COMPARISON OF ENTROPY AND ENSEMBLE-BASED FEATURE SELECTION THROUGH NETWORK ANALYSIS OF ALZHEIMERS DISEASE-ASSOCIATED VARIANTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

SEVDA RAFATOV

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
BIOINFORMATICS

FEBRUARY 2022

Approval of the thesis:

ANALYSIS OF ALZHEIMERS DISEASE-ASSOCIATED VARIANTS FOR THE DISCOVERY OF AFFECTED BIOLOGICAL PATHWAYS

Submitted by **Sevda Rafatov** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Supervisor, **Health Informatics, METU** _____

Examining Committee Members:

Assist. Prof. Aybar Can Acar
Health Informatics Dept., METU _____

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics Dept., METU _____

Assoc. Prof. Tunca Doğan
Computer Engineering Dept., Hacettepe University _____

Date: **08.02.2022**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : SEVDA RAFATOV

Signature : _____

ABSTRACT

COMPARISION OF ENTROPY AND ENSEMBLE BASED FEATURE SELECTION THROUGH NETWORK ANALYSIS OF ALZHEIMERS DISEASE-ASSOCIATED VARIANTS

Rafatov, Sevda

MSc., Bioinformatics

Supervisor: Assist. Prof. Dr. Yeşim Aydın Son

February 2022, 120 pages

Alzheimer's Disease (AD) is a complex, progressive and irreversible brain disorder that slowly destroys memory and thinking skills and eventually loses the ability to do daily tasks. Our group is currently developing in-silico AD models in which genotyping and phenotyping data are integrated for the differential diagnosis Late-On-Set AD (LOAD) cases. Meta-analysis of four different LOAD data sets provided by ADNI and dbGAP, which includes the genotyping data of more than 5000 LOAD patients, is done. In this study, we provided the biological interpretation of the variants selected through two different approaches, namely entropy and ensemble modeling. First, the LOAD-associated variants are annotated for their genomic location, consequence, gene and protein products, and biological pathways. The protein-coding variants prioritized were selected for experimental validation based on their relationship with LOAD-related biological pathways after network, PPI, and enrichment analysis. For 32 variants, pyrosequencing primers were designed, and sequencing primers were optimized. As a part of the study, a case-control group with 43 LOAD diagnosed and 38 healthy participants were formed, and genotyping for the prioritized variants was completed. We have shown that machine learning models capture hidden, new, and informative patterns by considering nonlinear interactions where multiple variants determine the risk. Further analysis of interconnected networks for selected genes and proteins can identify affected biological pathways underlying the molecular etiology of AD susceptibility. Understanding the affected molecular pathways can reveal potential causative variants that lead to novel preventative therapeutics for AD.

Keywords: Alzheimer's Disease, biological networks, functional enrichment analysis

ÖZ

ALZHEİMER İLE İLİŞKİLİ VARYANTLARIN AĞ ANALİZİ ÜZERİNDEN ENTROPY VE ENSEMBLE BAZLI DEĞİŞKEN SEÇİMİNİN KARŞILAŞTIRILMASI

Rafatov, Sevda

Yüksek Lisans, Biyoenformatik

Tez Yöneticisi: Doç. Dr.Yeşim Aydın Son

Şubat 2022, 120 sayfa

Alzheimer Hastalığı (AD) , hafıza ve düşünme becerilerini yavaş yavaş yok eden ve sonunda günlük işleri yapma yeteneğini kaybetmeye neden olan karmaşık, ilerleyici ve geri dönüşü olmayan bir beyin hastalığıdır. Grubumuz şu anda, Geç Başlangıçlı Alzheimer Hastalığı (LOAD) vakalarının ayırıcı tanısı için genotipleme ve fenotipleme verilerinin entegre edildiği in-siliko AD modelleri geliştirmektedir. ADNI ve dbGAP tarafından sağlanan ve 5000'den fazla LOAD hastasının genotipleme verilerini içeren dört farklı LOAD veri setinin meta analizi devam etmektedir. Bu çalışmada, entropi ve ensembl modelleme olmak üzere iki farklı yaklaşımla seçilen varyantların biyolojik olarak yorumlanmasını sağladık. İlk olarak, LOAD ile ilişkili varyantlar, genomik konumları, sonuçları, gen ve protein ürünleri ve biyolojik yolakları ile ilişkilendirilmiştir. Öncelik verilen protein kodlama varyantları, ağ, PPI ve zenginleştirme analizinden sonra LOAD ile ilgili biyolojik yollar ile ilişkilerine dayalı olarak deneysel doğrulama için seçilmiştir. 32 varyant için pirosekanslama primerleri tasarlanarak ve pirosekanslama primerleri optimize edilmiştir. 43 LOAD ve 38 sağlıklı katılımcıdan oluşan vaka-kontrol grubunda öncelikli varyantlar için genotiplendirilmiştir. Makine öğrenimi modellerinin, birden çok değişkenin riski belirlediği doğrusal olmayan etkileşimleri göz önünde bulundurarak gizli, yeni ve bilgilendirici örüntüler tanımlanmıştır. Seçilen genler ve proteinler için birbirine bağlı ağların ileri analizleri, AD yatkınlığının moleküler etiyolojisinin altında yatan etkilenen biyolojik yolları tanımlamada yardımcı olabilir. Etkilenen moleküler yolakları anlamak, AD için yeni önleyici terapötiklere yol açabilecek potansiyel nedensel varyantları ortaya çıkarabildiği önerilmektir.

Anahtar Sözcükler: Alzheimer hastalığı, biyolojik ağlar, fonksiyonel zenginleştirme analizi

To My Family,

ACKNOWLEDGMENTS

First of all, I would like to express my gratitude to my adviser Dr. Yeşim Aydın Son. I am grateful to her for guiding me with patience, wisdom, and compassion during my master's degree. She set an example for me not only as a scientist but also with her humanity. I feel lucky to have had the opportunity to work with her. I would not have been able to do this work without her support, encouragement, and guidance.

Furthermore, I would like to express my gratitude to professors Dr.Aybar Can Acar and Dr. Tolga Can. The courses I took from them made me understand better what I was curious about and what I wanted to focus on in the future. I am grateful for their knowledge and for generously sharing it with us.

I want to express my gratitude to my colleagues Burcu Yaldız and Onur Erdoğan. They never hesitated to help whenever I needed some support. I thank them for being there for me whenever I needed some support, for their understanding and helpfulness. Moreover, I want to thank my dearest friends Aylin Şen, Hazal Özen, Kübra Altınok and Gizem Buldu. It was not an easy experience to write a thesis during a global pandemic. My dearest friends have always been by my side during this time, supported me with their compassion, love and encouraged me to be my best self.

Lastly, I would like to thank my family for their love, patience, and endless belief in me. I would not have been able to do this work if not for the peaceful and loving environment they provided for me. I am endlessly grateful for my mother, Nigar. Her love, support, and belief in me kept me going through tough times. I am endlessly thankful for my father, İsmail. I am grateful to him for teaching me the love of science, scientific curiosity, exploration, and creativity. Moreover, I want to thank my siblings Telman ve Sevinç. They are the light and joy of my life. I am grateful to them for being there for me at my best and my worst; life would not be bearable without them. Also, I owe a special thanks to my dearest Ali Fatih Kuloğlu, who has been by my side through all the stages of my undergrad and grad education. Supported and encouraged me with endless love and patience. I am endlessly grateful for your support and patience. I am thankful and blessed for having these people in my life.

I acknowledge the Scientific and Technological Research Council of Turkey (TUBİTAK), for which we conducted this thesis under the grant TUBİTAK ARDEB 1003 - 216S468.

