER_5	17	0.6101	2.016	0.261
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	64	0.3975	1.95	0.418
GOLUB_ALL_VS_AML_DN	18	0.5596	1.888	0.591
CHIARETTI_T_ALL_REFRACTORY_TO_THERAPY	23	0.499	1.879	0.608
SHIN_B_CELL_LYMPHOMA_CLUSTER_5	15	0.541	1.772	0.864
ZUCCHI_METASTASIS_DN	35	0.4231	1.769	0.873
KIM_HYPOXIA	22	0.4561	1.704	0.964
DELPUECH_FOXO3_TARGETS_DN	37	0.3995	1.672	0.985
NIELSEN_LIPOSARCOMA_UP	15	0.5139	1.65	0.993

Discussion

GWAS of AD and AD-related endophenotypes have discovered and replicated associations with more than 60 genes (**Table 2.1**), many of which have roles in AD-related pathways (amyloid β aggregation, inflammation, cholesterol transport, immune response, etc.). To identify additional AD-related genes, we hypothesized that genes having suggestive evidence for association from a genome-wide screen and protein-level interactions (both direct and indirect) are more likely to replicate. This idea has been referred to as functional linkage [97]. To test this hypothesis, we developed a novel approach for improving the prioritization of candidate disease genes that incorporates a network diffusion of scores from known disease genes using a protein network and integration with GWAS risk scores. We tested this approach on a large AD GWAS dataset and validated the performance of the methodology using bootstrap aggregation. Several novel AD genes were predicted including *CR2*, *SHARPIN*, and *PTPN2*.

Part of the motivation for our approach was to identify genes that are more obviously biologically relevant to AD. This is exemplified by *SHARPIN*, whose principal known function is to form the LUBAC complex and prevent inflammation, a major process through which amyloid aggregation and AD are thought to develop [48]. Similarly, *CR2*, a homolog of *CR1* which is a well-established AD gene [35], is involved in immune response. Many immune response genes are differentially expressed between healthy and AD brains, and

investigations into the connection between expression in cell types and the presence of AD has led to growing interest in the role microglial cells (a first responder in the immune response pathway) [98]. Finally, *PTPN2* is involved in multiple AD-related pathways; it has roles in negatively regulating inflammation and de-phosphorylation of key glucose metabolism kinases including *INSR* and *EGFR* [99]. The AD-related roles of each of our novel AD gene predictions, in combination with their strong network and genetic scores, make them highly promising candidates.

One biological form of functional linkage that does not require direct physical interaction is membership in the same signaling pathway or protein complex. For example, our study identified interaction between *FOXO* and *INSR* that is consistent with evidence of a multi-link signaling pathway comprised of direct physical interactions in the insulin-signaling pathway [92]. By comparison, neighborhood enrichment approaches (i.e., testing a gene's direct interactions) cannot detect indirect interactions. Furthermore, neighborhood enrichment approaches are unreasonable for AD because some RAD genes are network hubs (e.g., *APP* has more than 2000 interactions) which would result in an unreasonably high number of genes having AD-enriched neighborhoods.

Some distance metrics capture indirect interactions by calculating the proximity between a pair of genes based upon short paths between them in the network.

However, after testing a simple distance metric known as average shortest path (ASP), we observed that hub genes were still the top-ranked predicted genes. Since hub genes have many interactions, they tend to have short overall paths to any genes in a network, although their functions are highly generic and unlikely tied to a particular disease. Ubiquitin C (*UBC*), for example, has nearly 9,000 interactions; however, this is simply because protein degradation is essential for regulating the vast majority of proteins. Therefore, a more nuanced network propagation approach can aid in making disease specific inferences.

Network diffusion is a widely used class of spectral graph clustering methods that have been applied to many computational disciplines [74]. We used this approach to propagate evidence in the form of AD scores throughout the network. A protein in the network that has a short “diffusion distance” to one or more well-established AD genes will receive a high network risk score. Notably, we observed that network diffusion down-weights hubs while simultaneously outperforming ASP distance when applying leave-one-out cross-validation to the RAD genes. Many diffusion kernels have been proposed in graph theory, however the Regularized Laplacian [74] approach used in this study has the highly desirable properties of requiring very little parameterization (in fact, only a single parameter is required to be set) and also more computationally efficient than other diffusion kernels. Network diffusion methods have been applied in other genetics research contexts such as labeling somatic network mutations in

cancer [100], characterizing gene sets [101], and predicting risk genes for amyotrophic lateral sclerosis [27].

We also observed that genes with high diffusion scores tended to replicate more frequently in our 125 pairs of bootstrapped discovery and replication datasets. However, network Z-scores and GWAS Z-scores in the full dataset were not strongly correlated. Taken together, these observations indicate the importance of considering jointly protein interaction data and genetic results even though they are independent because the integration of both types of information will likely yield noticeable improvement in replicability of findings. Since our bootstrapping procedure required splitting the original dataset, the simulations were conducted using datasets that contained only one-half of the total sample. This suggests that our network scores aided in determining which genetic associations were real in datasets with reduced power. We note that our bootstrapping approach was performed on the same data from which we derived the GWAS Z-scores used to train the SVM. Therefore, the selection of combination weights may have been biased in favor of GWAS Z-scores. Furthermore, it is unclear whether the weight combination used in this study (0.297/0.703) would be appropriate for combining genetic and network data for other disorders or traits.

The GWAS approach has a very limited capability to identify the entire set of genes which contribute to the risk of a complex disease like AD, even in datasets containing up to 100,000 individuals, because some genes do not contain variants that are sufficiently frequent and/or exert a large enough effect to yield a statistically significant association. To overcome this limitation, we developed a novel SVM approach to integrate the genetic and network scores by propagating GWAS Z-scores in a PPI network. In the AD example presented here, we initialized the RAD genes to have an identical high score in the network, thereby allowing re-prioritization of genes in any AD dataset regardless of the internal Z-scores of the RAD genes.

We acknowledge that our initial choice to treat each RAD gene equally may be controversial. Arguably, we could have seeded our analyses with GWAS Z-scores for each RAD gene from the original studies. However, our approach permits unbiased exploration of interactions of all plausible AD genes and does not require adjustment to these Z-scores for sample size or allele frequencies. Moreover, results derived from weighted RAD genes would be dominated by interactions with *APOE* for which the significance level exceeded a $-\log(\text{p-value})$ of more than 100 in several datasets (compared to < 10 for most other RAD genes in the total group of datasets. Also, several key AD-related genes (e.g., *APP*, *PSEN1* and *PSEN2*) which show little evidence for association with individual SNP or gene-based tests for AD would be undervalued in analyses

using weighted Z-scores. In order to make our software maximally flexible and support weights derived from confidence in the seed genes, we implemented an option for users to specify unequal weights on the seed genes at their own discretion.

A potential concern about our results is the strategy for selecting RAD genes because many significant GWAS findings include variants located in intergenic regions. The most parsimonious explanation is that the variant responsible for the association peak influences the nearest gene, but there is abundant evidence suggesting this assumption is often incorrect. To address this issue, we repeated our analyses using a more restricted set of RAD genes that included only those supported by genome-wide significant evidence of association with AD risk and replication in independent datasets or by other genetic evidence plus experiments linking them to AD-related pathophysiology. Our leave-one-out cross validation approach demonstrated that the genes in the restricted RAD set had closer network proximity to each other than would be expected by chance ($p = 5.932e-05$). The statistical support for the novel genes *CR2* ($p=4.09 \times 10^{-7}$), *SHARPIN* ($p=1.10 \times 10^{-5}$), and *PTPN2* ($p=2.41 \times 10^{-5}$) remained the same. Finally, combined Z-scores that were derived using diffusion from the more conservative RAD gene set yielded similar AD-related pathways such as Fx03 targets (FWER $p=0.064$), antigen processing (FWER $p=0.02$), and hippocampal development (FWER $p = 0.065$). These results confirm that the genes with a clear functional

role in AD produce network diffusion-based predictions that are consistent with the results presented here. Curiously, the inclusion or exclusion of the portion of RAD genes that have an ambiguous or non-validated functional role in AD did not affect our results.

We also acknowledge that several of the novel putative AD genes may have been erroneously prioritized because they are in the same locus with RAD genes. This concern is unlikely noting that there are several instances where a genetic association peak includes multiple genes that may have a possible functional role in AD (e.g., the *MS4A* gene cluster [35]). Although one of our novel AD genes, *CR2*, is located close to *CR1*, which is an unambiguous RAD gene given its robust replication in GWAS and effect on deposition of neuritic amyloid plaque, *CR2* is also an intriguing AD candidate gene because it has been shown to regulate hippocampal neurogenesis [63]. Thus, our findings suggest that our approach will aid in predicting truly multiple AD-related genes at a locus, however additional biological evidence may be required in some instances to make this distinction.

Previous AD studies have implicated inflammation and immune response genes, but we did not observe any enrichment for these pathways when using the GWAS-only scores in our dataset. However, these and other known AD-related pathways emerged after applying our network re-prioritization method (Table 2.6)

suggesting that incorporation of network data can help minimize discrepancies in predictions across different genetic datasets. However, a few well-established AD-related pathways were not detected by our analysis including cholesterol metabolism and endocytosis. Upon further investigation, enrichment for the cholesterol homeostasis pathway is not significant when applying GSEA to the genetic data only (FWER $p = 1$). The cholesterol homeostasis pathway as defined in MSigDB is general and therefore contains a high number of genes with weak associations to AD that diminish the enrichment of the set. The evidence for this pathway is greater in the analysis using only network scores (FWER $p = 0.18$), which indicates our method still improves the detection cholesterol homeostasis. We do however replicate enrichment for the HDL mediated lipid transport pathway using only our genetic data (FDR $q = 0.11$), but this is primarily driven by a strong signal from APOE. The seed genes from the network analysis, such as APOE, need to be ignored to prevent bias, and therefore the lack of enrichment of the HDL mediated lipid transport pathway after applying our network approach is simply explained by the removal of APOE.

It should also be noted that a simply connected network is a requirement for the diffusion algorithm to work properly.

Our approach offers several advantages in comparison to other network-based approaches including biological transparency, ease of integration with a variety of

GWAS methods, and the ability to balance data-driven statistics and biological prior probabilities. The extensive simulations we conducted provide a general basis for further establishing the practicality of genetic and network-based integration. Our network methodology is specifically developed with the goal of accommodating known complications of genetic analysis.

The software developed for this study is open source, accessible to most users (incorporated in an R package), and applicable to any set of variant- or gene-level disease association results. Importantly, it requires only a set of GWAS results and a list of previously known disease genes and, therefore, does not necessitate changes to previously established genetic analysis pipelines.

Although we used an SVM procedure to determine the weights for the score combination, a user can specify any weights or simply use our defaults that are based on the 0.297/0.703 ratio determined by SVM. Our package is accessible through GitHub (<https://github.com/lancour/ignition>).

**Chapter 3: Analysis of Brain Region-Specific Co-Expression Networks
Reveals Clustering Properties of Genes Associated with Alzheimer Disease
Risk and Identifies Novel Risk Gene Candidates**

Neurodegenerative (ND) diseases, such as Alzheimer disease (AD), Parkinson disease (PD), Huntington disease (HD), and amyotrophic lateral sclerosis (ALS), impair or damage neurons. Although many sub-cellular similarities between NDs have been identified [102], the regional differences between them are quite profound [103-106]. For example, neuronal cell death from HD is primarily localized to the basal ganglia, whereas both AD and PD result in cell death throughout the brain [106]. Furthermore, PD causes the most severe cell death in the substantia nigra [103] whereas AD most heavily affects the hippocampus, the frontal cortex, and the temporal lobe [105]. These studies highlight the importance of studying gene expression signatures and relationships of AD-associated genes in different brain regions. For instance, an increased correlation in gene expression among two AD-associated genes in brain structures such as the cortex as compared to other brain regions suggests either a functional relationship or cell/sub-region specific expression biases towards cell types where the disease tend to originate or progress most rapidly.

Altered functional connectivity between brain regions has been demonstrated for several neuropsychiatric diseases including schizophrenia, depression, and AD using functional magnetic resonance imaging (fMRI) [107-109]. Brain imaging and neuropathological studies indicate that the hippocampus, which has a role in memory formation, is one of the first structures showing marked neuronal loss in AD and, compared to other regions, suffers the largest relative reduction in volume by the latter stages of the disease [110]. Regional specificity is also evident by longitudinal patterning of the AD-related tau and amyloid- β proteins that aggregate into neurofibrillary tangles and senile plaques, respectively [111]. In the early stages of AD, a small number of tangles typically form in the brain stem, and then spread aggressively to most of the cerebrum by the latest stage. Amyloid plaques form in the opposite pattern, beginning primarily in the outer cortex and spreading inward and then to the brain stem [111]. Notably, very few protein aggregates form in the cerebellum even at the most severe stages of AD.

The aforementioned differences in AD severity between regions of the brain may be a consequence of a variety of factors. One such factor is the tissue specific expression patterns of genes throughout the body, which is a relevant consideration for the human brain given its vast complexity and compartmentalization [112]. An additional factor may also be the changing cell type fractions observed between major regions of the brain [113, 114]. Large

scale multi-omic approaches have been able to assist in understanding these complicated roots of neuropsychiatric disease [115, 116]. Furthermore, they have been able to identify novel disease-related gene candidates [27, 117, 118].

In this study, we applied network-based correlation methods to existing genome-wide association study (GWAS) data and gene expression data derived from brain in order to identify additional AD-related genes using network methodology. In addition to identifying several novel biologically relevant genes for AD, we show that the strength of correlations among previously established AD genes increases when the networks are restricted to the sub-regions of the brain that are most impacted by AD.

Methods

Acquisition of GWAS Data and Curation of AD Genes

We obtained summarized results from a GWAS for AD risk conducted using 16,175 AD cases and 17,175 controls of European ancestry that were derived as previously described [119]. Association evidence with each gene was derived from P-values for association with individual single nucleotide polymorphisms (SNPs) corrected for multiple testing using an approximation that has been shown to be a conservative adjustment for recombination hotspots, linkage disequilibrium, and gene size [88]. This correction can be expressed as

$$P_g^{Gene'} = 1 - (1 - P_g^{BestSNP})^{\frac{N+1}{2}}$$

where N is the number of SNPs existing within a gene. Analyses for this study were also predicated on a group of reproducible AD (RAD) genes which were previously curated from the literature [119].

Acquisition, Labeling and Processing of Brain Expression Data

Measurements of gene expression in the human brain were acquired from the Allen Brain Atlas (ABA). This database contains microarray data from 3,702 single tissue samples extracted from six neuropathologically healthy brains (ages 24, 26, 31, 39, 49, and 57). Data derived from each sample consists of an expression vector containing expression measurements from 45,000 probes in the extracted tissue. Each sample was annotated at three different levels of granularity, which are defined for the purpose of this study as low, mid, or high level structures in terms of region specificity. Principal components of the expression vectors across all samples were computed using the prcomp method from the R programming package [120]. Then each sample was annotated according to brain region based on the hierarchical labeling scheme described above. A scatterplot of the first and second principal components (PCs) was produced to ascertain whether the expression vectors of samples displayed batch effects related to either the region or the brains from which samples were derived.

Determining Gene Expression Within Each Brain Region

Because some genes are queried by multiple probes, the mean expression of all probes mapping to each gene was computed, resulting in a gene x sample expression matrix. Expression of each gene in each brain region was adjusted using a mixed model approach to account for repeated sampling of both individual brains and regions in the ABA dataset [121]. Each gene has an expression vector, E , of its expression levels in each sample. Each sample has an indicator of the *region* and the *brain* it was taken from. The model specification in R is then:

$$E \sim -1 + \text{region} + (1 | \text{brain})$$

The *region* coefficients of this model represent a single gene's expression in each of the 232 low level regions. This model is used for each individual gene, to ultimately determine each gene's expression within each of the low level regions ascertained in the ABA dataset.