The data/analyses presented in the current publication are based on the use of study data downloaded from the dbGaP web site, under dbGaP (phs000168.v1.p1 and phs000219.v1.p1), and ADNI (adni-info.org).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	Alzheimer’s Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
Aβ	B-Amyloid
BIND	The Biomelecular Interaction Network Database
DIP	Database of Interacting Proteins
DNA	Deoxyribonucleic acid
EFO	Experimental Factor Ontology
FDR	False Discovery Rate
GO	Gene Ontology
GRID	Global Research Identifier Database
GWAS	Genome-Wide Association Studies
HPO	Human Phenotype Ontology
HPRD	Human Protein Reference Database
KEGG	Kyoto Encyclopedia of Genes and Genomes
KG	Knowledge Graphs
LOAD	Late Onset Alzheimer’s Disease
MCL	Markov Cluster Algorithm
MINT	The Molecular Interaction Database
NCRAD	National Centralized Repository for Alzheimer's Disease
OMIM	Online Mendelian Inheritance in Man
PID	The Pathway Interaction Database
PPI	Protein-Protein Interaction Networks
RF	Random-Forest
RNA	Ribonucleic acid
SNP	Single Nucleotide Polymorphism
UniProt	The Universal Protein Resource

CHAPTER 1

1. INTRODUCTION

The most widespread form of neurodegenerative dementia is Alzheimer's disease, marked by progressive memory loss and cognitive, problem-solving, and language deficits (Cuyvers & Sleegers, 2016). Late-Onset Alzheimer's Disease (LOAD) is the most common type of dementia in the aging population, characterized by memory deterioration and other cognitive domains.

The epistatic nature of Alzheimer's disease and the complex genetic etiology of LOAD is still unclear, which restrains the early and differential diagnosis of LOAD. For that purpose, Genome-wide association studies (GWAS) are used by many researchers to find variants whose common occurrence signaled to AD phenotype (Bodily et al., 2016; Cuyvers & Sleegers; Tosto & Reitz, 2013; Lambert et al., 2013; Desikan et al., 2017; Escott-Price et al., 2014). GWAS allows exploration of the statistical interactions of individuals' variants, but the univariate analysis oversees interactions between variants. More sophisticated studies are needed to distinguish variants that collectively influence the phenotype (Bodily et al., 2016).

The machine learning algorithms can capture hidden, novel, and significant patterns considering nonlinear interactions between variants to understand the genetic predisposition for complex genetic disorders, where multiple variants determine the risk.

Our previous studies have used three different GWAS datasets provided by ADNI and the dbGAP's controlled accessed datasets from GenADA NCRAD initiatives. First statistical analysis is done by PLINK, and initial dimension reduction is completed by p-value filtering. Following GWAS, two-step Random Forest (RF-RF) modeling was implemented with 5-fold cross-validation (CV) using RANGER package in R. After PLINK-RF-RF analysis of LOAD GWAS datasets, a novel Ensemble and Entropy-based

methods for meta-analysis were optimized as a part of two other graduate studies completed within our group by Onur Erdoğan and Burcu Yaldız.

Based on their biological interpretation, this study compares the gene sets produced by these two meta-analyses approaches, Ensemble and Entropy. The network analysis approach to identify genes in Alzheimer's disease is currently studied in several works (Talwar et al.,2014; Rao et al., 2013; Rahman et al., 2019; Meng et al., 2020). There are several different methods to compare networks. Here, biological networks are constructed for the Ensemble and Entropy gene sets, and global parameters are compared to conclude the networks' connectivity to assess the usability of both meta-analysis methods.

The global parameters used for comparison are network density, average path length, and efficiency are compared by GeneMANIA based PPI network analysis and STRING network analysis. CROssBAR knowledge graphs show associations between genes and pathways for each gene set. Further, these associations are investigated by functional enrichment analysis of the gene sets. That led us to find novel correlations between our genes and Alzheimer's Disease. So, gene-AD networks are constructed to investigate the known associations of our genes in literature with Alzheimer's Disease.

Chapter 2 covers the effects of mutations on Alzheimer's disease phenotype, molecular etiology of Alzheimer's Disease, biological networks, methods to compare biological networks, pathway enrichment analysis, and knowledge graphs.

Chapter 3 explained how data was obtained and the methodology in detail. Firstly, an overview of the methods is given. After explaining how data is obtained, protein-protein interaction networks, GeneMANIA based PPI network analysis, network gene-training disease interaction networks, prediction of network genes-Alzheimer's disease interaction networks by Cytoscape are explained. Subsequently, building knowledge graphs by CROssBAR, pathway enrichment analysis, visualization, and functional enrichment analysis interpretation are explained.

In Chapter 4, the results are presented and explained in detail. This chapter explains protein-protein interaction networks, prediction of gene functions, network genes- training diseases networks, knowledge graphs, pathway-based enrichment analysis, network genes-Alzheimer's disease networks, and STRING enrichment analysis.

In Chapter 5, the genotyping constructed on the prioritized variants is explained in details.

In Chapter 6, we discussed our results and interpreted the importance of using a network-based approach to predict the biological meaningfulness of gene sets.

CHAPTER 2

2. LITERATURE REVIEW

2.1. Alzheimer’s Disease

A mutation is described as a change in the nucleotide sequence of an organism's genome, virus genome, or extrachromosomal DNA. Any change in DNA can cause different diseases such as Mendelian diseases or complex diseases. (Amberger, Bocchini & Hamosh, 2011). Alzheimer's disease is genetically complex, heritable, and the most common form of dementia, which increases as the human lifespan increases and is the cause of approximately 65% of all dementias, burdening both health and socioeconomic aspects in aging societies. Close to 50 million people worldwide live with Alzheimer's Disease (AD) (Bright Focus Foundation). These cases exhibit a late-onset complex genetic inheritance (Kamboh, 2018). Alzheimer's disease can be classified genetically into two types: a rare familial form that affects less than 1% of all patients and is inherited by autosomal dominant inheritance, and a sporadic multifactorial form that is assumed to be caused by a combination of environmental exposures and genetic susceptibility (Cuyvers & Sleegers, 2016). Besides memory loss, other cognitive dysfunctions develop in the AD process, amyloid plaques and neurofibrillary tangles accumulate in the brain, and cerebral atrophy, more commonly in the temporal region, is observed.

Thus far, some genes found to be implicated in the etiology of Alzheimer’s disease. Even though the use of molecular pathways in diagnosing Alzheimer's disease is unclear, several causes are believed to be involved in the disease's pathogenesis. (Moradifard et al., 2018). Genes that have been recently identified to be associated with AD are highly expressed in glial cells. Therefore, the effect of glial cells on the pathogenesis of Alzheimer's disease has recently been the focus of attention. Astrocytes are the most numerous cell type within the Central Nervous System (CNS). They are essential for the maintenance of brain homeostasis and neuronal protection. Astrocytes are essential in the synaptogenesis, the release of neurotransmitters, cognition, neuroinflammation, glycogen storage, formation of Blood-Brain Barrier (BBB), clearance of toxic substances such as glutamate excess and K+ spatial buffering, release of trophic factors for neurons and other

brain cells in both physiological and disease conditions (Volterra and Meldolesi, 2005; Hamby and Sofroniew, 2010; Kimelberg and Nedergaard, 2010; Barreto et al., 2011; Cabezas et al., 2012, 2014; Posada-Duque et al., 2014). Astrocytic activities constitute a three-party synapse with presynaptic and post-synaptic neurons, based on their intense relationship to the neurons at both molecular and morphological levels through their endfeet. (Perea et al., 2009; Perez-Alvarez and Araque, 2013; Coulter and Steinhauser, 2015). Astrocytic metabolic dysregulation in this aspect is a characteristic feature of neurodegenerative diseases like Alzheimer's Disease (AD) (Volterra and Meldolesi, 2005; Maragakis and Rothstein, 2006; Hamby and Sofroniew, 2010; Kimelberg and Nedergaard, 2010; Parpura et al., 2011).

It is believed that pathogenic Aβ and tau types can cause gliosis and neuroinflammation. In addition, it is thought that glial cells may regulate the pathogenesis of inflammation, Aβ, and tau. Neuropathologically, extracellular senile plaques containing β-amyloid (Aβ) and intracellular neurofibrillary tangles containing hyperphosphorylated tau protein are described (Harold et al., 2009).

The complex genetic etiology of LOAD is still unclear, limiting the early and differential diagnosis of LOAD. Making the differential diagnosis with other causes of dementia and making the differential diagnosis from age-related forgetfulness and mild cognitive impairment are often insufficient in making the diagnosis at an early stage; the diagnosis of the disease can be made partially with clinical evaluation and imaging methods. The definitive diagnosis of AD can be made by tissue diagnosis, that is, by brain biopsy or autopsy after the patient's death. Early differential diagnosis with early detection of cognitive decline is vital for improving disease management and even reversing symptoms.

So far, studies on Alzheimer's disease have focused on precursor proteins such as tau, presenilin, and amyloid and have been carried out through molecular mechanisms (Tan et al., 2019). It is observed that LOAD, which has a complex genetic transition, is 80% hereditary (Emahazion et al., 2001; Gatz et al., 2006), but studies on its molecular mechanisms are still ongoing. Today, early differential diagnosis of LOAD patients from aging-related dementia patients is not possible when they usually enter the clinic with dementia symptoms. This situation causes patients and their families to be late in planning treatment processes and other measures for AD.

Genome-wide association studies have been conducted to determine the genetic factors that cause Alzheimer's disease (Kamboh et al., 2012; Zhang et al., 2012; Mukherjee et al., 2014; Sherva et al., 2014). Variations associated with AD, especially APOE, can only explain about 25% of this complex genetic background (Musani et al., 2007; So et al.,

2011). The most critical LOAD risk factor is the first identified APOE ε4 allele. Genome-wide analysis studies conducted in the last ten years have identified candidate gene loci such as CLU, PICALM, CR1, BIN, ABCA7, EPHA1, CD33, CD2AP, ATP5H/KCTD2, and MS4A (Chouraki et al., 2014). As a result of the I-GAP study, the results were published in 2013 and performed in nearly 75,000 patients; 11 new loci, 19 in total, were identified (Lambert et al., 2013). In the second follow-up study, four of these candidates (DSG2; PTK2B; SORL1; SLC24A4) were highly significant in repeated studies (Ruiz et al., 2014).

Alzheimer's disease-associated new genes were discovered in the recent study by Prokopenko et al., 2021. 13 new candidate AD loci were identified, genes mapping to these loci are; FNBP1L, SEL1L, LINCOO298, PRKCH, C15ORF41, C2CD3, KIF2A, APC, LHX9, NALCN, CTNNA2, SYTL3, and CLSTN2.

In the study by Jansen et al. 2019, nine novel genes significantly associated with AD are published as; ADAMTS4, HESX1, CLNK, CNTNAP2, ADAM10, APH1B, KAT8, ALPK2, and AC074212.3. In the most recent study by de Rojas 2021, six variants associated with Alzheimer's disease risk near APP, CHRNE, PRKD3/NDUFAF7, PLCG2, and two exonic variants in the SHARPIN gene were discovered.

2.2. GWAS

Genome-Wide Association Studies (GWAS) explore the statistical association of SNPs in complex genetic disorders using high-dimensional datasets (Cantor, R. M., Lange, K., & Sinsheimer, J.S. (2010), Marees, et al., 2018). GWAS have shown success in many studies in identifying associated single-nucleotide polymorphism (SNP) profiles in diseases with complex genetic structure.

The Genome-Wide Association Studies (GWAS) approach tests a univariate hypothesis. GWAS does not evaluate the potential relevance of each genetic marker and assigns statistical significance based on statistical assumptions of data distribution. First, these associations are identified by single-focus approaches, where each SNP is tested individually for the association. Although this standard method provides information on novel loci for a particular complex disease, some limitations exist. GWAS does not consider the genetic interactions of each biomarker; this means that the expected specific mutation or non-mutation combinations are not observed together.

However, the univariate approach alone cannot explain genetic inheritance alone in most complex diseases. It quickly identifies diseases related to a single gene/mutation/variation

caused by genetic differences that substantially affect association studies. However, it is insufficient for complex genetic diseases that have weak effects and are caused by many variations. Identifying the effects of the epistatic interaction of multiple genetic variations on specific genes plays an important role in diagnosing and treating these complex human diseases. As both the number of genetic variations and the nature of interactions increase, such complex phenotypic traits are characterized by a high level of unpredictability.

Genetic risk factors, which cause susceptibility to diseases with complex genetic inheritance, interact together, although they are in different genetic regions. Complex interactions such as SNP-SNP, gene-gene, and gene-environment interactions that may contribute to the resolution of missing heritability are ignored during GWAS (Cordell, 2009). Moreover, although it reveals the top-ranking significant correlations associated with a particular disease, the method does not provide a predictive model that presents statistically significant biomarkers.

A challenging aspect of genomic studies is understanding the biological impact of inherited genetic variations in DNA structure between individuals and the molecular etiology of a complex disease. Single nucleotide variations (SNVs) have received much attention in disease prediction among genetic variations. SNVs that are the source of individual differences can be used as biomarkers and located on or near genes associated with particularly complex diseases. Discovering SNV biomarkers at different loci may improve early diagnosis accuracy and prevent these diseases through clinical decision-making.

The large size of the genetic data used in GWAS studies has necessitated the use of data mining methods in order to identify the interactions of genetic changes with the disease, prioritize their relationships and use them effectively in clinical applications to support decision making in diagnosis. It is a critical and promising area for data mining methods to identify representative SNPs to predict variability between individuals with the phenotype of interest.

2.3. Ensemble and Entropy methods

2.3.1. Ensemble method

The ensemble method is a machine learning technique that combines several base models to produce one optimal predictive model. Ensemble methods usually produce more accurate solutions than a single model would. Before our study, information from three different data mining models from different datasets is integrated with the ensemble method proposed by Onur Erdoğan et al., which is one of the first studies which uses

higher space (SNVs to genomic locations on bands) ensemble techniques to integrate multi models that offer significant information after GWAS and RF analysis. An Ensemble scoring algorithm is proposed to calculate information regarding bands that emphasize disease-related variations. This is the first ensemble methodology proposed by Onur Erdoğan et al. for different datasets whose attributes also differ. In this point, the novelty of this scoring algorithm is to ensemble multi-data mining methods in terms of knowledge extraction for each dataset of Alzheimer's disease. The output of multi-platform prediction models can be used as prior knowledge to merge all information to reach the posterior ensemble model with model minimization to detect Alzheimer's disease-related complex structure of genetic variations and other risk factors. (Erdoğan O. et al., 2022)

2.3.2. Entropy method

Shannon entropy is a non-linear function that measures the uncertainty of random variables (Shannon, 1948). Methods based on Shannon entropy measure the strength of predicting a combination of variables in explaining an event. Different combinations of variables are said to interact when the power of the joint prediction ability of this combination in describing an event is larger than the sum of the individual prediction abilities of these variables (Cover and Thomas, 1991). Fan. et al. proposed an information gain approach based on mutual information and used an interaction-information gain approach for three-way interactions. The prioritized SNPs in each dataset (ADNI, GenADA, NCRAD) are investigated using this entropy-based three-way interaction information method (3WII).

Two-way mutual information and three-way interaction information are entropy-based methods that measure the interaction between two markers and the information common to all three attributes (Yaldız B. et al., 2022).

D=0 denotes the disease status of an individual for healthy individuals and D=1 for affected ones in a case-control study design. The difference between the mutual information in the affected population and the general population is defined as information gain:

$$IG(X,Y\backslash D)=I(X,Y\backslash D)-I(X,Y)$$

Interaction information gain of markers X, Y, and Z are defined similarly:

$$IIG(X,Y,Z\backslash D)=I(X,Y,Z\backslash D)-I(X,Y,Z)$$

Information gain-based test statistic (T_{IG}) is calculated by dividing IG or IIG by a specific normalization factor of variance Λ. The resulting test statistics is centrally chi-square distributed with 1 degree of freedom under the null hypothesis that the markers are independent of the disease.

The test statistic for each dataset was calculated using the interaction information gain for prioritized SNPs. Since we look for the interactions common to all three variants that cannot be explained by two-way mutual information gain, the triplets with SNP combinations with significant two-way mutual information gain are excluded (Yaldız B. et al., 2022).

2.4. Biological networks and methods to compare biological networks

Graph theory is a mathematical discipline that underpins the study of complex networks in biological and other applications. It has been successfully extended to the study of biological network topology, from a global perspective of their scale-free, small-world, hierarchical existence to a zoomed-in view of interaction motifs, clusters, and modules as fundamental interactions between different biomolecules. Biological networks show that their structure is not random but instead linked to a function. (Pavlopoulus et al., 2011).