Construction of Region-Specific Brain Co-Expression Networks

We constructed a co-expression network based upon correlations between all pairs of genes across the cerebrum, cerebellum, and the brain stem. In this instance, each node of the network is a gene and each edge between genes is the absolute value of the Pearson correlation coefficient between a pair of genes.

Due to the high impact of AD on the cerebrum, two additional correlations networks were created by subdividing the cerebrum into two non-overlapping subsets of regions representing Braak stages [122]. In total, there were 79 regions of the cerebrum in the ABA dataset that showed some evidence of AD at Braak stage 1, which we refer to as the early stage regions, and 37 regions of the cerebrum that showed more pronounced AD pathology at Braak stage 3, which we refer to as late stage regions. Correlations between all gene pairs were computed separately for early and late stage regions. In order to compare correlations between sets of genes of interest across networks, we normalized the correlations within each network. For this we applied a novel metric, referred to as median ranking by correlation (MRC), that is derived using a “leave one out” strategy to normalize the distribution of correlations into a uniform distribution of ranks that is comparable across networks. Let RAD be the set of RAD genes and let G-RAD be the set of other genes.

1. For each gene G_i in RAD let RAD- G_i be the set of genes in RAD different from G_i . For each G_i we compute the total sum of the correlations of G_i to the other genes in RAD (RAD- G_i). We refer to this as SUM-COR(G_i , RAD- G_i). Now for each gene G_j in G-RAD (genes outside the RAD set) we compute the full distribution of sum of correlations to genes SUM-COR(G_j , RAD- G_i).

2. Let the Rank($G, \text{RAD-}G_i$) be the rank of SUM-COR($G_i, \text{RAD-}G_i$) in the full distribution of SUM-COR($G_j, \text{RAD-}G_i$).
3. Let MRC be the median such rank.

If a gene set is clustered, the median of these rankings (MRC) of the RAD genes will exceed the expectation of 0.5.

Verifying Consistency of Gene Rankings Across Correlation Networks

Due to the small number of brain specimens in the ABA dataset, we tested the consistency of gene rankings by constructing a gene x gene correlation network for each brain. This procedure uses the same correlation approach described above, except that the expression levels of genes in each region were determined based on measurements from a single brain at a time. Next, each non-RAD gene was ranked by its correlation (RC) to the RAD seed genes within each of the six individual correlation networks. Finally, a Kendall Tau rank correlation matrix was derived based upon all possible combinations of these six ranked lists.

Network-Based Ranking of Novel AD Genes

Based on the observation that the RAD genes tend to be highly correlated, we hypothesized that other genes showing high correlation with established AD genes are likely to be AD-related genes. Therefore, a summed absolute Pearson

correlation with all the RAD genes was computed for each non-RAD gene. Next, a percentile rank of each non-RAD gene based upon these sums was computed and converted to Z-scores, which we refer to as network scores. If N genes are being ranked, then the percentile rank for each gene is: Percentile = (rank) / (N+1), ranging from 0 to 1. These percentiles form a uniform distribution, which are converted to Z-scores using `qnorm(Percentile, lower.tail = F)` in R. These network scores were then combined with genetic association Z-scores derived by a GWAS for AD risk including approximately 30,000 individuals using a combination procedure from the meta-analysis tool METAL that was modified to equally weight both scores [95]:

$$Z_{combined} = \frac{0.5 * Z_{gwas} + 0.5 * Z_{network}}{\sqrt{0.5^2 + 0.5^2}}$$

Further ranking was performed by integrating phenotypic information from gene orthologs in Flybase to focus on genes which when knocked out in a model organism result in an AD-related phenotype including premature aging, defective memory, defective aging, and oxidative stress [123].

Results

Principal component analysis did not reveal any batch effects related to sub-region or brain specimen, however clustering of the samples was observed due to high level brain structure resembling the cerebrum, cerebellum, and brain stem

(**Figure 3.1**). The extent to which these samples are similar cannot be exactly determined with a PCA approach since the expression of many genes is not independent, but the clustering we observed suggests the presence of a batch effect. Further analysis revealed that RAD genes tended to have homogenous expression in the cerebellum and brain stem regardless of the changes in the mid-level structure (**Figure 3.2**). However, expression of these genes was highly variable across mid-level structures in the cerebrum.



Figure 3.1: Principal Component Analysis Indicates Clustering by High Level Structure. The principal components were computed for all samples in the dataset. Clustering was checked for low level structure, individual, and high level structure. Only high level structure, as labeled here, appeared to form clusters in the principal component plot (CX = cerebrum, BS = brain stem, CB = cerebellum)

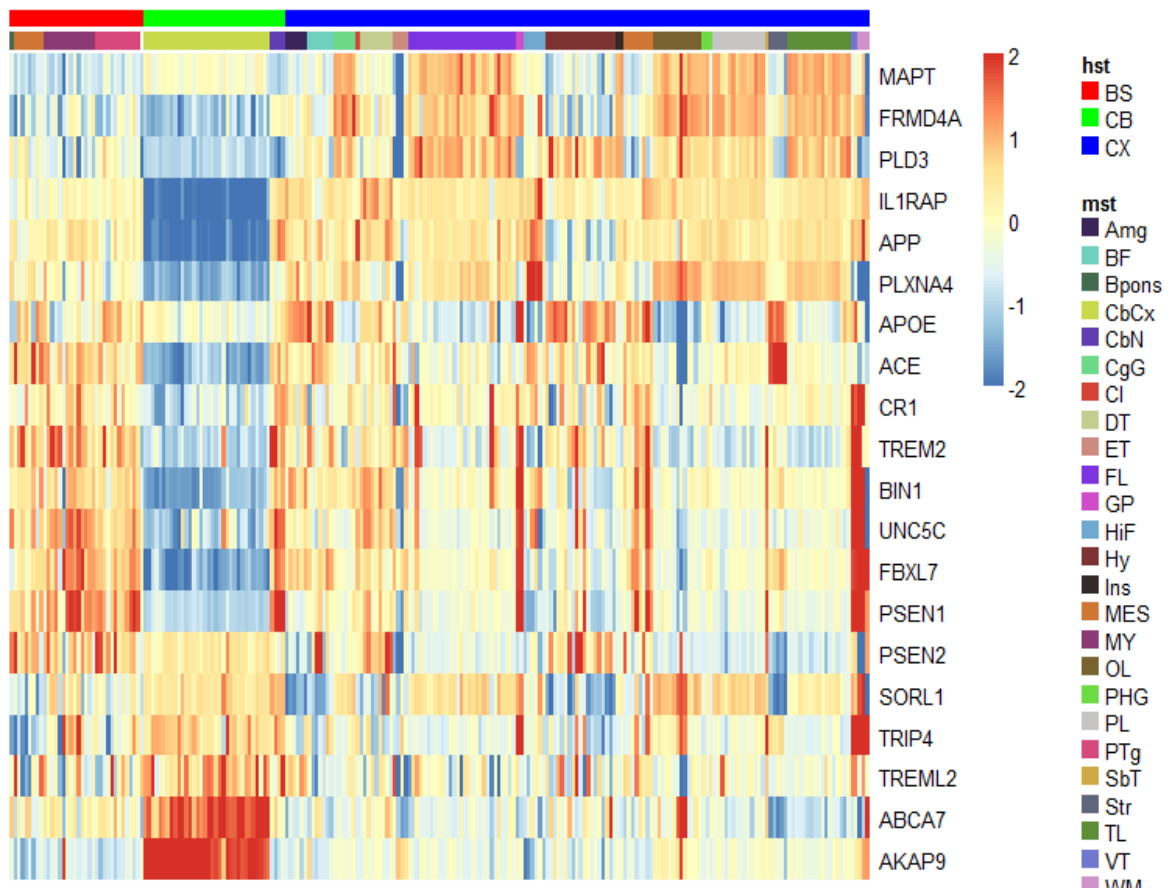


Figure 3.2: RAD Genes Have Region-Specific Expression Levels. 20 of the commonly studied RAD genes were input into a heatmap clustering model. The RAD genes appear to cluster by high level structure (hst), which matches our expectation from the PCA. The brainstem and the cerebellum tend to have homogenous expression for individual genes, whereas the cerebrum has highly varied expression across mid-level structures (mst's).

Comparison of co-expression of RAD genes across high level brain regions revealed higher correlation ranks (RC) in the cerebrum (0.748) than in the brain stem (0.648) and cerebellum (0.574, Table 3.1). These differences appear to be due largely to a few genes including *APOE* and *MAPT* which showed much greater co-expression in the cerebrum (RC = 0.745 and 0.863 respectively) than in the cerebellum (RC = 0.280 and 0.542, respectively) and brain stem (RC = 0.216 and 0.337, respectively). Surprisingly, the RC for *APP* was much higher in the cerebellum (0.99) than in the brain stem (0.505) and cerebrum (0.376). Multiple RAD genes including *PSEN2*, *EPHA1*, *LMX1B*, *TPBG*, *CLU*, *AKAP9*, *ZNF804B*, *PDGFRL*, and *ABCA7* were not meaningfully co-expressed with other RAD genes in any of the structures.

Table 3.1. RC of RAD Genes in the Cerebrum, Cerebellum, and Brain Stem.

The RC of each of the RAD genes was computed in each of the correlation networks for each high level structure. From these, the MRC is computed and determined to be highest in the cerebrum (MRC = 0.749) in comparison to the cerebellum (MRC = 0.574) and brain stem (MRC = 0.648). Genes are ordered in the table based upon the variance of their RC across the three structures.

Highlighted rows are examples of well-established AD genes that have dramatically different RC between brain structures.

Gene	Percentile Ranking by Correlation		
	Brain Stem	Cerebellum	Cerebrum
<i>ZCWPW1</i>	0.991	0.080	0.783
<i>TRIP4</i>	0.997	0.226	0.772
<i>SORCS1</i>	0.076	0.602	0.807
<i>OSTN</i>	0.753	0.196	0.858
<i>PLXNA4</i>	0.182	0.601	0.854
<i>SORCS2</i>	0.785	0.136	0.605
<i>CASP8</i>	0.222	0.720	0.853
<i>AKAP9</i>	0.541	0.877	0.214
<i>APP</i>	0.505	0.990	0.376
<i>ABCG1</i>	0.392	0.325	0.897
<i>TP53INP1</i>	0.880	0.394	0.968
<i>PFDN1</i>	0.557	0.998	0.406
<i>COBL</i>	0.902	0.323	0.662
<i>APOE</i>	0.216	0.280	0.745
<i>SORCS3</i>	0.134	0.376	0.706
<i>EPHA1</i>	0.219	0.739	0.294
<i>ACE</i>	0.778	0.667	0.249
<i>CR1</i>	0.068	0.344	0.603
<i>MAPT</i>	0.337	0.542	0.863
<i>KCNMB2</i>	0.021	0.530	0.401

<i>CLU</i>	0.732	0.582	0.219
<i>PDGFRL</i>	0.634	0.395	0.126
<i>PLD4</i>	0.948	0.463	0.770
<i>SORL1</i>	0.935	0.462	0.612
<i>CD2AP</i>	0.351	0.691	0.772
<i>GALNT7</i>	0.390	0.823	0.699
<i>SLC10A2</i>	0.329	0.224	0.640
<i>PILRA</i>	0.601	0.896	0.494
<i>CASS4</i>	0.040	0.204	0.438
<i>PLD3</i>	0.536	0.355	0.752
<i>MS4A6A</i>	0.603	0.623	0.934
<i>C1QTNF4</i>	0.804	0.714	0.451
<i>MS4A4A</i>	0.728	0.499	0.859
<i>ABI3</i>	0.685	0.567	0.908
<i>NCR2</i>	0.776	0.664	0.975
<i>UNC5C</i>	0.797	0.517	0.765
<i>PTK2B</i>	0.688	0.979	0.899
<i>MEF2C</i>	0.812	0.695	0.986
<i>ABCA7</i>	0.082	0.338	0.095
<i>LMX1B</i>	0.003	0.078	0.276
<i>TREM2</i>	0.793	0.820	0.963
<i>ECHDC3</i>	0.805	0.703	0.670
<i>BIN1</i>	0.663	0.798	0.724
<i>PSEN2</i>	0.242	0.256	0.356
<i>CD33</i>	0.876	0.832	0.950
<i>ZNF804B</i>	0.073	0.141	0.162
<i>PICALM</i>	0.912	0.922	0.989
<i>HLA-DRB5</i>	0.973	0.910	0.986
<i>PSEN1</i>	0.879	0.924	0.866
<i>INPP5D</i>	0.937	0.987	0.982
<i>TPBG</i>	0.220	0.240	0.249
<i>PLCG2</i>	0.943	0.963	0.945

The MRC of the RAD genes was appreciably higher for the late stage network (0.733) than early stage network (0.615), but the RCs for many individual genes including *APOE*, *APP*, and *MAPT* were similar across these two networks (Table 3.2). Comparison of correlation networks in the cerebrum constructed for each individual showed that RAD genes tend to have low variability in RC among individuals within this dataset (**Figure 3.3**). The patterns of co-expression across the individual brains is moderately high and consistent with RC values between 0.455 and 0.652 (**Figure 3.4**).

Table 3.2: RC of RAD Genes in Late and Early Stage Correlation Networks.

To determine if there was an observable relationship between known AD pathology and the high MRC of the RAD genes, we next divided the cerebrum network into two parts: one part containing all of the early stage AD regions (Braak = 1), and the other part containing the later stage AD regions (Braak = 3). For both late and early stage, a network was constructed across the relevant brain regions and the MRC of the RAD genes was computed. The late state network had higher MRC (0.733) in comparison to the early stage network (0.615).

Gene	Percentile Ranking by Correlation	
	Early	Late
<i>UNC5C</i>	0.127	0.997
<i>TP53INP1</i>	0.093	0.931
<i>ZCWPW1</i>	0.156	0.979
<i>ABCG1</i>	0.192	0.956
<i>PLD4</i>	0.233	0.963
<i>KCNMB2</i>	0.954	0.269
<i>ABCA7</i>	0.831	0.180
<i>CLU</i>	0.822	0.235
<i>TPBG</i>	0.735	0.163
<i>CASS4</i>	0.202	0.774
<i>SLC10A2</i>	0.284	0.851
<i>ECHDC3</i>	0.951	0.396
<i>INPP5D</i>	0.363	0.874
<i>AKAP9</i>	0.647	0.195
<i>SORL1</i>	0.975	0.529
<i>SORCS1</i>	0.865	0.419
<i>LMX1B</i>	0.588	0.158

<i>CASP8</i>	0.425	0.851
<i>OSTN</i>	0.080	0.505
<i>NCR2</i>	0.492	0.888
<i>SORCS3</i>	0.482	0.864
<i>PICALM</i>	0.640	1.000
<i>C1QTNF4</i>	0.105	0.428
<i>TRIP4</i>	0.619	0.896
<i>PSEN1</i>	0.712	0.984
<i>PSEN2</i>	0.594	0.380
<i>PILRA</i>	0.489	0.289
<i>ABI3</i>	0.671	0.867
<i>PLCG2</i>	0.984	0.794
<i>ACE</i>	0.065	0.249
<i>APOE</i>	0.987	0.807
<i>PTK2B</i>	0.756	0.579
<i>PDGFRL</i>	0.111	0.272
<i>SORCS2</i>	0.481	0.641
<i>MS4A6A</i>	0.728	0.884
<i>MS4A4A</i>	0.881	0.727
<i>TREM2</i>	0.849	0.999
<i>GALNT7</i>	0.737	0.882
<i>EPHA1</i>	0.330	0.467
<i>CD2AP</i>	0.531	0.667
<i>HLA-DRB5</i>	0.847	0.948
<i>COBL</i>	0.830	0.740
<i>PLXNA4</i>	0.847	0.760
<i>CR1</i>	0.610	0.691
<i>PLD3</i>	0.584	0.647
<i>MEF2C</i>	0.800	0.739
<i>PFDN1</i>	0.596	0.657
<i>ZNF804B</i>	0.219	0.168
<i>APP</i>	0.676	0.705
<i>BIN1</i>	0.949	0.925
<i>CD33</i>	0.977	0.988
<i>MAPT</i>	0.534	0.526

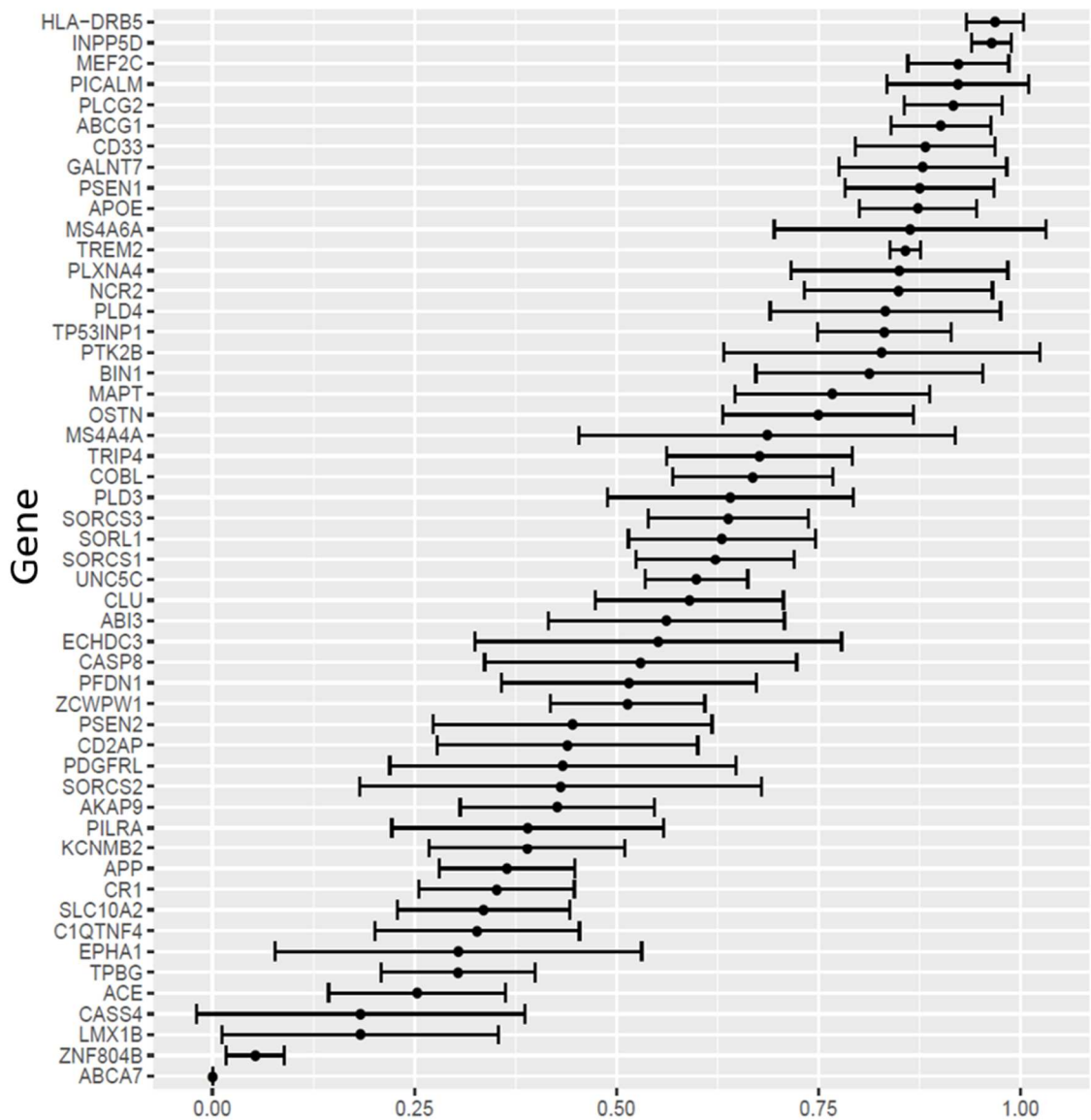


Figure 3.3: The RAD Gene Set Has Consistently High MRC in Each Individual. Due to the small number of individuals in the dataset (six), a correlation network of the cerebrum was constructed for each of the six brains. The MRC for each RAD gene was computed in each of the six individual

correlation networks. Depicted here are the MRC for each RAD gene averaged across the six individuals. The points denote the mean MRC, and the bars denote the 95% confidence intervals.

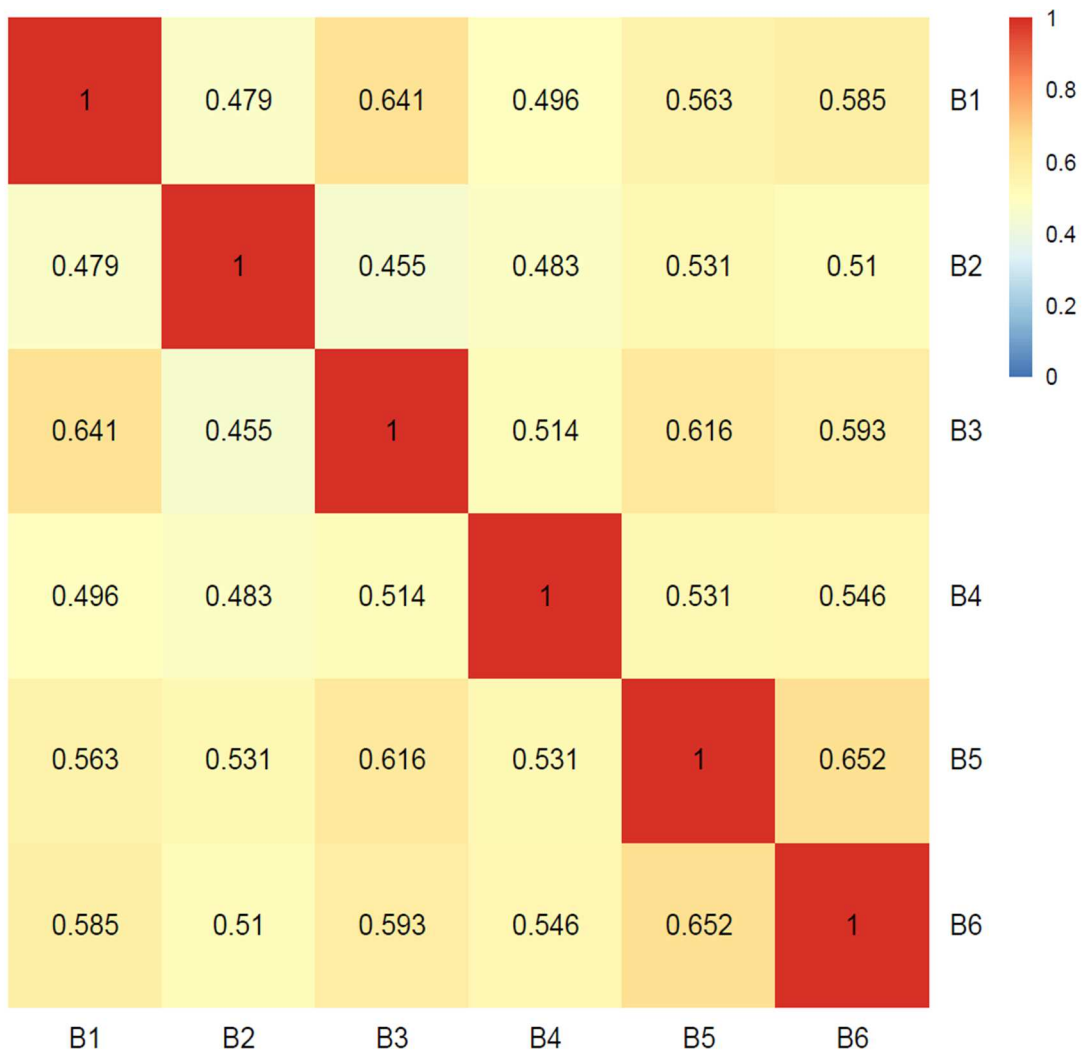


Figure 3.4: Rankings Using RAD Genes Are Consistent in Each Individual

Brain. In order to test if the network gene scores produced from the RAD genes were robust, we computed all genes scores from the RAD genes in each of the six individual networks. We then computed the Kendall Tau rank correlation, ordering the genes by their scores, for each pairing of individuals. Gene score orderings are strongly correlated between all possible pairings of individuals.

In order to predict novel AD genes based upon the above observations, a network score was produced for each non-RAD gene using the cerebrum correlation network in which clustering of the RAD genes was strongest. These network scores were then combined with GWAS Z-scores resulting in re-ordered AD gene rankings. A normal approximation was used to evaluate the significance of the combined scores. These results were filtered using gene knockout information from Flybase to limit the focus to genes which have functional evidence for producing AD-related phenotypes. Of the remaining 654 genes after the final filtering step, there was significant evidence for two novel AD-related genes including *EPS8* (FDR q-value = 8.77×10^{-3}) and *HSPA2* (FDR q-value = 0.245) (Table 3.3). Several previously reported AD genes also had high rankings but were not significant after FDR correction including *ADAM10* (FDR q-value = 0.401) and *HDAC1* (FDR q-value = 0.791). Another potential novel candidate gene, although not as statistically supported, is *CAT* (FDR q-value = 0.791) due to it influencing three total AD-related phenotypes in flies, which we did not observe of any other high ranking genes.

Table 3.3. Combined Z-Scores Reveal Novel AD Gene Candidates. Each non-RAD gene was assigned a new Z-score based upon an equal weight combination of its network and GWAS Z-scores. To further enhance our confidence in network gene scores, we then integrated information to include only scores for genes that are reported as inducing AD-related phenotypes (defective memory = MD, defective aging = DA, oxidative stress = OS, premature aging = PA, Alzheimer disease = AD) in flies. In total, 654 genes had phenotypic evidence in flies. Here we show all genes which had an unadjusted p-value of < 0.05.

		One-Tailed Z-Score			P-value	
Gene Name	Phenotype	GWAS	Network	Combined	Unadjusted	FDR
<i>EPS8</i>	DM	3.16	2.78	4.20	1.34E-05	8.77E-03
<i>HSPA2</i>	DA	2.98	1.51	3.17	7.51E-04	2.45E-01
<i>ADAM10</i>	AD	2.61	1.50	2.90	1.84E-03	4.01E-01
<i>HSPA6</i>	DA	2.13	1.63	2.66	3.88E-03	6.34E-01
<i>CAMK2A</i>	DM	0.94	2.60	2.50	6.13E-03	7.91E-01
<i>HDAC1</i>	AD,OS	1.40	1.97	2.38	8.55E-03	7.91E-01
<i>MAPK10</i>	AD,OS	1.58	1.68	2.30	1.06E-02	7.91E-01
<i>CAT</i>	AD,OS	2.31	0.89	2.27	1.17E-02	7.91E-01
<i>FXR1</i>	DM	0.58	2.60	2.24	1.24E-02	7.91E-01
<i>CD164</i>	DM	0.90	2.23	2.21	1.34E-02	7.91E-01
<i>HSPB1</i>	OS	1.58	1.52	2.19	1.43E-02	7.91E-01
<i>FBXW7</i>	OS	0.66	2.34	2.12	1.69E-02	7.91E-01
<i>DAGLB</i>	OS	1.52	1.48	2.12	1.71E-02	7.91E-01
<i>NFE2L3</i>	OS	1.71	1.20	2.06	1.98E-02	7.91E-01
<i>MAFB</i>	OS	1.56	1.31	2.03	2.11E-02	7.91E-01
<i>ITGAX</i>	DM	1.67	1.20	2.03	2.11E-02	7.91E-01
<i>SETBP1</i>	AD	1.26	1.60	2.02	2.14E-02	7.91E-01
<i>ACHE</i>	DM	2.12	0.71	2.00	2.28E-02	7.91E-01

<i>ITGAM</i>	DM	1.56	1.19	1.94	2.59E-02	7.91E-01
<i>ITPR1</i>	AD	1.01	1.72	1.93	2.68E-02	7.91E-01
<i>HBB</i>	OS	2.68	0.02	1.91	2.81E-02	7.91E-01
<i>PLK3</i>	AD	2.16	0.50	1.88	2.98E-02	7.91E-01
<i>TRIB3</i>	DM	1.35	1.29	1.87	3.08E-02	7.91E-01
<i>RCAN1</i>	AD,DM,OS	1.24	1.40	1.87	3.10E-02	7.91E-01
<i>GABARAP</i>	DM	0.74	1.87	1.85	3.23E-02	7.91E-01
<i>NIPBL</i>	DM	0.72	1.90	1.85	3.24E-02	7.91E-01
<i>GPD1</i>	OS	1.54	1.06	1.84	3.26E-02	7.91E-01
<i>GRIN2A</i>	DM	0.98	1.59	1.82	3.47E-02	8.10E-01
<i>ITPKA</i>	OS	0.57	1.95	1.78	3.78E-02	8.29E-01
<i>DNM1</i>	DM	1.02	1.46	1.76	3.94E-02	8.29E-01
<i>PGC</i>	AD	1.28	1.19	1.75	4.00E-02	8.29E-01
<i>CIDEC</i>	DM	1.60	0.87	1.74	4.06E-02	8.29E-01
<i>TXNRD2</i>	DA	1.85	0.59	1.73	4.21E-02	8.34E-01
<i>BZW2</i>	DM	1.60	0.80	1.69	4.51E-02	8.67E-01
<i>CBX3</i>	AD	1.54	0.81	1.66	4.81E-02	8.91E-01
<i>PCNA</i>	OS	2.31	0.03	1.65	4.91E-02	8.91E-01

Discussion

Previous studies using correlation or other network strategies have increased discovery and understanding the functional roles of novel disease related genes across many biological contexts [19, 26, 27, 124]. In this study, we applied an integrative network strategy to capture complex relationships between RAD genes across relevant regions of the brain and to aid discovery novel AD-related genes. This approach entailed integration of AD GWAS data, gene expression measures in multiple brain regions, and phenotypic information (i.e., memory and aging-related outcomes) from gene knockout studies in *Drosophila* [123]. By separating regions of the brain according to established patterns of AD-related pathology including neurodegeneration and protein aggregation, we showed that the correlation of expression between previously established AD genes is highest in regions severely impacted by AD, noting gene expression data were derived from brains without AD pathology. In addition, we identified potential novel AD genes by numerically combining results from analysis of co-expression of established AD genes and other genes in relevant brain regions with summary statistics from a large AD GWAS.

The most robust novel gene identified by our approach is *EPS8*. This gene encodes epidermal growth factor receptor substrate 8 which is involved in actin cytoskeleton regulation and is abundantly expressed in many brain regions [125].

The accumulation of filamentous actin (F-actin) is associated with tau-induced neurodegeneration in *Drosophila* and mouse tauopathy models [126]. Deletion of *Eps8* in mice leads to a reduction in hippocampal synaptic plasticity and impaired cognitive performance [127]. Three genes encoding heat shock proteins (*HSPA2*, *HSPA6*, and *HSPB1*) also emerged among our top findings. Notably, *HSPA2* was also identified as related to AD in a recent network analysis in an independent dataset [128]. Heat shock proteins have a major role in handling misfolded proteins including amyloid- β . [129] Although expression of heat shock protein genes has been well studied in AD, [130] there is little evidence for association of AD risk with polymorphisms in any members of this gene family with AD risk [131].