The term "network" is a common name for a collection of linked or interacting elements. An example of a network in biology is protein-protein interaction networks (PPI). Proteins are biological tissues' primary catalysts, structural components, signaling messengers, and molecular machines. PPIs play a critical role in coordinating the events in a cell and are the foundation of many signal transduction pathways and transcriptional regulatory networks in a cell. (Raman et al., 2010).

The principal measures that influence graphs are the number of nodes (N) and the networks' average degree (k). The type of network topology, which is generally unknown for experimental data, determines the direct nature of that influence. Therefore, direct comparisons of graph measures between empirical networks with different nodes (N) or edges (k) can give falsified results (Van Wijk et al., 2010). One of the challenges of comparing networks is that it is not easy to compare them as total entities. Global properties and summary statistics, such as network density, degree distribution, transitivity, average shortest path length, and others, can be used to compare networks. The studies comparing networks with different measures are done by (Van Wijk, Stam & Daffertshofer, 2010; Steuer and Lopez, 2008 and Valdeolivas, 2019).

2.5. GeneMANIA

GeneMANIA (Warde-Farley et al., 2010) is a web server predicting the functions of genes and gene sets. GeneMANIA searches many large, publicly available biological datasets to find related genes. These include protein-protein, protein-DNA and genetic interactions, pathways, reactions, gene and protein expression data, protein domains, and phenotypic screening profiles. The data in GeneMANIA is regularly updated. The explanation of network names is given below.

Physical interaction is a protein-protein interaction data; two gene products are linked if they were found to interact in a protein-protein interaction study. These data are collected from primary studies found in protein interaction databases, including BioGRID and PathwayCommons. Genetic interactions; two genes are functionally associated if the effects of perturbing one gene were found to be modified by perturbations to a second gene. These data are collected from primary studies and BioGRID. Co-localization; genes expressed in the same tissue or proteins found in the exact location. Two genes are linked if they are both expressed in the same tissue or if their gene products are both identified in the exact cellular location. Co-expression; is gene expression data. Two genes are linked if their expression levels are similar across conditions in a gene expression study. Most of these data are collected from the Gene Expression Omnibus (GEO); only data associated with a publication is collected. Shared protein domains; is protein domain data. Two gene products are linked if they have the same protein domain. These data are collected from domain databases, such as InterPro, SMART, and Pfam.

2.6. STRING analysis

STRING (Szklarczyk et al., 2019) is a functional protein association network used for building protein-protein interaction networks. The experimental data is extracted from BIND, DIP, GRID, HPRD, IntAct, MINT, and PID. The curated data is extracted from BioCarta, BioCyc, GO, KEGG, and Reactome databases. The detailed information about statistics terms is given below.

An **average node degree** is a number of how many interactions that a protein has on the average in the network. The **clustering coefficient** measures how connected the nodes in the network are. Highly connected networks have high values.

The **expected number of edges** gives how many edges are to be expected if the nodes were selected randomly. A small PPI enrichment p-value indicates that the nodes are not random and that the observed edges are significant. **Network diameter** is the shortest distance between the two most distant nodes in the network. A **radius** of the graph exists only if it has a diameter. The minimum among all the maximum distances between a vertex to all other vertices is the radius. **The characteristic path length** is the average shortest path length between all pairs of nodes in the network. **The clustering coefficient** measures the degree to which nodes in a graph tend to cluster together. (Holland and Leinhardt, 1971). The **network density** is determined by its ratio of links to the nodes. The higher the ratio, the denser the network. **Network efficiency** is measured by 1/L, where L represents average path length (Latora and Marchiori, 2001).

2.7. Pathway enrichment analysis

Quantification of biological samples, DNA, RNA, or protein, is becoming a standard in molecular genetic research. Moreover, extensive data is produced through high-yield quantification methods, and analysis of these data helps researchers to discover novel biological functions, genotype-phenotype relationships, and disease mechanisms. (Lander, 2011). Researchers use pathway enrichment analysis to gain mechanistic insight into gene lists produced by omics experiments. Pathways are statistically evaluated for over-representation in the experimental gene list in comparison to what would be predicted by chance, using a variety of conventional statistical analyses that take into account the number of genes found in the trial, their relative ranking, and the number of genes annotated to a pathway of interest (Reimand et al., 2019). In pathway enrichment analysis protocol Lander 2011, the *p*-value of the enrichment of a pathway is computed using a Fisher's exact test, and Benjamini & Hochberg –FDR test applies multiple-test correction. Multiple-testing correction methods are used to decrease the significance of each P-value derived from a series of tests. Fisher's exact test, based on hypergeometric distribution, is a standard statistical test used for pathway enrichment analyses of a gene list. It indicates whether the fraction of genes of concern in the pathway is more significant than the fraction of genes outside the pathway. Fisher's exact test is generally used for non-ranked gene lists, and since our gene lists are not ranked, as we used in this study.

An "enrichment map" is a network visualization that displays the overlaps across enriched pathways. On the other hand, "EnrichmentMap" is a Cytoscape application that creates visualization. Related pathways are automatically grouped into main biological themes by network layout and clustering algorithms. This method is used in several studies (Merico, Isserlin, Bader, 2011; Isserlin et al., 2014; Reimand et al., 2019; Karagiannis et al., 2013).

2.8. Knowledge graphs

Knowledge Graphs (KG) are an effective tool for data science in many data formats that obtain information (Ehrlinger and Wöß, 2016). The knowledge graph represents a collection of interlinked descriptions of entities – objects, events, or concepts. Knowledge graphs put data in context via linking and semantic metadata and, this way, provide a framework for data integration, unification, analytics, and sharing.

Knowledge Graphs can be used to represent biological entities. In the CROssBAR Knowledge Graphs by Dogan et al., biological entities are represented by nodes; relationships between the same and different types of biological entities are expressed by the graph's edges (Dogan et al., 2020).

2.9. Pyrosequencing

Pyrosequencing is a method developed by Pal Nyren in 1986 (Nyren, 2007). It is faster than Sanger sequencing and has advantages in convenience, staff, time, and cost. It performs sequence analysis by adding nucleotides one by one to the reaction during the synthesis of the target DNA product from a single strand (ssDNA) template. It is an accurate quantitative sequencing technique based on detecting pyrophosphates released by a biotin-labeled sequencing primer during DNA synthesis. Since it is still the fastest and most accurate sequencing method today, it is the most suitable method for detecting the presence of target mutations and SNPs in clinical samples (Ahmadian et al., 2006; Royo et al., 2007; Arnold et al., 2005; Vengen et al., 2012; Ballester LY, 2016).

The basic principle of pyrosequencing is the sequential addition of dNTPs to newly formed DNA and the determination of the sequence by detecting pyrophosphate (PPi), which is released when the corresponding DNA template is added. The single-stranded DNA template is hybridized to the sequencing primer and incubated with DNA polymerase, ATP sulfurylase, luciferase, apyrase enzymes, and substrates of adenosine 5' phosphosulfate (APS) and luciferin. Adding one of the nucleotides Adenine, Guanine, Cytosine, and Thymine (A, G, C, T) releases pyrophosphate (PPi) if a nucleotide matches the DNA template. ATP sulfurylase converts PPi to ATP in adenosine 5' phosphosulfate. This ATP formed serves as a substrate for the conversion of luciferin to oxyluciferin in the presence of luciferase. Oxyluciferin is the substance that emits light. There is a linear relationship between the generated ATP and oxyluciferin. In order to observe the resulting light oscillation, the pyrogram graph is examined. The enzyme apyrase degrades unbound nucleotides and ATP.

The pyrosequencing method, a high-performance and fast method, has become one of the clinical diagnostic tests used in medical biology and genetics laboratories to determine mutations and SNPs in certain diseases. Pyrosequencing can identify variations with very low margins of error, whose incidence is as low as 5% in the population (Alqahtani QM et al., 2016). In addition, pyrosequencing is a more adaptable, fast, and economic analysis than other high-throughput sequencing methods.