Several other top-ranked genes in our study have directly or indirectly been linked to AD. *ADAM10* encodes disintegrin and metalloproteinase 10 which is a synaptic enzyme that has been previously shown to limit amyloid- β_{1-42} peptide formation in AD. A variant in *ADAM10* recently achieved genome-wide significance in one of the largest genetic studies of AD containing more than 95,000 individuals [132, 133]. The catalase protein encoded by *CAT* binds with amyloid and inhibition of this interaction has been reported to protect cells from toxic protein aggregation [134, 135]. Several genes in the *HDAC* family have been reported to impair memory in animal models, and inhibitors of several

members of the *HDAC* gene family, including *HDCA1* identified for the first time in our study as an AD candidate gene, have been gaining support as a therapeutic approach for treating AD [136-139]. In humans, loss of *HDAC5* impairs memory function and variants in *HDAC9* have been associated with a dual outcome of neurofibrillary tangles and amyloid angiopathy [138, 140]. We also obtained mild evidence supporting a role for the gene encoding acetylcholinesterase (*ACHE*). This is a noteworthy finding in light of inconsistent and generally negative reports of association for AD with *ACHE* and related genes encoding choline acetyltransferase (*CHAT*) and butyrylcholinesterase (*BCHE*), despite the fact that AD is characterized by an extensive loss of cholinergic neurons from the basal forebrain area and the wide use of cholinesterase inhibitors to treat the early stages of cognitive decline [141].

A major motivation for our approach was to determine if the regional-specific effects that AD exhibits biologically can be detected using a correlation network approach. Recent work indicates that cell type compositions of brain regions are highly variable in aging brains, so cross-regional analysis is able to capture important properties such as changing cell fractions that may explain why the biological symptoms of AD are not uniformly present throughout the brain [114]. The high MRC of the RAD genes in the cerebrum supports this notion, given that the cerebrum tends to be the most major structure in the brain affected by AD

[105]. Further evidence for this is also provided by the low MRC of the RAD genes in the other brain regions (brain stem, cerebellum) where the effect of AD is far less severe. Notably, these patterns appear to be consistent in our study of cognitively healthy individuals (**Figure 3.3**).

Our findings also highlight several interesting patterns among several well established RAD genes. Expression of *APOE* and *MAPT* is highly correlated with other RAD genes in the cerebrum to the other RAD genes, but much less in the cerebellum and brain stem which is consistent with our previous observation of the RAD gene set as a whole. However, expression of *APP* and other RAD genes is highly correlated in the cerebellum (0.99), but not in the cerebrum (0.38). One explanation for this peculiar observation is that the hypothesis of RAD genes clustering via a correlation metric relies on coordinated changes in expression between RAD genes across regions of the brain. *APP* has consistent expression across most regions of the cerebrum, as evidenced in the Gene Tissue Expression (GTEx) portal, and thus would not appear correlated with genes which have more variable expression in the same regions, such as the other RAD genes [142]. A clearer understanding of this pattern will require focused analysis of gene co-expression within specific regions in the cerebrum.

Interpretation of our results has several caveats. First, we analyzed a dataset that has few individuals but a high number of brain regions in which expression was measured. However, expression patterns were consistent across individual brains in the dataset. If we had chosen instead a publicly available dataset containing a larger number of individuals but expression measurements in fewer regions, we would not have observed the high variation in expression of the RAD genes across regions of the cerebrum. This underscores the need for larger samples of brains with expression data in more precisely defined regions. Second, the present study did not include any brains from AD individuals. Although we utilized known Braak staging to characterize regions, it could be meaningful to have measurements of gene expression from these regions at varying stages of AD progression in order to understand how the presence of AD influences the observations we have noted of the RAD genes. Finally, we did not uncover a high number of genes that remained significant after an FDR-correction. In part, this was due to a power issue as the both the expression and genetic datasets used could greatly benefit from the inclusion of additional individuals.

This work establishes a strong foundation for many potential follow-up investigations. As larger fine-grained brain region expression datasets become available it will be important to confirm that the patterns we observed understand

how these patterns differ among brains with various stages of AD-related pathology. Although highly granular regional expression data from AD brains is not readily available, efforts are in progress by the AMP-AD consortium to profile expression of various regions of AD brains [143]. Validated differences in cross-regional correlation patterns between healthy and AD brains would improve understanding of mechanisms underlying the progression of AD and inform strategies for developing more effective therapeutic targets.

Chapter 4: Analysis of Median Ranking by Diffusion of General Phenotype Sets and the Effects of Cell Type and Disease on the Clustering Properties of Alzheimer Genes

The multi-omic network approaches employed in earlier chapters have had a demonstrable role in clarifying the joint behaviors of the known AD genes as well as suggesting several novel candidate AD genes. Associations with several of the proposed candidate AD genes have been identified with genome-wide statistical evidence in recent studies with vastly larger sample sizes. This suggests that the application of network methodology can aid in the robust identification of disease-related genes in genetic datasets that would otherwise be limited by statistical power. Although genetic studies of diseases such as AD, Type 2 diabetes, and several cancers have analyzed sample sizes up to several hundred thousand, there remain a wide array of diseases and phenotypes for which the identification of candidate genes by GWAS has been hampered by sample size. Therefore, the extension of our methodology to these phenotypes with more limited samples could vastly aid in expanding their genetic understanding.

In order to improve our understanding of the general effectiveness of the MRD approach, we extended the application of our methodology to the abundant number of phenotype gene sets available in Flybase [144]. Although Flybase

does not contain the genetic data required to test the integration aspect of our strategy, the wide variety of phenotype sets certainly allows for testing the network approaches that were applied in previous chapters in the context of AD.

An additional motivation of this research was to further the understanding of the biological basis for progression of AD-related pathology in the brain. In chapter 3, we determined that the co-expression of the RAD genes appears to be related to the known pattern of regional progression of AD within the human brain. A logical next step is to identify biological differences between brain regions that could explain this behavior. Recently, several studies reported measurable differences in cell type compositions between regions of the adult human brain [114].

Additionally, there is increasing support that microglial and immune cells serve a central role in AD progression [145]. Collectively this supports the idea that a cell type-related factor may contribute to the differences in expression of the RAD genes observed in chapter 3. Finally, in order to understand how relevant factors such as aging and the presence of AD can influence the conclusions from analyses of healthy brains (see chapter 3), here we extend our methodology to a dataset with a wider variety of clinical information for each subject.

Methods

Acquisition and Mapping of Flybase Phenotype Sets to the Human Protein-Protein Interaction Network

The Flybase phenotype database was accessed to acquire new gene sets for additional testing of the MRD approach [144]. In order to ensure compatibility with the existing networks established in prior chapters, the fly genes associated with phenotypes needed to be matched to their closest human gene orthologs. This matching was accomplished using the DRSC Integrative Ortholog Prediction Tool (DIOPT) [146]. Briefly, the DIOPT is a consensus strategy that utilizes the high number of ortholog prediction approaches. Each mapping algorithm was applied to each fly gene and the human gene ortholog that was predicted most commonly across all approaches was selected and was assigned a score based upon the number of algorithms that agreed with the mapping. Orthologs with a DIOPT score of ≥ 3 were deemed high quality [146]. After applying the DIOPT approach to the genes in the Flybase phenotype set, 6,300 phenotypes were matched with a minimum of two human gene orthologs.

Computing the Median Ranking by Diffusion of Flybase Phenotypes

The human-mapped Flybase phenotypes were filtered to ensure phenotypes being tested had adequate representation in the human PPI network that was assembled and described in Chapter 2. All phenotypes that had a minimum of five genes in the human PPI network were retained for further analysis. A total of

3100 phenotypes remained after this filtering step. Next, the MRD of each phenotype was computed within the human PPI network, using the methodology described in Chapter 2. Briefly, MRD is a leave one out style approach that determines the overall proximity of a set of genes within a network normalized to account for the distribution of scores for all genes. The frequencies of different ranges of MRD were illustrated in a histogram and several phenotypes were highlighted to represent either essential (cell lethality and DNA repair defectiveness) or AD-associated phenotypes (oxidative stress and premature or defective aging).

Cell Type Deconvolution in Late and Early Stage Regions of AD

Cell type deconvolution as implemented in the CellMix R package [147] was applied to the ABA gene expression data, described in chapter 3, in each region using gene markers for microglia (*TLR2*, *CX3CR1*, *IL1A*), neurons (*STMN2*, *SYN1*, *SYT1*, *GAD1*, *CCK*), and astrocytes (*GFAP*, *ALDH1L1*, *AQP4*, *GJA1*, *SOX9*) [148-151] in order to obtain frequencies for each cell type in each brain sub-region. The regions were then partitioned into two groups corresponding to early and late stage AD regions as described in chapter 3. Regions were sorted by computing the Shannon entropy among cell types in each region [152]. Briefly, entropy is a measure of the variability of a region. Regions with an equal mixture

of many cell types would therefore have high entropy, and regions that are dominated by mostly one cell type have low entropy.

Determining Changes in MRC of RAD Genes in Brains Due to Aging and AD

In order to determine the possible effects AD and other AD-related risk factors (such as age) on the MRC of the RAD genes, the approach described above was applied to an independent dataset from the Mount Sinai Brain Bank (MSBB). The MSBB dataset contains gene expression measurements for as many as 21 sub-regions of the brain per individual and is comprised of approximately 40 AD cases and 40 controls. The MSBB dataset was accessed via the publicly available AMP-AD web portal on Synapse ([synapse.org](https://www.synapse.org)). Individuals who had measurements in five or fewer regions or were > 90 years of age were excluded since all ages above 90 are reported as 90 to maintain anonymity of subjects, and mislabeled ages would sharply interfere with fitting a linear model. A total of 32 individuals remained after applying these filtering criteria. For each remaining individual, the absolute correlation was computed for all possible gene pairs across available brain regions. Then the MRC of the RAD genes was determined for each of these individual correlation networks. Finally, the resulting MRCs were regressed against the age and AD status of the individuals.

Results

Flybase Phenotypes Have High MRDs in Human PPI Network

After applying the MRD approach for each qualifying Flybase phenotype, the majority of the gene sets had many inner-set connections between gene members. Nearly perfect (≥ 0.98) MRDs were observed for small gene sets ($n < 10$), and this result was not surprising since the genes in these small sets often form densely connected neighborhoods or linear chains in the human PPI network. Most of the remaining gene sets had MRDs between 0.85 and 0.90. The MRDs for the AD-related phenotypes (**Figure 4.1**) were calculated as 0.75 (oxidative stress), 0.82 (premature aging), 0.83 (defective aging), and 0.9 (defective DNA repair) which is lower than the average of Flybase phenotypes in general (**Figure 4.2**).

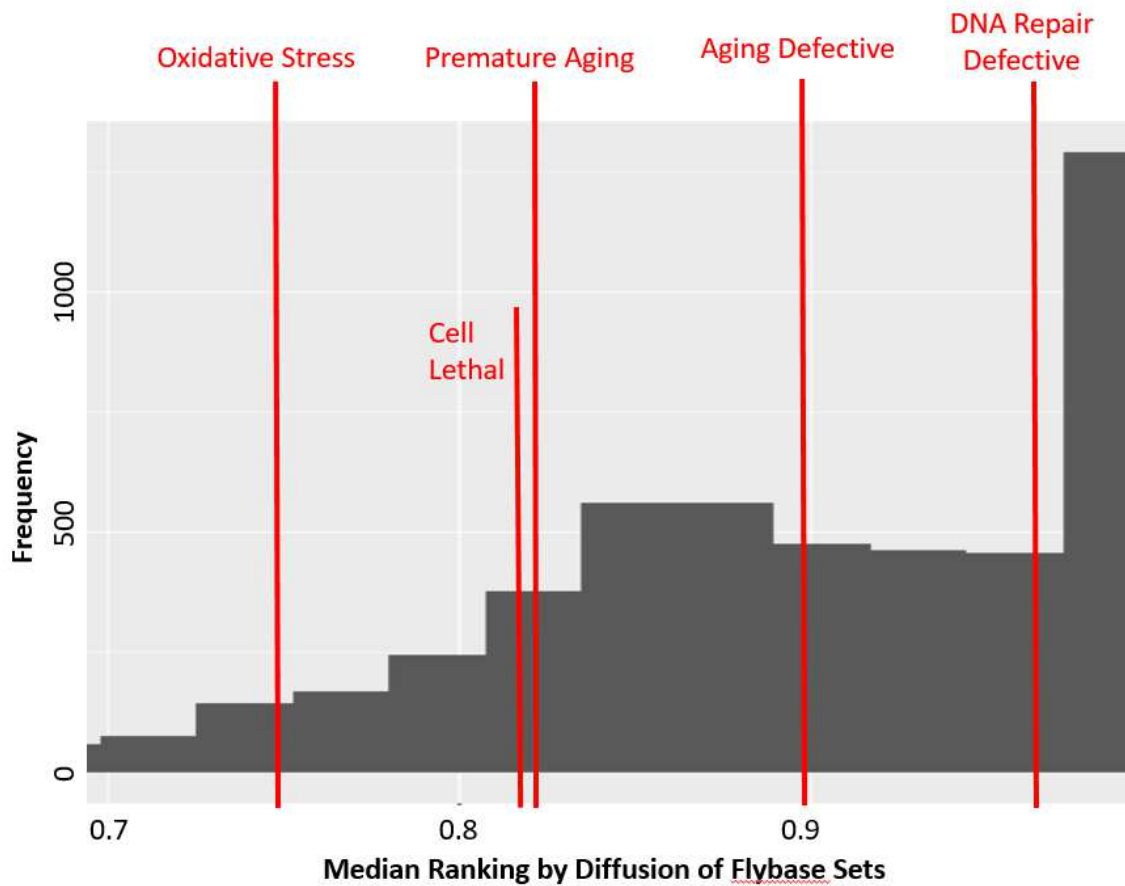


Figure 4.1: Overview of MRDs of Important AD and Survival Phenotypes.

The MRDs of the AD-related phenotypes used in chapter 3 (Oxidative Stress, Premature Aging, and Aging Defective) as well as two representative survival phenotypes (Cell Lethality, DNA Repair) were examined. The MRDs of the selected sets (highlighted in red) were above a random set, but notably at or below average of the Flybase sets as a whole with the exception of DNA repair.

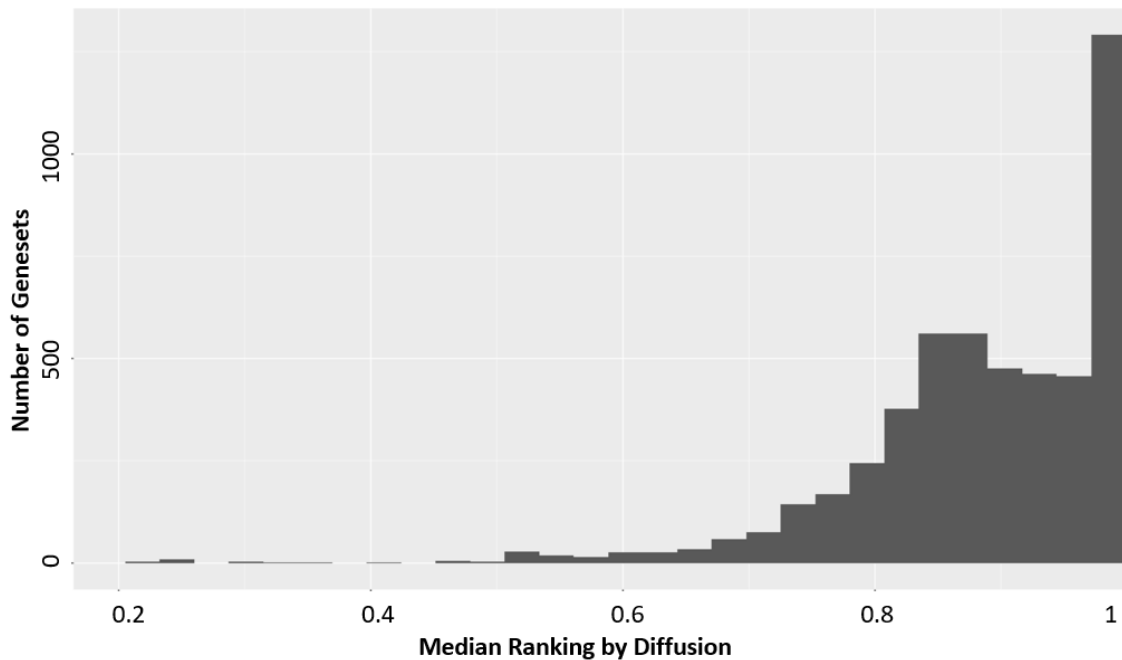


Figure 4.2: Flybase phenotype sets have high MRD in human PPI network.