CHAPTER 3

3. MATERIALS AND METHODS

3.1. Overview of the method

Our study obtains SNPs from GWAS datasets provided by ADNI, GenADA, and NCRAD initiatives. The SNP prioritization is conducted by novel Ensembl and Entropy-based data mining methods (Erdoğan O et al. 2022 and Yaldız B et al., 2022).

This study aims to construct and compare biological networks of the variants obtained from these two methods and do functional enrichment analysis to discover the affected biological pathways. The obtained figures are built using databases; STRING 11.0, GeneMANIA, CROssBAR, g: Profiler, and software Cytoscape with plugins EnrichmentMap and Autoannotate. A detailed explanation of the methods is provided below.

In this study, for genotyping, the pyrosequencing method is used. Genomic DNA was obtained from 43 participants from LOAD groups and 38 participants from control groups. The workflow of Pyrosequencing consists of 3 stages;

1) DNA isolation; 2) PCR; Creating a single-stranded pattern; 3) Pyrosequencing.

3.2. Obtaining data

This section explains all data used in this study in detail. The gene sets used in this study are provided by our team working on these genes in their research. The final five gene sets are named: Ensembl >2.31, Ensembl >3.21, and Entropy genes obtained from three databases; ADNI, GenADA, and NCRAD. The number of genes in these gene sets

respectively are; 238, 63, 138, 19, 15, and 74. The process of obtaining these gene sets is explained below.

3.2.1. Genotyping Data

Three different high dimensional datasets from Alzheimer's Disease Neuroimaging Initiative (ADNI), GenADA (dbGaP Study Accession: phs000219.v1. p1), and the National Centralized Repository for Alzheimer's Disease and Related Dementias (NCRAD) (dbGaP study accession: phs000168.v1. p1.) are obtained via dbGaP control access (Filippini et al., 2009; Li et al., 2008). Affymetrix Mapping250K_Nsp and Mapping250K_Sty with 620901 SNPs Illumina Human610_Quadv1_B 500K with 410969 SNPs and Illumina Human610-Quad BeadChip with 590247 SNPs platforms are used by these initiatives, respectively. Two hundred ten controls and 344 cases for ADNI, 777 controls and 798 cases for GenADA, 1310 controls, and 1289 cases for NCRAD are genotyped using these platforms.

3.2.2. SNP Prioritization

For the initial dimension reduction to discover statistically significant SNP's for building a LOAD model from each dataset, GWAS analysis is done. PLINK analysis is completed for identifying the independent statistical significance of variations related to the LOAD (Chang et al., 2015; Purcell et al., 2007). PLINK results are used for filtering and reducing redundant SNPs that are not directly related to the disease.

After filtering, SNP's significantly associated with LOAD are used as an input for modeling a multistep RF strategy. One RF is used for dimension reduction, and in the next step, another RF is conducted as the modeling algorithm. RF is implemented with 5-fold cross-validation (CV) using the RANGER package in R (Wright & Ziegler, 2017).

3.2.3. ENSEMBLE and ENTROPY Scoring Algorithm

The multistep RF-RF modeling selected a set of prioritized variants out of thousands of variants. Following the SNP prioritization step, our team implemented two meta-analysis approaches, Ensemble, and Entropy, for model minimization after RF- RF modeling.

The Ensemble is a scoring algorithm in which chromosomal locations of SNVs are obtained by mapping to cytogenetic bands affinities between pairs (Onur Erdoğan p.c., 2022). 719 LOAD-associated variants from three different sequencing platforms are selected based on Ensemble scores.

The Entropy approach measures the difference between the mutual information in case and control groups. Three-way interaction information (3WII) calculated by entropy analysis can assess third-order interactions. The 3WII describing the amount of information common to all variables not present in any other subset alone are calculated for all three datasets to reveal triplets of variants significantly associated with LOAD (Bucru Yaldız p.c.,2022).

3.2.4. Annotation of Variants

The prioritized variants obtained by Ensemble and Entropy methods are annotated using the SNPNexus tool. (Oscanoa et al., 2020; Chelala et al., 2009; Dayem Ullah et al., 2012, 2013, 2018). For some analysis, an extended version of gene sets is used, consisting of overlapped genes, nearest downstream genes, and nearest upstream genes, and for some analysis, only overlapped gene sets are used.

3.2.5. Obtaining Gene Lists

After obtaining lists of rsID's from Onur Erdoğan (p.c.,2022) and Burcu Yaldız (p.c., 2022) these variants are mapped to genes. The gene lists that are used in this study are added to the Appendix D. The number of genes for Ensemble >2.31 is 238, for Ensemble 3.21> is 63, for Entropy-all 138, for Entropy-ADNI 19, for Entropy-GenADA 15 and for Entropy-NCRAD 74.

3.3. GeneMANIA based PPI network analysis

Genemania (Warde-Farley et al., 2010) is a web server predicting the functions of genes and gene sets. Co-expression, shared protein domains, physical interactions, genetic interactions, pathways, and co-localization are the network categories created by GeneMANIA. We constructed gene networks based on their functions by using GeneMANIA for our gene sets; Ensemble, Entropy, NCRAD, ADNI, and GenADA.

3.4. Building protein-protein interaction networks by STRING

STRING (Szklarczyk et al., 2019) is a functional protein association network used for building protein-protein interaction networks. The experimental data is extracted from BIND, DIP, GRID, HPRD, IntAct, MINT, and PID. The curated data is extracted from BioCarta, BioCyc, GO, KEGG, and Reactome databases. A PPI enrichment p-value shows that nodes are not connected randomly, and the edges between the nodes are significant.

This study built gene-gene interaction networks for gene sets Ensemble, Entropy, ADNI, GenADA, and NCRAD. These networks were created with a confidence (score) cutoff of 0.40. The cutoff score determines the minimum interaction score required for being included in the prediction network. Additional interactors are added until the PPI enrichment is less or equal to 0.05, which means that our proteins have more interactions among themselves than what would be expected for a collection of proteins drawn at random from the genome that are identical in size. Such an enrichment indicates that the proteins are at least partially biologically connected as a group, where all networks have PPI enrichment less or equal to 0.05. NetworkAnalyzer (Assenov et al., 2008) is a Cytoscape plugin used to obtain summary statistics. It is a plugin that computes the topological attributes for biological networks. Summary statistics created by NetworkAnalyzer are used to compare the networks created.

3.5. Building network gene-training disease interaction networks by Cytoscape

Cytoscape (Shannon et al., 2003) is an open-source software platform for visualizing molecular interaction networks and biological pathways and incorporating annotations, gene expression profiles, and other state data into these networks. In this study, we used this software to obtain gene-disease interaction networks. HGPEC (Le et al., 2017) application is used for building gene-disease interaction networks. HGPEC is based on a random walk with a restart algorithm through a heterogeneous network of genes and diseases. This application provides a heterogeneous network of genes/proteins and a phenotypic disease similarity network to prioritize network genes and diseases. Novel disease-gene and disease-disease associations can be identified based on the rankings. The gene-disease associations are obtained from DisGeNET (Pinero et al., 2019), the largest publicly available discovery platform. Alzheimer's disease is selected as the disease of interest, and a training gene (genes that have functional associations with Alzheimer's disease) list is created. The diseases selected for the training list are indicated in **Table 3.1**

Table 3.1. Diseases selected as training diseases for HGPEC.

Disease ID	Name
MIM104300	ALZHEIMER DISEASE; AD
MIM104310	ALZHEIMER DISEASE 2
MIM125320	DEMENTIA/PARKINSONISM WITH NON-ALZHEIMER AMYLOID PLAQUES

MIM602096	ALZHEIMER DISEASE 5
MIM605055	ALZHEIMER DISEASE, FAMILIAL EARLY-ONSET, WITH COEXISTING AMYLOID AND PRION PATHOLOGY
MIM605526	ALZHEIMER DISEASE 6
MIM606889	ALZHEIMER DISEASE 4
MIM607822	ALZHEIMER DISEASE 3
MIM608907	ALZHEIMER DISEASE 9

Network gene sets are created by manual input, and these genes are prioritized by back probability 0.5, jumping probability 0.6, and sub-network importance weight 0.7. These networks are built for Ensemble, Entropy, ADNI, GenADA, and NCRAD genes with the number of genes respectively, 37, 38, 11, 8 and 20. The detailed summary of attributes used in HGPEC analysis is indicated in **Table 3.2**. The Cytoscape plugin Diffusion is used to broaden node selection using network propagation algorithms. The Diffusion algorithm in Cytoscape aims to find the most critical nodes from a group of nodes and an entire interaction network. By using this algorithm, networks are constructed from our genes of interest and training diseases.