6,297 phenotype gene sets were mapped to the human PPI network. The above histogram depicts the frequency of MRD values across these sets. Most of the phenotype sets from Flybase tended to have MRD above the random model of 0.5, with a high number being small sets of very closely clustered genes resulting in a large number of sets having MRDs that were in the 0.98 to 1 range.

Cell Types Compositions are Notably Different between Late and Early Stage

Regions

Deconvolution was applied to the ABA data in order to identify potential cell type differences between the various regional co-expression networks. Overall the early stage regions tended to be composed predominantly of neurons, with a much smaller fraction of microglia and astrocytes (**Figure 4.3**). By contrast, the late regions displayed a more balanced representation of these three cell types (**Figure 4.4**).

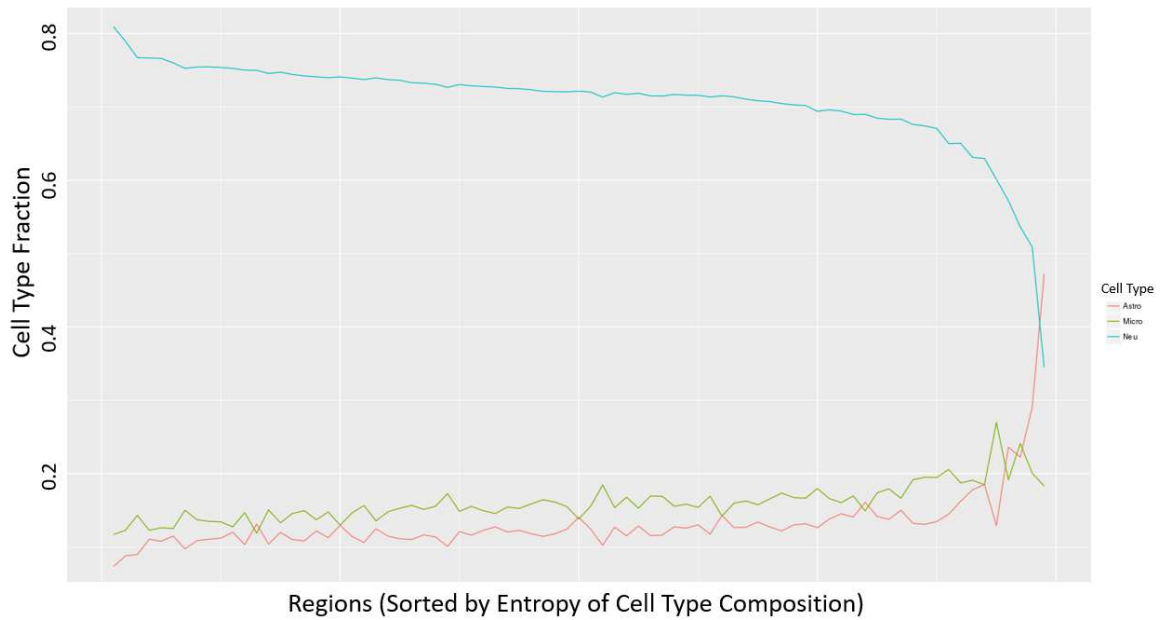


Figure 4.3: Cell Type Compositions of the Early Stage Regions: The cell type composition of 96 early stage regions was determined using deconvolution. Each column corresponds to a single region, and the sum of the three cell types for each region sum to one. Neurons (blue) were the dominant cell type in the vast majority of the early stage regions. Microglia (green) and Astrocytes (red) were slightly represented in most regions.

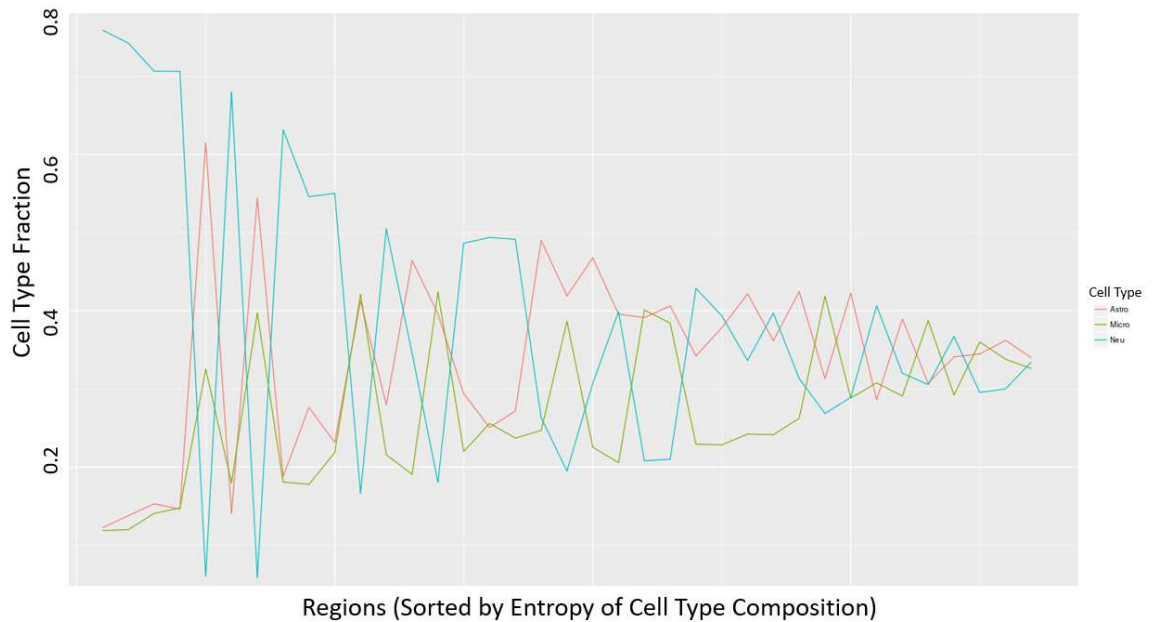


Figure 4.4: Cell Type Compositions of the Late Stage Regions. The cell type compositions of 37 early stage regions were determined using deconvolution. Each column corresponds to a single region, and the sum of the three cell types for each region sum to one. While neurons (blue) tended to be the most common, there are also a notable number of regions in which microglia (green) or astrocytes (red) were the dominant cell type. Overall the late stage regions display much higher variability in cell type in comparison to the early stage regions (**Figure 4.3**).

RAD Genes Have High MRC in Healthy and Young Individuals

The MRD of the RAD genes was determined in each individual, and regressed against their age and AD status. At high ages, there does not appear to be a meaningful distinction between AD cases and control (**See Figure 4.5**). However, at younger ages, there appears to be a much larger separation. Due to limited sample size, it cannot be determined if this trend is robust, given that the 95% confidence intervals (dark gray in **Figure 4.5**) for the slopes of the regression lines clearly overlap.

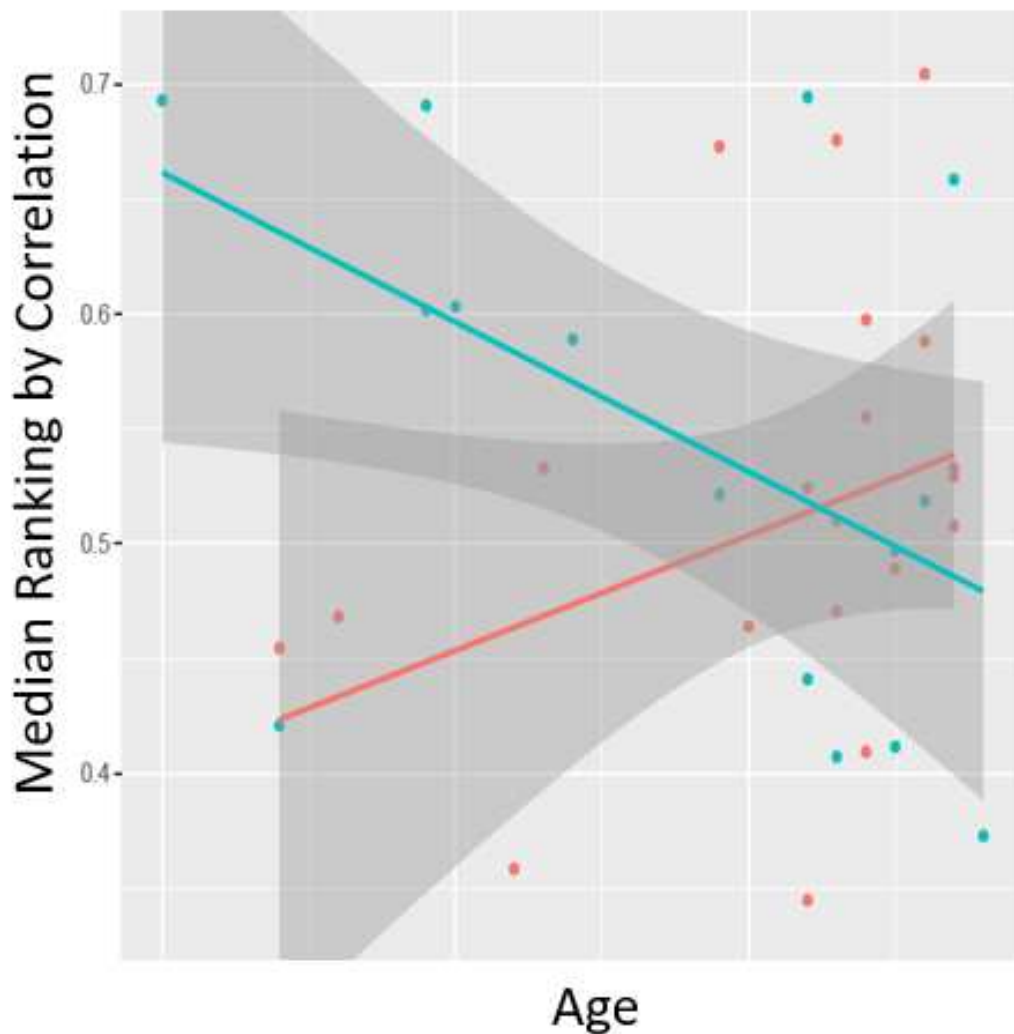


Figure 4.5: The Effects of Age and AD on the MRC of the RAD Genes. The MRC was computed within each individual in the MSBB dataset. AD cases (red) tended to have a much lower MRC for the RAD genes than age equivalent controls (blue). At higher ages, the MRC of the RAD genes tended to be indistinguishable in either group.

Discussion

In the earlier chapters, methodology for identifying novel disease-associations to genes was applied to an AD disease set. Ideally, this approach should be applied to curated datasets for other diseases in order to generalize its utility. However, publicly available data for many other diseases are either outdated or compiled using less stringent criteria for establishing genes as associated with disease. The assembly of a RAD gene set carried out using a manual assembly approach (see Chapter 2) required literature curation of several decades of research and thus may not be practical for some diseases. A more efficient strategy would be to apply our methodology to phenotype sets that can be assembled more efficiently using high throughput knockout screening in a suitable model organism such as *Drosophila* or *c. elegans*.

There are several explanations for our observation of a very high MRD of Flybase phenotypes on average (**Figure 4.1**). First, phenotypes in Flybase correspond to very specific human traits that can be screened in flies and, thus, may be more closely related than disease sets such as the RAD set (described in chapter 2) which may contain genes fulfilling a wider variety of biological roles. Whereas RAD genes may span a larger portion of the PPI network, Flybase phenotypes are likely localized to very specific areas of the interaction network. Second, there is a tautology between model organism screening and the addition of interactions to human PPI networks, i.e. when a small set of genes are shown

to be functionally related in animal models, the human orthologs of these genes will be tested subsequently and more likely be added to PPI databases than other genes.

It is surprising that a phenotype as detrimental as cell lethality had a lower MRD than an average Flybase phenotype (**Figure 4.2**). This may be due in part to the size of gene sets which can vary widely based upon the number and complexity of biological pathways related to the trait. In addition, some extreme phenotypes such as lethality are less specific because they can be induced by the shutdown of many essential processes and are therefore not localized to any particular pathway.

Genes with cell type specific expression can serve as markers that form the basis for several deconvolution approaches which approximate cell type fractions based upon the relative expression of these genes. Marker-based strategies have the advantage of being applicable *in silico* to any expression dataset without requiring the original biological samples for lab-based testing, which enabled us to determine the cell types underlying the ABA data. Single cell sequencing technologies have identified numerous markers that can be used for a wide variety of phenotypes.

The results we observe after deconvolution suggest that the underlying cell types of the brain regions could be responsible for the changes in clustering of the RAD

genes that was previously observed in chapter 3. The early stage regions, where the MRC of the RAD genes was slightly lower, are dominantly composed of neuronal cell types (**Figure 4.3**). The later stage regions, where the MRC of the RAD genes was higher, are a balanced mixture of the three tested cell types (**Figure 4.4**). This would suggest that the RAD genes tend to have stronger co-expression to one another in microglia and astrocytes than in neurons. It is also worth noting that cell type seems to be a fairly good indicator of either early or late stage regions. The regions of the brain that are most susceptible to AD in its early phases tend to be predominantly composed of neurons. The regions which are not affected until much later in AD progression, have a much smaller fraction of neurons. Given that AD is by definition a neurodegenerative disease, this would seem consistent with what is known biologically. There may also be protective effects resulting from the higher levels of astrocytes and microglial cells that insulate these regions from AD-related damage early on in the disease progression.

We observed that a reasonable separation occurs in MRD between young AD cases and controls (**Figure 4.5**). As individuals age however, this separation diminishes (**Figure 4.5**). This could be for a variety of reasons. Age is the most critical risk factor for AD, given that AD cases tend to not begin until around age 65 (except for rare familial early onset cases), after which point the prevalence

rate of AD increases rapidly to nearly 40% by age 80 [153]. Therefore, healthy older individuals still share aging-related biological changes with the older AD individuals that might mask and overwhelm any disease related differences.

It is notable that young and healthy individuals tended to have the highest MRCs of the RAD genes in the MSBB (**Figure 4.5**) and the ABA (chapter 3) datasets. Another key observation is that the MRC was greater in brain sub-regions which are spared early in the disease. Both of these observations are in agreement that co-expression of RAD genes is strongest in regions or individuals that are resilient to developing AD. RAD genes may simply be poorly co-expressed in AD individuals as a consequence of the underlying damage to the brain specimens. However, it would be highly informative if it was determined that the direction of causation is the opposite in that a breakdown in RAD gene co-expression elevates AD risk. If the direction of causation in this relationship can be determined in future studies, it could greatly improve our understanding of the therapeutic benefits of influencing RAD gene expression.