Table 3.2. Summary of HGPEC attributes.

Summary	**ENSEMBLE**	**ENTROPY**	**ADNI**	**GenADA**	**NCRAD**
Disease	Alzheimer	Alzheimer	Alzheimer	Alzheimer	Alzheimer
Back probability	0.5	0.5	0.5	0.5	0.5
Jumping probability	0.6	0.6	0.6	0.6	0.6
Subnetwork importance	0.7	0.7	0.7	0.7	0.7
Disease Network size	|V|=5080, |A|=38467	|V|=5080, |A|=38467	|V|=5080, |A|=38467	|V|=5080, |A|=38467	|V|=5080, |A|=38467
Gene/Protein Network Size	|V|=10486, |A|=50791	|V|=10486, |A|=50791	|V|=10486, |A|=50791	|V|=10486, |A|=50791	|V|=10486, |A|=50791

Number of Training Genes	1183	1183	1183	1183	1183
Number of Training Diseases	9	9	9	9	9
Number of Network Genes	37	38	11	8	20

3.6. Prediction of network gene-Alzheimer's disease interaction networks by Cytoscape

Gene-Alzheimer's disease interaction networks are built by the HGPEC plugin on Cytoscape. The networks constructed for genes from gene sets Ensemble, Entropy, ADNI, GenADA, and NCRAD predict a relationship to Alzheimer's Disease. The gene-disease associations are obtained from DisGeNET. Alzheimer's Disease is selected as the disease of interest, and a training gene (known Alzheimer's Disease-related genes) list is created. The disease IDs and names selected as training diseases are represented in **Table 3.3**. Furthermore, a detailed summary of HGPEC attributes is given in **Table 3.4.**

Table 3.3. Disease IDs and names of training diseases.

Disease ID	Name
MIM104300	ALZHEIMER DISEASE; AD
MIM104310	ALZHEIMER DISEASE 2
MIM125320	DEMENTIA/PARKINSONISM WITH NON-ALZHEIMER AMYLOID PLAQUES
MIM602096	ALZHEIMER DISEASE 5
MIM605055	ALZHEIMER DISEASE, FAMILIAL EARLY-ONSET, WITH COEXISTING AMYLOID AND PRION PATHOLOGY
MIM605526	ALZHEIMER DISEASE 6
MIM606889	ALZHEIMER DISEASE 4
MIM607822	ALZHEIMER DISEASE 3
MIM608907	ALZHEIMER DISEASE 9

Table 3.4. HGPEC attributes for network gene-Alzheimer's disease networks.

Summary	**ENSEMBLE**	**ENTROPY**	**ADNI**	**GenADA**	**NCRAD**
Disease	Alzheimer	Alzheimer	Alzheimer	Alzheimer	Alzheimer
Back probability	0.5	0.5	0.5	0.5	0.5
Jumping probability	0.6	0.6	0.6	0.6	0.6
Subnetwork importance	0.7	0.7	0.7	0.7	0.7
Disease Network size	|V|=5080, |A|=38467	|V|=5080, |A|=38467	|V|=5080, |A|=38467	|V|=5080, |A|=38467	|V|=5080, |A|=38467
Gene/Protein Network Size	|V|=10486, |A|=50791	|V|=10486, |A|=50791	|V|=10486, |A|=50791	|V|=10486, |A|=50791	|V|=10486, |A|=50791
Number of Training Genes	1183	1183	1183	1183	1183
Number of Training Diseases	9	9	9	9	9
Number of Network Genes	37	38	11	8	20

3.7. Functional enrichment analysis and visualization

Pathway enrichment and visualization of the data were done following a protocol written by Reimand et al. (2019).

3.7.1. Pathway enrichment analysis by g:Profiler

g: Profiler (Raudvere et al., 2019) is a public web server for characterizing and manipulating gene lists. g: GOSt is the core of the g: Profiler; it provides statistical enrichment analysis to analyze the given gene list. In this study, g: GOSt is used to obtain Gene Ontology (GO) Molecular Function, GO Cellular Component, GO Biological Process and Reactome pathways for Ensemble, Entropy, GenADA, ADNI, and NCRAD gene sets.

All analyses were done with default attributions with a significance threshold of 0.05, and multiple analyses correction was done with Bonferroni.

3.7.2. Visualization by EnrichmentMap

EnrichmentMap (Merico et al., 2010) is a Cytoscape plugin used for functional enrichment visualization. Overrepresented functional groups derived from functional annotation such as Gene Ontology (GO) can be identified by enrichment analysis. Gene sets, such as pathways and GO terms, are organized into networks. In this study, EnrichmentMap is used to create networks from GO annotations and Reactome pathways for Ensemble, Entropy gene sets. The *P*-value of the enrichment of a pathway is computed using Fisher's exact test and Benjamini-Hochberg FDR (Q value) used as multiple test correction. All analyses were done with a *p-value* of 0.05 FDR q-value cutoff 0.1 and edge similarity cutoff 0.25 (Jaccard metric).

3.7.3. Interpretation by AutoAnnotate

AutoAnnotate (Kucera et al., 2016) is a Cytoscape plugin that finds clusters and annotates them visually with labels and groups. This study used AutoAnnotate to create clusters after functional enrichment analysis. The clusters are obtained using the Markov Cluster Algorithm (MCL) cluster annotation algorithm, and labels are generated automatically based on the word frequencies of selected attributes. GO Molecular Function, GO Cellular Component, GO Biological Process, and Reactome were used for label calculation.

3.8. Building knowledge graphs by CROssBAR

CROssBAR (Dogan et al., 2020) is a database constructing knowledge graphs for biological entities and relationships between them, represented by nodes and edges. This study created knowledge graphs for Ensemble genes, Entropy genes, ADNI, GenADA, and NCRAD genes with the disease query "Alzheimer's Disease." We used knowledge graphs to obtain literature information related to our genes and the pathways in which they occur. The observed gene-pathway associations are investigated in pathway enrichment analysis. The genes and query parameters details are given in **Table 3.5 in Appendix A**. The biomedical data sources used in the CROssBAR database are; UniProt, IntAct, InterPro, DrugBank, ChEMBL, PubChem, Reactome, KEGG, OMIM, Orphanet, Experimental Factor Ontology (EFO), and Human Phenotype Ontology (HPO).

3.9. Experimental validation

3.9.1. Primer Design

Primer design for PCR primer and Pyrosequencing primer was done using Pyromark Assay Design 2.0 by Qiagen. After the design of the primer, the suitability of the designed primers was checked by using the BLAT tool (Kent, 2002) from USCS Genome Browser. In the study, the design of PCR primers is done by following standard rules. It is designed that the temperature is in the range of 62-64°C, the primer length is 18-25 bases, the formation of dimers and loops is prevented, and the primers with as equal G/C-A/T as possible. In order to bind to streptavidin-coated magnetic beads, one of the primers was biotin-labeled and the other unmarked.

3.9.2. DNA Isolation

The saliva samples are collected from Turkish LOAD patients and controls at the Hacettepe University Geriatric Clinic. Saliva samples received by METU Bioinformatics Systems Biology Laboratory were used for genomic DNA isolation. Genomic DNA isolations of the samples of the participants were obtained with the optimized protocol based on the Norgen Saliva DNA isolation protocols. Genomic DNA was obtained from 43 LOAD groups and 38 control groups in total.

3.9.3. NanoDrop Spectrophotometry

After the DNA isolation of the saliva samples, concentration and quality evaluations were completed with NanoDrop Spectrophotometry. Samples with good quality and concentration were used for the next step.