Chapter 5: Conclusions and Future Projects

The work presented in the chapters of this thesis provide further support for the benefit of multi-omic integration in the study of phenotypes and disease. Often times the difficulty of appropriately incorporating additional data types into an analysis discourages doing so. However, in this work we have demonstrated a very intuitive and simple to implement framework for re-prioritizing analysis results from existing statistical approaches that does not require any substantial changes to the existing methodology. The specific emphasis of this work has been on re-prioritization of genes from GWAS, however the framework can just as easily be applied to any other form of gene-level statistics relating to disease, such as differential expression analysis, expression Quantitative Trait Loci (eQTL) amongst many others. Several of the top genes proposed by the various applications of our integration approach have been replicated in very recent studies involving substantially larger sample sizes which demonstrates the robustness multi-omic analyses can achieve. Reaching massive sample sizes in sequencing studies is a costly and time-consuming endeavor, and so it stands to reason that making the best use of all presently available resources is an efficient exercise in the interim.

There are several clear opportunities for further development of this methodology that have not yet been fully explored. One such option is to determine the

potential benefits of simultaneously utilizing both the PPI and regional co-expression networks within a single analysis (See **Figure 5.1**). Biologically, it is not necessarily the case that two genes which interact necessarily need to be co-expressed. For example, genes in the blue region would correspond to such genes which interact but are not expressed in the brain, but rather are active in alternative tissues. By contrast, purple would be genes which are co-expressed in the brain, but do not have known protein interactions potentially because they have not been tested in proper environmental conditions.

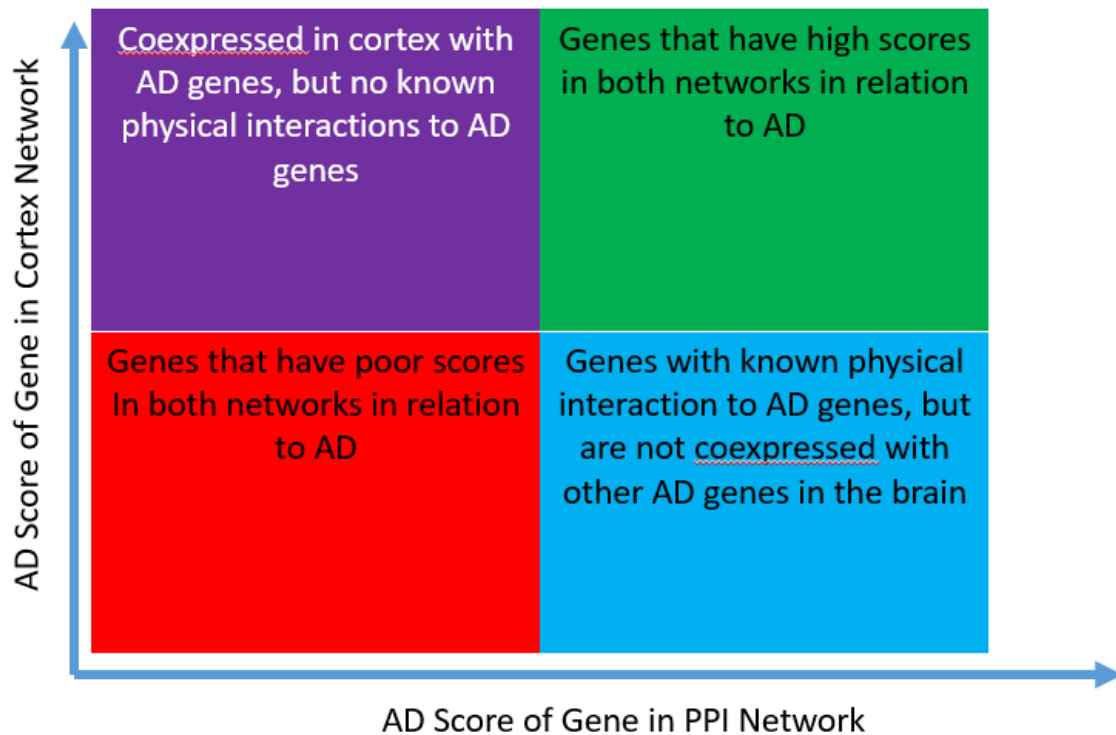


Figure 5.1: Overview of Possible Combinations of Protein Interactions and Co-Expression. Two different types of networks were constructed in this work. First, a protein interaction network (X-axis) and second, co-expression networks in the brain (Y-axis). It is not necessarily the case that genes would receive similar scores in both of these networks, as highlighted here.

Genes which are similarly ranked in both networks are arguably the most simple to interpret, however this should not lead to overlooking the potential information that can be obtained from following up on the cases of rankings which are discrepant between the PPI and coexpression networks. The development of tissue specific protein interactions network has been a growing area of research that seeks to clarify this issue, such as the recently developed TissueNet database [154]. Many protein-protein interactions are known to be heavily influenced by the environment in which they occur [154]. Therefore, the general PPI network (chapter 2) could contain interactions which do not occur in the essential areas of the brain where AD pathology is most active (chapter 3). The usage of tissue specific PPI data to further refine general networks to be more appropriate for a particular disease is therefore one potential improvement upon our existing methodology. However, there will still likely be genes that have discrepant rankings between a tissue/region specific PPI and a coexpression network of the relevant tissue. Many steps are involved in the activation of a biological pathway, usually beginning with a change in the cellular environment that activates a signaling cascade. The chain of interactions that occurs as part of a signaling cascade tend to resemble a ripple effect. So whereas a PPI network will generally represent all interactions underlying a pathway that would occur over time as the pathway is activating, a regional coexpression network reflects a static snapshot of the expression landscape at a given point in time.

Therefore this should be kept in mind during the design of any approach that attempts to harmonize all the earlier chapters of this work.

Another potential improvement to the methodology is related to how the seeded disease genes (the RAD genes, for example) are represented. Initially, we have treated all genes in the RAD set as having an equally important role in AD.

Genetically, this is known to be false given that *APOE* explains more of the heritability of late onset AD than any of the other known RAD genes [6]. It could therefore be optimal to weight the individual RAD genes based upon some metric of their statistical significance to AD. However, there are a tremendous number of factors that need to be accounted for when doing this such as sample size, appropriate correction for confounding variable, ethnic-specific effects, amongst many others. The choice to leave the RAD genes unweighted allowed us to focus on gauging the general effectiveness of the approach, but exploring different options for the initialization of the method would certainly be ideal in the future.

Finally, there are many different pathways that underlie diseases as biologically complex as AD. Therefore, it may make sense to divide the RAD gene set into subsets that reflect specific known biology. This however can be difficult, given that overall there are a high number of proposed AD-related pathways, and it isn't fully clear how each individual RAD gene corresponds to them. Clustering can be done to determine subsets of the RAD genes automatically, but this can

be heavily influenced by the choice of algorithm and parameters. In our initial exploration of simple clustering approaches, we found that each sub-cluster did appear enriched for specific AD-related processes, such as aging, inflammation, and neurodegeneration. This style of approach could therefore allow for a more pathway-driven investigation into disease than we are considering presently.

References

1. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell*. 1993;72(6):971-983. Epub 1993/03/26. doi: 10.1016/0092-8674(93)90585-e. PubMed PMID: 8458085.
2. Sandhoff K. Variation of beta-N-acetylhexosaminidase-pattern in Tay-Sachs disease. *FEBS Letters*. 1969;4(4):351-354. Epub 1969/08/01. doi: 10.1016/0014-5793(69)80274-7. PubMed PMID: 11947222.
3. Steinberg MH, Sebastiani P. Genetic modifiers of sickle cell disease. *American Journal of Hematology*. 2012;87(8):795-803. Epub 2012/05/30. doi: 10.1002/ajh.23232. PubMed PMID: 22641398; PubMed Central PMCID: PMC4562292.
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-753. Epub 2009/10/09. doi: 10.1038/nature08494. PubMed PMID: 19812666; PubMed Central PMCID: PMC4562292.
5. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*. 2017;169(7):1177-86. Epub 2017/06/18. doi: 10.1016/j.cell.2017.05.038. PubMed PMID: 28622505; PubMed Central PMCID: PMC5536862.
6. Ridge PG, Mukherjee S, Crane PK, Kauwe JS, Alzheimer's Disease Genetics C. Alzheimer's disease: analyzing the missing heritability. *PLoS One*. 2013;8(11):e79771. Epub 2013/11/19. doi: 10.1371/journal.pone.0079771. PubMed PMID: 24244562; PubMed Central PMCID: PMC3820606.
7. Fouss F, Francoise K, Yen L, Pirotte A, Saerens M. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural Networks*. 2012;31:53-72. Epub 2012/04/14. doi: 10.1016/j.neunet.2012.03.001. PubMed PMID: 22497802.
8. Lafferty J, Lebanon G. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*. 2005;6:129-63. PubMed PMID: WOS:000236328800005.
9. Norris JR. *Markov chains*. 1st pbk. ed. Cambridge, UK ; New York: Cambridge University Press; 1998. xvi, 237 p.

10. Jordan MI. Learning in graphical models. Cambridge, Mass.: MIT Press; 1999. vii, 634 p.
11. Marcoulides GA. The elements of statistical learning: Data mining, inference and prediction. *Structural Equation Modeling*. 2004;11(1):150-151. doi: 10.1207/S15328007sem1101_10. PubMed PMID: WOS:000221781000010.
12. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*. 1998;30(1-7):107-117. doi: 10.1016/S0169-7552(98)00110-X. PubMed PMID: WOS:000073360600013.
13. Jackson MO. Social and economic networks. Princeton, NJ: Princeton University Press; 2008. xiii, 504 p.
14. Ideker T, Nussinov R. Network approaches and applications in biology. *PLoS Computational Biology*. 2017;13(10):e1005771. Epub 2017/10/13. doi: 10.1371/journal.pcbi.1005771. PubMed PMID: 29023447; PubMed Central PMCID: PMC5638228.
15. Mezlini AM, Goldenberg A. Incorporating networks in a probabilistic graphical model to find drivers for complex human diseases. *PLoS Computational Biology*. 2017;13(10):e1005580. Epub 2017/10/13. doi: 10.1371/journal.pcbi.1005580. PubMed PMID: 29023450; PubMed Central PMCID: PMC5638204.
16. Guven-Maiorov E, Tsai CJ, Nussinov R. Structural host-microbiota interaction networks. *PLoS Computational Biology*. 2017;13(10):e1005579. Epub 2017/10/13. doi: 10.1371/journal.pcbi.1005579. PubMed PMID: 29023448; PubMed Central PMCID: PMC5638203.
17. Schadt EE. Molecular networks as sensors and drivers of common human diseases. *Nature*. 2009;461(7261):218-223. doi: 10.1038/nature08454. PubMed PMID: WOS:000269654600035.
18. Krogan NJ, Lippman S, Agard DA, Ashworth A, Ideker T. The Cancer Cell Map Initiative: Defining the hallmark networks of cancer. *Molecular Cell*. 2015;58(4):690-698. doi: 10.1016/j.molcel.2015.05.008. PubMed PMID: WOS:000355154000015.
19. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezchnikov AA, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell*. 2013;153(3):707-720. Epub 2013/04/30. doi: 10.1016/j.cell.2013.03.030. PubMed PMID: 23622250; PubMed Central PMCID: PMC3677161.

20. Parikshak NN, Gandal MJ, Geschwind DH. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews. Genetics*. 2015;16(8):441-458. Epub 2015/07/08. doi: 10.1038/nrg3934. PubMed PMID: 26149713; PubMed Central PMCID: PMC4699316.
21. Gaiteri C, Mostafavi S, Honey CJ, De Jager PL, Bennett DA. Genetic variants in Alzheimer disease – molecular and brain network approaches. *Nature Reviews. Neurology*. 2016;12(7):413-427. Epub 2016/06/11. doi: 10.1038/nrneurol.2016.84. PubMed PMID: 27282653; PubMed Central PMCID: PMC45017598.
22. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*. 2001;292(5518):929-934. Epub 2001/05/08. doi: 10.1126/science.292.5518.929. PubMed PMID: 11340206.
23. Readhead B, Haure-Mirande JV, Funk CC, Richards MA, Shannon P, Haroutunian V, et al. Multiscale analysis of independent Alzheimer's cohorts finds disruption of molecular, genetic, and clinical networks by human herpesvirus. *Neuron*. 2018;99(1):64–82.e7. doi: 10.1016/j.neuron.2018.05.023. PubMed PMID: WOS:000438378100010.
24. Hormozdiari F, Penn O, Borenstein E, Eichler EE. The discovery of integrated gene networks for autism and related disorders. *Genome Research*. 2015;25(1):142-154. doi: 10.1101/gr.178855.114. PubMed PMID: WOS:000347373200013.
25. Parikshak NN, Gandal MJ, Geschwind DH. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews Genetics*. 2015;16(8):441-458. doi: 10.1038/nrg3934. PubMed PMID: WOS:000358075900007.
26. Liu M, Liberzon A, Kong SW, Lai WR, Park PJ, Kohane IS, et al. Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genetics*. 2007;3(6):e96. Epub 2007/06/19. doi: 10.1371/journal.pgen.0030096. PubMed PMID: 17571924; PubMed Central PMCID: PMC1904360.
27. Novarino G, Fenstermaker AG, Zaki MS, Hofree M, Silhavy JL, Heiberg AD, et al. Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders. *Science*. 2014;343(6170):506-511. Epub 2014/02/01. doi: 10.1126/science.1247363. PubMed PMID: 24482476; PubMed Central PMCID: PMC4157572.

28. Lander ES. Initial impact of the sequencing of the human genome. *Nature*. 2011;470(7333):187-197. Epub 2011/02/11. doi: 10.1038/nature09792. PubMed PMID: 21307931.
29. Hayes B. Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). *Methods in Molecular Biology*. 2013;1019:149-169. Epub 2013/06/13. doi: 10.1007/978-1-62703-447-0_6. PubMed PMID: 23756890.
30. Alzheimer's A. 2011 Alzheimer's disease facts and figures. *Alzheimers & Dementia*. 2011;7(2):208-244. Epub 2011/03/19. doi: 10.1016/j.jalz.2011.02.004. PubMed PMID: 21414557.
31. Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature*. 1991;349(6311):704-706. Epub 1991/02/21. doi: 10.1038/349704a0. PubMed PMID: 1671712.
32. Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature*. 1995;375(6534):754-760. Epub 1995/06/29. doi: 10.1038/375754a0. PubMed PMID: 7596406.
33. Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, Pettingell WH, et al. Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science*. 1995;269(5226):973-977. Epub 1995/08/18. PubMed PMID: 7638622.
34. Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PH, Pericak-Vance MA, Joo SH, et al. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*. 1993;43(8):1467-1472. Epub 1993/08/01. PubMed PMID: 8350998.
35. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*. 2013;45(12):1452-1458. Epub 2013/10/29. doi: 10.1038/ng.2802. PubMed PMID: 24162737; PubMed Central PMCID: PMC3896259.
36. Rogaeva E, Meng Y, Lee JH, Gu Y, Kawarai T, Zou F, et al. The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. *Nature Genetics*. 2007;39(2):168-177. Epub 2007/01/16. doi: 10.1038/ng1943. PubMed PMID: 17220890; PubMed Central PMCID: PMC3896259.