3.9.4. Polymer Chain Reactions (PCR)

The polymerase chain reaction (PCR) is used to amplify small segments of DNA. Because the obtained DNA amount after DNA isolation is not enough for the necessary molecular analysis, which is pyrosequencing in our case.

3.9.5. Pyrosequencing

The pyrosequencing method, which is fast and has an error rate of less than 1/1000, was preferred for genotyping. The pyrosequencing method is known as "sequencing by synthesis". It is a method based on detecting which base the enzyme adds as DNA polymerase produces new DNA. Pyrosequencing is based on detecting light emission

resulting from the chain reaction that occurs when pyrophosphate is released. Pyrosequencing is done by using the Qiagen Pyromark Q24 machine. In total, for 43 LOAD patients, 38 controls and 32 variants runs were constructed.

CHAPTER 4

4. RESULTS

This study compares the variant lists produced by two feature selection approaches, Ensemble and Entropy, based on their biological interpretation with experimental validation on the variant lists provided by Onur Erdoğan and Burcu Yaldız. PLINK association analysis is performed in these prior studies for genotyping data of three LOAD GWAS datasets. Then RF-RF is conducted for modeling. The prioritized variants after PLINK-RF-RF steps are analyzed either ensemble or entropy approaches for model minimization.

The 719 variants proposed for LOAD classification with RF-RF models for three GWAS datasets are scored according to the Ensemble method. The variants with an Ensemble score of 2.31 and higher are categorized as two groups for investigation as an Ensemble >2.31 and Ensemble >3.21 (Onur Erdoğan et al., personal communication).

Next, for the Entropy selection, variants selected by RF-RF models are prioritized by three-way-interaction analysis (3WI). These results are investigated under four groups Entropy-ALL, Entropy-ADNI, Entropy-GENADA, and Entropy-NCRAD. The number of selected variants for these groups were 145, 39, 25, and 78, respectively (Burcu Yaldız et al., personal communication).

In all six analysis groups, variants overlapping with a protein-coding gene and LOAD-related biological pathways were selected for experimental validation based on the examination results in terms of biological networks, protein-protein interactions, and functional enrichment. These network analyses are planned to show the functional relevance of RF-RF model variants associated with LOAD risk as potential causative variants. The advantages and disadvantages of ensemble and entropy-based approaches are discussed.

Finally, for LOAD-associated variants selected with model minimization, pyrosequencing primer design is completed, and sequencing primers were optimized. Out of the 32 genes,

3 were controls, the Ensemble model prioritized 11 genes, and 18 were selected by the Entropy model of GenADA and ADNI studies. Results of the genotyping experiments and their analysis are presented and discussed.

4.1. Prediction of gene functions by GeneMANIA

By using GeneMANIA, gene functions are predicted for our gene sets. According to the co-expression results, ADNI has the highest percentage with 100%, GenADA follows with 83.04%, and the other gene sets follow with; Entropy: 79.31%, Ensemble: 71.15%, and NCRAD: 63.68%. A detailed summary of GeneMANIA network results is given in **Table 4.1** and a detailed table of the genes found in the GeneMANIA networks is given in **Table 4.2**, Appendix B.

Table 4.1. GeneMANIA network summary: The statistical analysis of the GeneMANIA results for all five gene lists are summarized.

SUMMARY STATISTICS	**ENSEMBLE >2.31**	**ENSEMBLE >3.21**	**ENTROPY**	**Entropy-ADNI**	**Entropy-GENADA**	**Entropy-NCRAD**
#related genes	20	20	20	20	20	20
#total genes	218	70	66	34	30	42
#attributes	0	0	0	0	0	0
#total links	4250	353	269	61	114	109
Co-expression	41.47%	59.19%	**85.18%**	94%	77.90%	81.07%
Shared protein domains	2.12%	**10.96%**	N/A	N/A	22.10%	N/A
Physical interactions	**19.19%**	13.18%	5.57%	N/A	N/A	18.34%
Genetic Interactions	**6.00%**	3.14%	1.44%	6.33%	N/A	0.59%
Pathway	4.69%	N/A	N/A	N/A	N/A	N/A
Co-localization	**26.52%**	2.64%	7.81%	N/A	N/A	N/A

In Table 4.1., co-expression, shared protein domains, physical interactions, genetic interactions, and co-localization features are investigated to compare our gene lists. The highest values corresponding to these features are highlighted in bold.

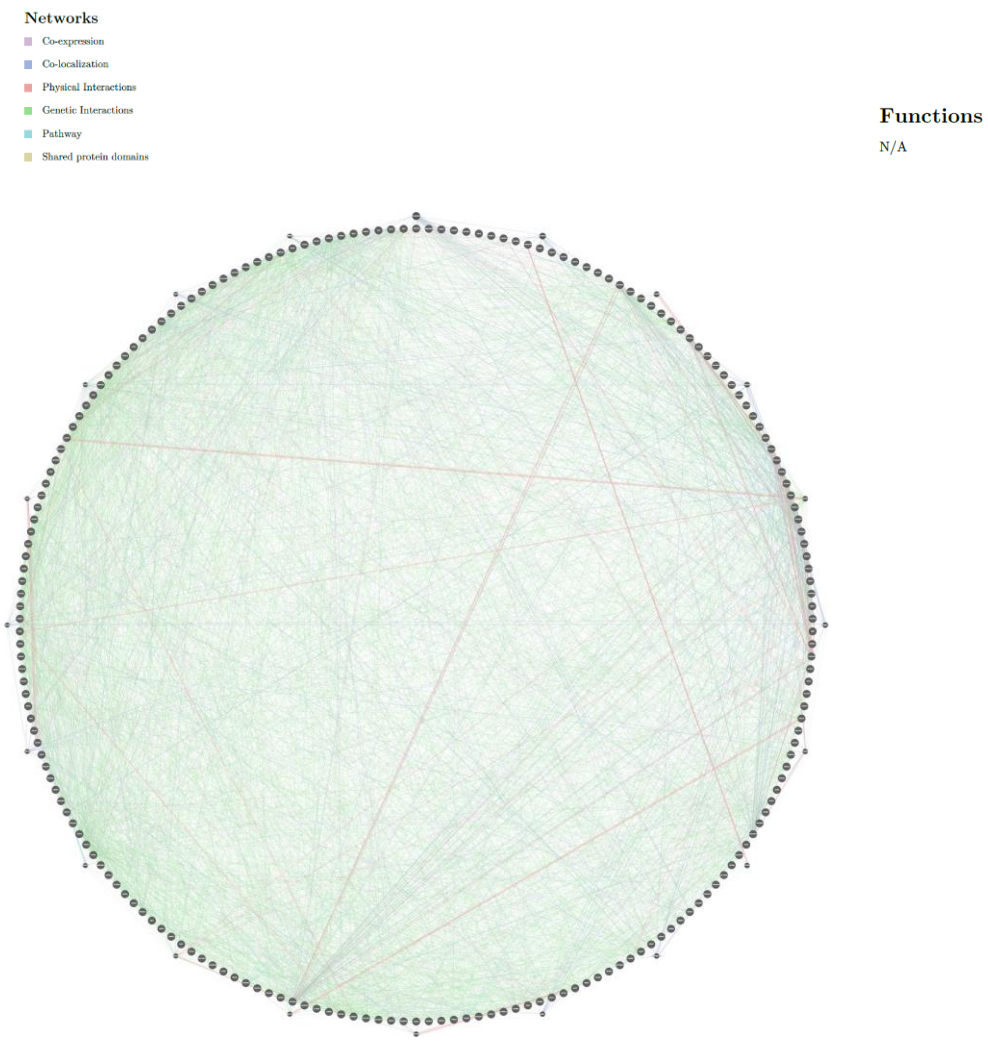

Figure 4.1. Topological relationship for Ensemble > 2.31 scored only overlapped genes and related genes obtained by GeneMANIA. Purple strings represent co-expression, pink strings represent physical interactions, yellow strings represent shared protein domains, green strings represent genetic interactions, and blue strings represent pathways between nodes. Nodes represent genes, and we have searched genes are indicated with stripes.