37. Jia P, Zhao Z. Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Human Genetics*. 2014;133(2):125-138. Epub 2013/10/15. doi: 10.1007/s00439-013-1377-1. PubMed PMID: 24122152; PubMed Central PMCID: PMC3943795.
38. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*. 2011;21(7):1109-1121. Epub 2011/05/04. doi: 10.1101/gr.118992.110. PubMed PMID: 21536720; PubMed Central PMCID: PMC3129253.
39. Beyer A, Bandyopadhyay S, Ideker T. Integrating physical and genetic maps: from genomes to interaction networks. *Nature Reviews. Genetics*. 2007;8(9):699-710. doi: 10.1038/nrg2144. PubMed PMID: WOS:000248882400016.
40. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Research*. 2011;39(Database issue):D712-717. Epub 2010/11/13. doi: 10.1093/nar/gkq1156. PubMed PMID: 21071422; PubMed Central PMCID: PMC3013724.
41. Razick S, Magklaras G, Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*. 2008;9:405. Epub 2008/10/01. doi: 10.1186/1471-2105-9-405. PubMed PMID: 18823568; PubMed Central PMCID: PMC3013724.
42. Rolland T, Tasan M, Charlotiaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-scale map of the human interactome network. *Cell*. 2014;159(5):1212-1226. Epub 2014/11/25. doi: 10.1016/j.cell.2014.10.050. PubMed PMID: 25416956; PubMed Central PMCID: PMC4266588.
43. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*. 2003;19 Suppl 1:i197-204. Epub 2003/07/12. PubMed PMID: 12855458.
44. Kolaczyk ED. Statistical Analysis of Network Data: Methods and Models. *Statistical Analysis of Network Data: Methods and Models*. 2009:1-386. doi: 10.1007/978-0-387-88146-1. PubMed PMID: WOS:000266858000011.
45. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*. 2007;3:140. Epub 2007/10/18. doi: 10.1038/msb4100180. PubMed PMID: 17940530; PubMed Central PMCID: PMC2063581.

46. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, et al. Variations in DNA elucidate molecular networks that cause disease. *Nature*. 2008; 452(7186):429-35. Epub 2008/03/18. doi: 10.1038/nature06757. PubMed PMID: 18344982; PubMed Central PMCID: PMCPMC2841398.
47. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*. 2015;47(6):569-576. Epub 2015/04/29. doi: 10.1038/ng.3259. PubMed PMID: 25915600; PubMed Central PMCID: PMCPMC4828725.
48. Hardy J. Amyloid, the presenilins and Alzheimer's disease. *Trends in Neurosciences*. 1997;20(4):154-159. Epub 1997/04/01. PubMed PMID: 9106355.
49. Jun G, Naj AC, Beecham GW, Wang LS, Buross J, Gallins PJ, et al. Meta-analysis confirms CR1, CLU, and PICALM as Alzheimer disease risk loci and reveals interactions with APOE genotypes. *Archives of Neurology*. 2010;67(12):1473-84. Epub 2010/08/11. doi: 10.1001/archneurol.2010.201. PubMed PMID: 20697030; PubMed Central PMCID: PMCPMC3048805.
50. Pullabhatla V, Roberts AL, Lewis MJ, Mauro D, Morris DL, Odhams CA, et al. De novo mutations implicate novel genes in Systemic Lupus Erythematosus. *Human Molecular Genetics*. 2017. Epub 2017/11/28. doi: 10.1093/hmg/ddx407. PubMed PMID: 29177435.
51. Scheuner D, Eckman C, Jensen M, Song X, Citron M, Suzuki N, et al. Secreted amyloid beta-protein similar to that in the senile plaques of Alzheimer's disease is increased in vivo by the presenilin 1 and 2 and APP mutations linked to familial Alzheimer's disease. *Nature Medicine*. 1996; 2(8):864-870. Epub 1996/08/01. PubMed PMID: 8705854.
52. Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo MM, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nature Genetics*. 2011; 43(5):429-435. Epub 2011/04/05. doi: 10.1038/ng.803. PubMed PMID: 21460840; PubMed Central PMCID: PMCPMC3084173.
53. Mez J, Chung J, Jun G, Kriegel J, Bourlas AP, Sherva R, et al. Two novel loci, COBL and SLC10A2, for Alzheimer's disease in African Americans. *Alzheimer's & Dementia*. 2017;13(2):119-129. Epub 2016/10/23. doi: 10.1016/j.jalz.2016.09.002. PubMed PMID: 27770636; PubMed Central PMCID: PMCPMC5318231.

54. Jun G, Asai H, Zeldich E, Drapeau E, Chen C, Chung J, et al. PLXNA4 is associated with Alzheimer disease and modulates tau phosphorylation. *Annals of Neurology*. 2014;76(3):379-392. Epub 2014/07/22. doi: 10.1002/ana.24219. PubMed PMID: 25043464; PubMed Central PMCID: PMC4830273.
55. Rohn TT, Head E, Nesse WH, Cotman CW, Cribbs DH. Activation of caspase-8 in the Alzheimer's disease brain. *Neurobiology of Disease*. 2001; 8(6):1006-1016. Epub 2001/12/14. doi: 10.1006/nbdi.2001.0449. PubMed PMID: 11741396.
56. Beecham GW, Hamilton K, Naj AC, Martin ER, Huentelman M, Myers AJ, et al. Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. *PLoS Genetics*. 2014;10(9): e1004606. Epub 2014/09/05. doi: 10.1371/journal.pgen.1004606. PubMed PMID: 25188341; PubMed Central PMCID: PMC4154667.
57. Escott-Price V, Bellenguez C, Wang LS, Choi SH, Harold D, Jones L, et al. Gene-wide analysis detects two new susceptibility genes for Alzheimer's disease. *PLoS One*. 2014;9(6):e94661. Epub 2014/06/13. doi: 10.1371/journal.pone.0094661. PubMed PMID: 24922517; PubMed Central PMCID: PMC4055488.
58. Otani Y, Yamaguchi Y, Sato Y, Furuichi T, Ikenaka K, Kitani H, et al. PLD4 is involved in phagocytosis of microglia: expression and localization changes of PLD4 are correlated with activation state of microglia. *PLoS One*. 2011;6(11):e27544. Epub 2011/11/22. doi: 10.1371/journal.pone.0027544. PubMed PMID: 22102906; PubMed Central PMCID: PMC3216956.
59. Cruchaga C, Kauwe JS, Harari O, Jin SC, Cai Y, Karch CM, et al. GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease. *Neuron*. 2013;78(2):256-268. Epub 2013/04/09. doi: 10.1016/j.neuron.2013.02.026. PubMed PMID: 23562540; PubMed Central PMCID: PMC3664945.
60. Chung J, Wang X, Maruyama T, Ma Y, Zhang X, Mez J, et al. Genome-wide association study of Alzheimer's disease endophenotypes at prediagnosis stages. *Alzheimer's & Dementia*. 2017. Epub 2017/12/24. doi: 10.1016/j.jalz.2017.11.006. PubMed PMID: 29274321.
61. Jun GR, Chung J, Mez J, Barber R, Beecham GW, Bennett DA, et al. Transethnic genome-wide scan identifies novel Alzheimer's disease loci. *Alzheimer's & Dementia*. 2017;13(7):727-38. Epub 2017/02/12. doi:

- 10.1016/j.jalz.2016.12.012. PubMed PMID: 28183528; PubMed Central PMCID: PMC5496797.
62. Ruiz A, Heilmann S, Becker T, Hernandez I, Wagner H, Thelen M, et al. Follow-up of loci from the International Genomics of Alzheimer's Disease Project identifies TRIP4 as a novel susceptibility gene. *Translational Psychiatry*. 2014;4:e358. Epub 2014/02/06. doi: 10.1038/tp.2014.2. PubMed PMID: 24495969; PubMed Central PMCID: PMC3944635.
 63. Wetzel-Smith MK, Hunkapiller J, Bhangale TR, Srinivasan K, Maloney JA, Atwal JK, et al. A rare mutation in UNC5C predisposes to late-onset Alzheimer's disease and increases neuronal cell death. *Nature Medicine*. 2014;20(12):1452-1457. Epub 2014/11/25. doi: 10.1038/nm.3736. PubMed PMID: 25419706; PubMed Central PMCID: PMC34301587.
 64. Hoekstra EJ, Mesman S, de Munnik WA, Smidt MP. LMX1B is part of a transcriptional complex with PSPC1 and PSF. *PLoS One*. 2013;8(1):e53122. Epub 2013/01/12. doi: 10.1371/journal.pone.0053122. PubMed PMID: 23308148; PubMed Central PMCID: PMC3537735.
 65. Sims R, van der Lee SJ, Naj AC, Bellenguez C, Badarinarayan N, Jakobsdottir J, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nature Genetics*. 2017;49(9):1373-1384. Epub 2017/07/18. doi: 10.1038/ng.3916. PubMed PMID: 28714976; PubMed Central PMCID: PMC5669039.
 66. Jun G, Ibrahim-Verbaas CA, Vronskaya M, Lambert JC, Chung J, Naj AC, et al. A novel Alzheimer disease locus located near the gene encoding tau protein. *Molecular Psychiatry*. 2016;21(1):108-117. Epub 2015/03/18. doi: 10.1038/mp.2015.23. PubMed PMID: 25778476; PubMed Central PMCID: PMC34573764.
 67. Narain Y, Yip A, Murphy T, Brayne C, Easton D, Evans JG, et al. The ACE gene and Alzheimer's disease susceptibility. *Journal of Medical Genetics*. 2000;37(9):695-697. Epub 2000/09/09. PubMed PMID: 10978362; PubMed Central PMCID: PMC1734696.
 68. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, et al. TREM2 variants in Alzheimer's disease. *New England Journal of Medicine*. 2013;368(2):117-127. Epub 2012/11/16. doi: 10.1056/NEJMoa1211851. PubMed PMID: 23150934; PubMed Central PMCID: PMC3631573.

69. Huang KL, Marcora E, Pimenova AA, Di Narzo AF, Kapoor M, Jin SC, et al. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. *Nature Neuroscience*. 2017;20(8):1052-1061. Epub 2017/06/20. doi: 10.1038/nn.4587. PubMed PMID: 28628103; PubMed Central PMCID: PMC5759334.
70. Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature*. 2014;505(7484):550-554. Epub 2013/12/18. doi: 10.1038/nature12825. PubMed PMID: 24336208; PubMed Central PMCID: PMC4050701.
71. Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA: The Journal of the American Medical Association*. 1997;278(16):1349-1356. Epub 1997/10/29. PubMed PMID: 9343467.
72. Logue MW, Schu M, Vardarajan BN, Farrell J, Bennett DA, Buxbaum JD, et al. Two rare AKAP9 variants are associated with Alzheimer's disease in African Americans. *Alzheimer's & Dementia*. 2014;10(6):609-618 e11. Epub 2014/08/31. doi: 10.1016/j.jalz.2014.06.010. PubMed PMID: 25172201; PubMed Central PMCID: PMC4253055.
73. Karch CM, Ezerskiy LA, Bertelsen S, Alzheimer's Disease Genetics C, Goate AM. Alzheimer's Disease Risk Polymorphisms Regulate Gene Expression in the ZCWPW1 and the CELF1 Loci. *PLoS One*. 2016;11(2):e0148717. Epub 2016/02/27. doi: 10.1371/journal.pone.0148717. PubMed PMID: 26919393; PubMed Central PMCID: PMC4769299.
74. Smola AJ, Kondor R. Kernels and regularization on graphs. *Learning Theory and Kernel Machines*. 2003;2777:144-158. doi: 10.1007/978-3-540-45167-9_12. PubMed PMID: WOS:000185937100011.
75. Kondor R, Lafferty J. Diffusion Kernels on Graphs and Other Discrete Input Spaces. In: University CM, editor. 2002.
76. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*. 2015;43(Database issue):D447-452. Epub 2014/10/30. doi: 10.1093/nar/gku1003. PubMed PMID: 25352553; PubMed Central PMCID: PMC4383874.

77. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. 1974;36(2):111-147. PubMed PMID: WOS:A1974U703600001.
78. Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buross J, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nature Genetics*. 2011; 43(5):436-+. doi: 10.1038/ng.801. PubMed PMID: WOS:000289972600012.
79. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genetics*. 2006;2(12):2074-2093. doi: ARTN e190 10.1371/journal.pgen.0020190. PubMed PMID: WOS:000243482100012.
80. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nature Genetics*. 2016;48(10):1284-1287. doi: 10.1038/ng.3656. PubMed PMID: WOS:000384391600026.
81. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*. 2016;48(10):1279-1283. doi: 10.1038/ng.3643. PubMed PMID: WOS:000384391600025.
82. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*. 2012;44(8):955-+. doi: 10.1038/ng.2354. PubMed PMID: WOS:000306854700025.
83. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics*. 2015;31(5):782-784. doi: 10.1093/bioinformatics/btu704. PubMed PMID: WOS:000352268500026.
84. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*. 2007;81(3):559-575. doi: 10.1086/519795. PubMed PMID: WOS:000249128200012.
85. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research*. 2014;42(D1):D756-D763. doi: 10.1093/nar/gkt1114. PubMed PMID: WOS:000331139800112.
86. Wang K, Li MY, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*.

2010;38(16). doi: ARTN e164 10.1093/nar/gkq603. PubMed PMID:
WOS:000281720500004.

87. Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Human Molecular Genetics*. 2007;16(1):36-49. Epub 2006/12/01. doi: 10.1093/hmg/ddl438. PubMed PMID: 17135278; PubMed Central PMCID: PMCPMC2270437.
88. Segre AV, Groop L, Mootha VK, Daly MJ, Altshuler D, Consortium D, et al. Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. *PLoS Genetics*. 2010;6(8). doi: ARTN e1001058 10.1371/journal.pgen.1001058. PubMed PMID: WOS:000281383800014.
89. Efron B, Tibshirani R. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*. 1997; 92(438):548-560. doi: Doi 10.2307/2965703. PubMed PMID: WOS:A1997XE29600020.
90. Cortes C, Vapnik V. Support-Vector Networks. *Machine Learning*. 1995; 20(3):273-297. doi: 10.1023/A:1022627411411. PubMed PMID: WOS:A1995RX35400003.
91. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(43):15545-15550. doi: 10.1073/pnas.0506580102. PubMed PMID: WOS:000232929400051.
92. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. 2000;28(1):27-30. doi: DOI 10.1093/nar/28.1.27. PubMed PMID: WOS:000084896300007.
93. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. *Nucleic Acids Research*. 2016;44(D1):D481-487. Epub 2015/12/15. doi: 10.1093/nar/gkv1351. PubMed PMID: 26656494; PubMed Central PMCID: PMCPMC4702931.
94. Dreyfus SE. An Appraisal of Some Shortest-Path Algorithms. *Operations Research*. 1969;17(3):395-412. doi: DOI 10.1287/opre.17.3.395. PubMed PMID: WOS:A1969D425700002.

95. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26(17):2190-191. doi: 10.1093/bioinformatics/btq340. PubMed PMID: WOS:000281738900017.
96. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. *Trends in Genetics*. 1997; 13(4):163. Epub 1997/04/01. PubMed PMID: 9097728.
97. Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding CM, Cantor CR, et al. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2004;101(9):2888-2893. doi: 10.1073/pnas.0307326101. PubMed PMID: WOS:000220065300045.
98. Solito E, Sastre M. Microglia function in Alzheimer's disease. *Frontiers in Pharmacology*. 2012;3:14. Epub 2012/03/01. doi: 10.3389/fphar.2012.00014. PubMed PMID: 22363284; PubMed Central PMCID: PMC3277080.
99. de la Monte SM, Wands JR. Alzheimer's disease is type 3 diabetes-evidence reviewed. *Journal of Diabetes Science and Technology*. 2008; 2(6):1101-1113. Epub 2009/11/04. doi: 10.1177/193229680800200619. PubMed PMID: 19885299; PubMed Central PMCID: PMC32769828.
100. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*. 2015; 47(2):106-114. Epub 2014/12/17. doi: 10.1038/ng.3168. PubMed PMID: 25501392; PubMed Central PMCID: PMC4444046.
101. Blatti C, Sinha S. Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks. *Bioinformatics*. 2016;32(14):2167-2175. Epub 2016/05/07. doi: 10.1093/bioinformatics/btw151. PubMed PMID: 27153592; PubMed Central PMCID: PMC4937193.
102. Rubinsztein DC. The roles of intracellular protein-degradation pathways in neurodegeneration. *Nature*. 2006;443(7113):780-786. Epub 2006/10/20. doi: 10.1038/nature05291. PubMed PMID: 17051204.
103. Elbaz A, Carcaillon L, Kab S, Moisan F. Epidemiology of Parkinson's disease. *Revue Neurologique (Paris)*. 2016;172(1):14-26. Epub 2016/01/01. doi: 10.1016/j.neurol.2015.09.012. PubMed PMID: 26718594.

104. Rowland LP. Amyotrophic lateral sclerosis. *Current Opinion in Neurology*. 1994;7(4):310-315. Epub 1994/08/01. PubMed PMID: 7952238.
105. Wenk GL. Neuropathologic changes in Alzheimer's disease: potential targets for treatment. *Journal of Clinical Psychiatry*. 2006;67 Suppl 3:3-7; quiz 23. Epub 2006/05/03. PubMed PMID: 16649845.
106. Browne SE, Bowling AC, MacGarvey U, Baik MJ, Berger SC, Muqit MM, et al. Oxidative damage and metabolic dysfunction in Huntington's disease: selective vulnerability of the basal ganglia. *Annals of Neurology*. 1997;41(5):646-653. Epub 1997/05/01. doi: 10.1002/ana.410410514. PubMed PMID: 9153527.
107. Hohenfeld C, Werner CJ, Reetz K. Resting-state connectivity in neurodegenerative disorders: Is there potential for an imaging biomarker? *NeuroImage. Clinical*. 2018;18:849-870. Epub 2018/06/08. doi: 10.1016/j.nicl.2018.03.013. PubMed PMID: 29876270; PubMed Central PMCID: PMC5988031.
108. Gur RE, McGrath C, Chan RM, Schroeder L, Turner T, Turetsky BI, et al. An fMRI study of facial emotion processing in patients with schizophrenia. *American Journal of Psychiatry*. 2002;159(12):1992-1999. Epub 2002/11/27. doi: 10.1176/appi.ajp.159.12.1992. PubMed PMID: 12450947.
109. Yao Z, Wang L, Lu Q, Liu H, Teng G. Regional homogeneity in depression and its relationship with separate depressive symptom clusters: a resting-state fMRI study. *Journal of Affective Disorders*. 2009;115(3):430-438. Epub 2008/11/15. doi: 10.1016/j.jad.2008.10.013. PubMed PMID: 19007997.
110. Mu Y, Gage FH. Adult hippocampal neurogenesis and its role in Alzheimer's disease. *Molecular Neurodegeneration*. 2011;6:85. Epub 2011/12/24. doi: 10.1186/1750-1326-6-85. PubMed PMID: 22192775; PubMed Central PMCID: PMC3261815.
111. Ross CA, Poirier MA. Protein aggregation and neurodegenerative disease. *Nature Medicine*. 2004;10 Suppl:S10-17. Epub 2004/07/24. doi: 10.1038/nm1066. PubMed PMID: 15272267.
112. Pierson E, Consortium GT, Koller D, Battle A, Mostafavi S, Ardlie KG, et al. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS Computational Biology*. 2015;11(5):e1004220. Epub 2015/05/15. doi: 10.1371/journal.pcbi.1004220. PubMed PMID: 25970446; PubMed Central PMCID: PMC4430528.

113. Mostafavi S, Gaiteri C, Sullivan SE, White CC, Tasaki S, Xu J, et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nature Neuroscience*. 2018;21(6):811-819. Epub 2018/05/29. doi: 10.1038/s41593-018-0154-9. PubMed PMID: 29802388; PubMed Central PMCID: PMC6599633.
114. Collado-Torres L, Burke EE, Peterson A, Shin J, Straub RE, Rajpurohit A, et al. Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. *Neuron*. 2019;103(2):203-216 e8. Epub 2019/06/09. doi: 10.1016/j.neuron.2019.05.013. PubMed PMID: 31174959.
115. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*. 2010;26(4):445-455. Epub 2010/01/08. doi: 10.1093/bioinformatics/btp713. PubMed PMID: 20053841; PubMed Central PMCID: PMC6599633.
116. Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, 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*. 2010;53(3):1051-1063. Epub 2010/01/27. doi: 10.1016/j.neuroimage.2010.01.042. PubMed PMID: 20100581; PubMed Central PMCID: PMC2892122.
117. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*. 2010;6(1):e1000641. Epub 2010/01/22. doi: 10.1371/journal.pcbi.1000641. PubMed PMID: 20090828; PubMed Central PMCID: PMC2797085.
118. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*. 2017; 18(9):551-562. Epub 2017/06/14. doi: 10.1038/nrg.2017.38. PubMed PMID: 28607512.
119. Lancour D, Naj A, Mayeux R, Haines JL, Pericak-Vance MA, Schellenberg GD, et al. One for all and all for One: Improving replication of genetic studies through network diffusion. *PLoS Genetics*. 2018;14(4):e1007306. Epub 2018/04/24. doi: 10.1371/journal.pgen.1007306. PubMed PMID: 29684019; PubMed Central PMCID: PMC6599633.
120. Team RC. R: A language and environment for statistical computing. 2019.

121. Sheiner LB, Grasela TH. An Introduction to Mixed Effect Modeling - Concepts, Definitions, and Justification. *Journal of Pharmacokinetics and Biopharmaceutics*. 1991;19(3):S11-S24. doi: Doi 10.1007/Bf01371005. PubMed PMID: WOS:A1991FU32600002.
122. Braak H, Braak E. Neuropathological staging of Alzheimer-related changes. *Acta Neuropathologica*. 1991;82(4):239-259. Epub 1991/01/01. PubMed PMID: 1759558.
123. Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, et al. FlyBase 2.0: the next generation. *Nucleic Acids Research*. 2019; 47(D1):D759-D765. Epub 2018/10/27. doi: 10.1093/nar/gky1003. PubMed PMID: 30364959; PubMed Central PMCID: PMC6323960.
124. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*. 2007;104(21):8685-8690. Epub 2007/05/16. doi: 10.1073/pnas.0701361104. PubMed PMID: 17502601; PubMed Central PMCID: PMC1885563.
125. Dianza A, Carlier MF, Stradal TE, Didry D, Frittoli E, Confalonieri S, et al. Eps8 controls actin-based motility by capping the barbed ends of actin filaments. *Nature Cell Biology*. 2004;6(12):1180-1188. Epub 2004/11/24. doi: 10.1038/ncb1199. PubMed PMID: 15558031.
126. Fulga TA, Elson-Schwab I, Khurana V, Steinhilb ML, Spires TL, Hyman BT, et al. Abnormal bundling and accumulation of F-actin mediates tau-induced neuronal degeneration in vivo. *Nature Cell Biology*. 2007;9(2):139-148. Epub 2006/12/26. doi: 10.1038/ncb1528. PubMed PMID: 17187063.
127. Wang YT, Huang CC, Lin YS, Huang WF, Yang CY, Lee CC, et al. Conditional deletion of Eps8 reduces hippocampal synaptic plasticity and impairs cognitive function. *Neuropharmacology*. 2017;112(Pt A):113-123. Epub 2016/10/25. doi: 10.1016/j.neuropharm.2016.07.021. PubMed PMID: 27450093.
128. Petyuk VA, Chang R, Ramirez-Restrepo M, Beckmann ND, Henrion MYR, Piehowski PD, et al. The human brainome: network analysis identifies HSPA2 as a novel Alzheimer's disease target. *Brain*. 2018;141(9):2721-2739. Epub 2018/08/24. doi: 10.1093/brain/awy215. PubMed PMID: 30137212; PubMed Central PMCID: PMC6136080.
129. Campanella C, Pace A, Caruso Bavisotto C, Marzullo P, Marino Gammazza A, Buscemi S, et al. Heat Shock Proteins in Alzheimer's Disease: Role and Targeting. *International Journal of Molecular Sciences*. 2018;19(9). Epub

2018/09/12. doi: 10.3390/ijms19092603. PubMed PMID: 30200516;
PubMed Central PMCID: PMC6163571.

130. Hamos JE, Oblas B, Pulaski-Salo D, Welch WJ, Bole DG, Drachman DA. Expression of heat shock proteins in Alzheimer's disease. *Neurology*. 1991;41(3):345-350. Epub 1991/03/01. doi: 10.1212/wnl.41.3.345. PubMed PMID: 2005999.
131. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Author Correction: Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nature Genetics*. 2019;51(9):1423-1424. Epub 2019/08/17. doi: 10.1038/s41588-019-0495-7. PubMed PMID: 31417202.
132. Colciaghi F, Marcello E, Borroni B, Zimmermann M, Caltagirone C, Cattabeni F, et al. Platelet APP, ADAM 10 and BACE alterations in the early stages of Alzheimer disease. *Neurology*. 2004;62(3):498-501. Epub 2004/02/12. doi: 10.1212/01.wnl.0000106953.49802.9c. PubMed PMID: 14872043.
133. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nature Genetics*. 2019;51(3):414-430. Epub 2019/03/02. doi: 10.1038/s41588-019-0358-2. PubMed PMID: 30820047; PubMed Central PMCID: PMC6463297.
134. Milton NG. Amyloid-beta binds catalase with high affinity and inhibits hydrogen peroxide breakdown. *Biochemical Journal*. 1999;344 Pt 2:293-296. Epub 1999/11/24. PubMed PMID: 10567208; PubMed Central PMCID: PMC61220643.
135. Habib LK, Lee MT, Yang J. Inhibitors of catalase-amyloid interactions protect cells from beta-amyloid-induced oxidative stress and toxicity. *Journal of Biological Chemistry*. 2010;285(50):38933-38943. Epub 2010/10/07. doi: 10.1074/jbc.M110.132860. PubMed PMID: 20923778; PubMed Central PMCID: PMC2998107.
136. Yang SS, Zhang R, Wang G, Zhang YF. The development prospect of HDAC inhibitors as a potential therapeutic direction in Alzheimer's disease. *Translational Neurodegeneration*. 2017;6:19. Epub 2017/07/14. doi: 10.1186/s40035-017-0089-1. PubMed PMID: 28702178; PubMed Central PMCID: PMC5504819.

137. Janczura KJ, Volmar CH, Sartor GC, Rao SJ, Ricciardi NR, Lambert G, et al. Inhibition of HDAC3 reverses Alzheimer's disease-related pathologies in vitro and in the 3xTg-AD mouse model. *Proceedings of the National Academy of Sciences of the United States of America*. 2018;115(47):E11148-E11157. Epub 2018/11/07. doi: 10.1073/pnas.1805436115. PubMed PMID: 30397132; PubMed Central PMCID: PMC6255210.
138. Agis-Balboa RC, Pavelka Z, Kerimoglu C, Fischer A. Loss of HDAC5 impairs memory function: implications for Alzheimer's disease. *Journal of Alzheimer's Disease*. 2013;33(1):35-44. Epub 2012/08/24. doi: 10.3233/JAD-2012-121009. PubMed PMID: 22914591.
139. Kilgore M, Miller CA, Fass DM, Hennig KM, Haggarty SJ, Sweatt JD, et al. Inhibitors of class 1 histone deacetylases reverse contextual memory deficits in a mouse model of Alzheimer's disease. *Neuropsychopharmacology*. 2010;35(4):870-880. Epub 2009/12/17. doi: 10.1038/npp.2009.197. PubMed PMID: 20010553; PubMed Central PMCID: PMC3055373.
140. Chung J, Zhang X, Allen M, Wang X, Ma Y, Beecham G, et al. Genome-wide pleiotropy analysis of neuropathological traits related to Alzheimer's disease. *Alzheimer's Research & Therapy*. 2018;10(1):22. Epub 2018/02/21. doi: 10.1186/s13195-018-0349-z. PubMed PMID: 29458411; PubMed Central PMCID: PMC5819208.
141. Hampel H, Mesulam MM, Cuello AC, Farlow MR, Giacobini E, Grossberg GT, et al. The cholinergic system in the pathophysiology and treatment of Alzheimer's disease. *Brain*. 2018;141(7):1917-1933. Epub 2018/06/01. doi: 10.1093/brain/awy132. PubMed PMID: 29850777; PubMed Central PMCID: PMC6022632.
142. Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and Biobanking*. 2015;13(5):307-308. Epub 2015/10/21. doi: 10.1089/bio.2015.29031.hmm. PubMed PMID: 26484569; PubMed Central PMCID: PMC4692118.
143. Wang M, Roussos P, McKenzie A, Zhou X, Kajiwara Y, Brennand KJ, et al. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Medicine*. 2016;8(1):104. Epub 2016/11/02. doi: 10.1186/s13073-016-0355-3. PubMed PMID: 27799057; PubMed Central PMCID: PMC5088659.
144. Drysdale RA, Crosby MA, FlyBase C. FlyBase: genes and gene models. *Nucleic Acids Research*. 2005;33(Database issue):D390-395. Epub

- 2004/12/21. doi: 10.1093/nar/gki046. PubMed PMID: 15608223; PubMed Central PMCID: PMC540000.
145. Mandrekar-Colucci S, Landreth GE. Microglia and inflammation in Alzheimer's disease. *CNS & Neurological Disorders Drug Targets*. 2010;9(2):156-167. Epub 2010/03/09. doi: 10.2174/187152710791012071. PubMed PMID: 20205644; PubMed Central PMCID: PMC540000.
146. Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, et al. An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*. 2011;12:357. Epub 2011/09/02. doi: 10.1186/1471-2105-12-357. PubMed PMID: 21880147; PubMed Central PMCID: PMC3179972.
147. Gaujoux R, Seoighe C. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics*. 2013;29(17):2211-2212. Epub 2013/07/05. doi: 10.1093/bioinformatics/btt351. PubMed PMID: 23825367.
148. Zhang Y, Sloan SA, Clarke LE, Caneda C, Plaza CA, Blumenthal PD, et al. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron*. 2016;89(1):37-53. Epub 2015/12/22. doi: 10.1016/j.neuron.2015.11.013. PubMed PMID: 26687838; PubMed Central PMCID: PMC4707064.
149. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *Journal of Neuroscience*. 2008;28(1):264-278. Epub 2008/01/04. doi: 10.1523/JNEUROSCI.4178-07.2008. PubMed PMID: 18171944; PubMed Central PMCID: PMC6671143.
150. Holtman IR, Raj DD, Miller JA, Schaafsma W, Yin Z, Brouwer N, et al. Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: a co-expression meta-analysis. *Acta Neuropathologica Communications*. 2015;3:31. Epub 2015/05/24. doi: 10.1186/s40478-015-0203-5. PubMed PMID: 26001565; PubMed Central PMCID: PMC4489356.
151. Li Z, Del-Aguila JL, Dube U, Budde J, Martinez R, Black K, et al. Genetic variants associated with Alzheimer's disease confer different cerebral cortex cell-type population structure. *Genome Medicine*. 2018;10(1):43. Epub 2018/06/09. doi: 10.1186/s13073-018-0551-4. PubMed PMID: 29880032; PubMed Central PMCID: PMC5992755.

152. Shannon CE. The mathematical theory of communication. 1963. M.D. Computing. 1997;14(4):306-317. Epub 1997/07/01. PubMed PMID: 9230594.
153. Barnes J, Dickerson BC, Frost C, Jiskoot LC, Wolk D, van der Flier WM. Alzheimer's disease first symptoms are age dependent: Evidence from the NACC dataset. *Alzheimer's & Dementia*. 2015;11(11):1349-1357. Epub 2015/04/29. doi: 10.1016/j.jalz.2014.12.007. PubMed PMID: 25916562; PubMed Central PMCID: PMC4619185.
154. Basha O, Barshir R, Sharon M, Lerman E, Kirson BF, Hekselman I, et al. The TissueNet v.2 database: A quantitative view of protein-protein interactions across human tissues. *Nucleic Acids Research*. 2017;45(D1):D427-D431. Epub 2016/12/03. doi: 10.1093/nar/gkw1088. PubMed PMID: 27899616; PubMed Central PMCID: PMC5210565.

Curriculum Vitae