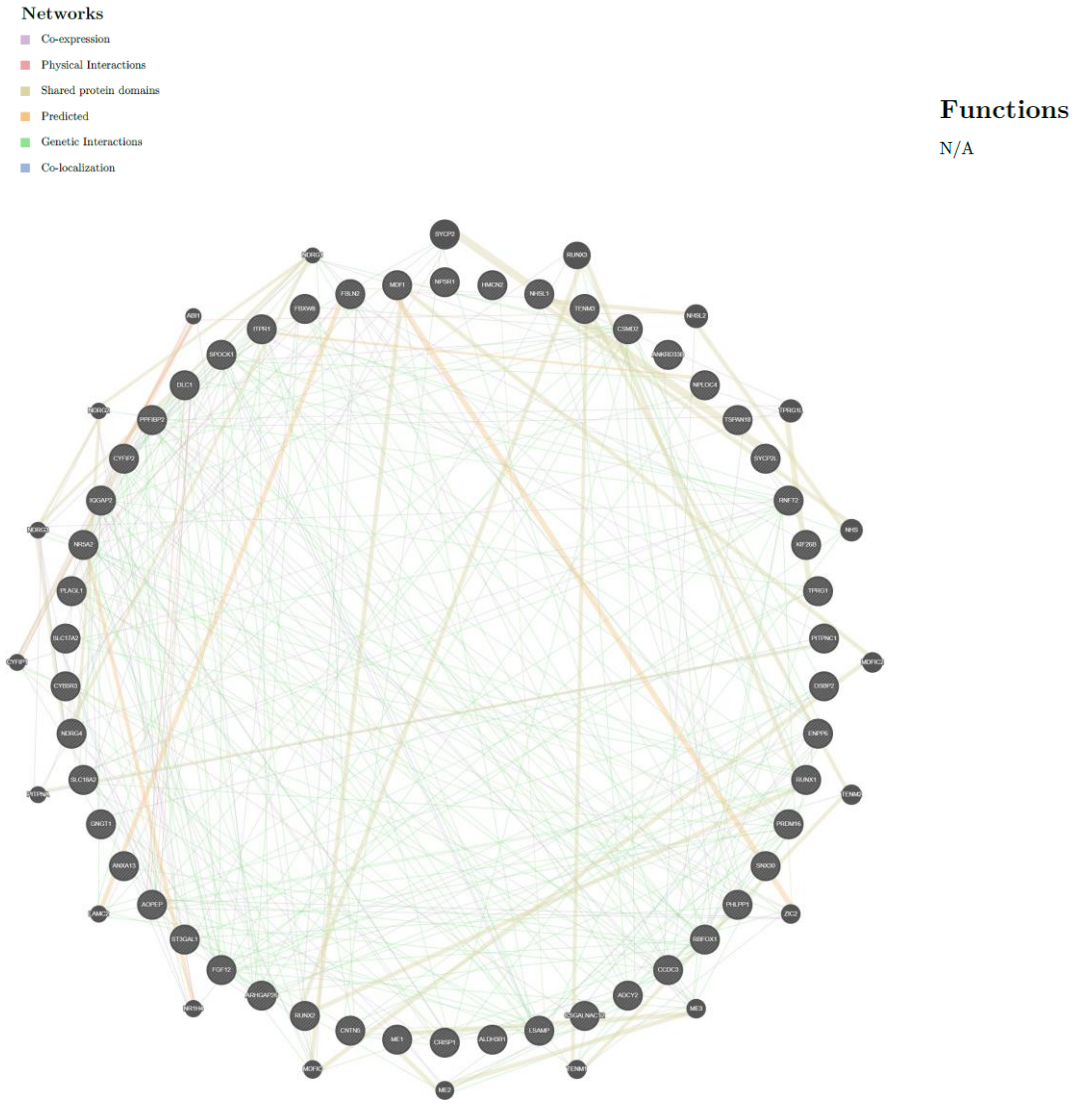

Figure 4.2. Topological relationship between Ensemble >3.21 genes. Purple strings represent co-expression, pink strings represent physical interactions, green strings represent genetic interactions, and blue strings represent colocalization between nodes. Nodes represent genes, and we have searched genes are indicated with stripes

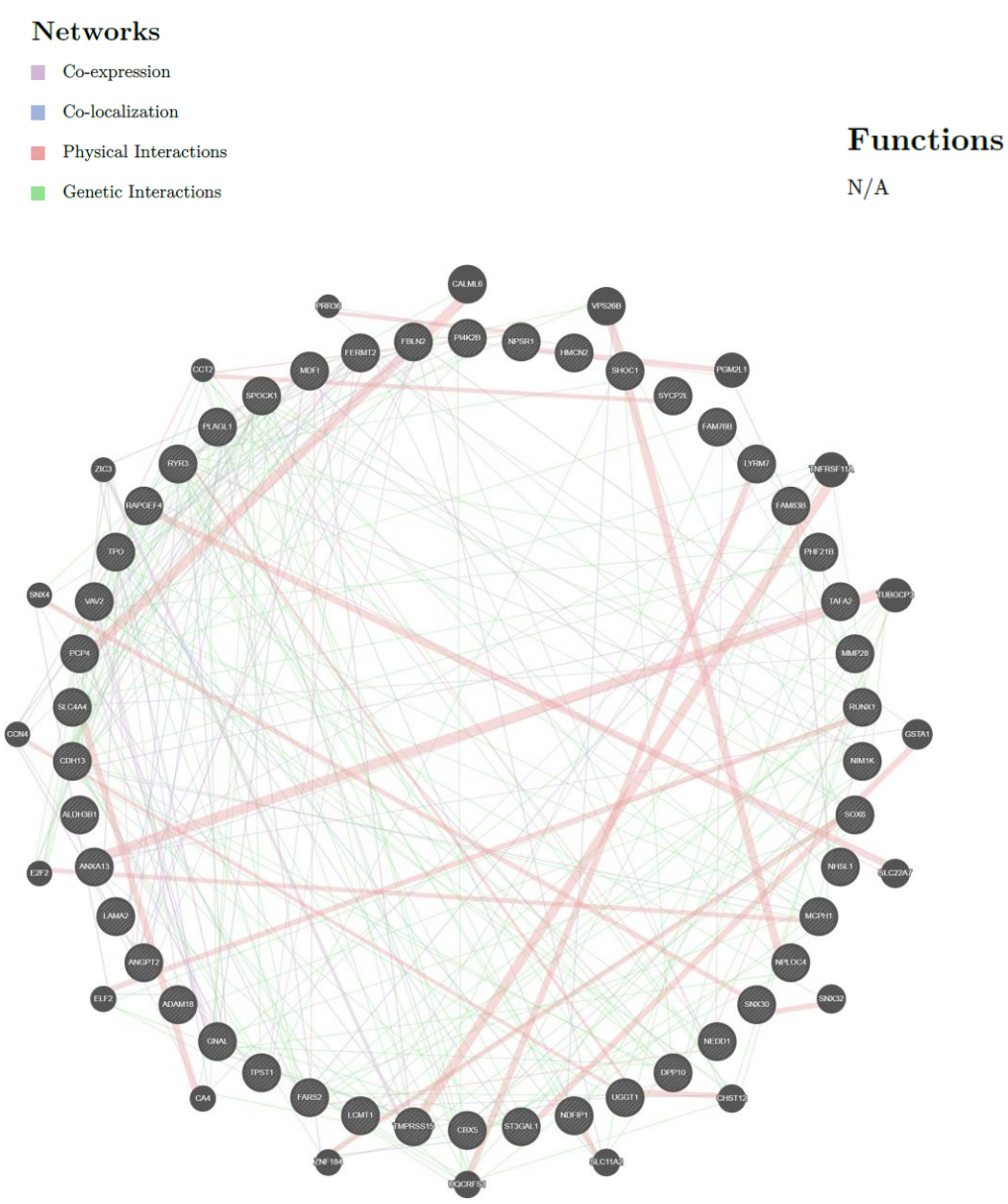

Figure 4.3. Topological relationship between Entropy genes. Purple strings represent co-expression, pink strings represent physical interactions, green strings represent genetic interactions, and blue strings represent colocalization between nodes. Nodes represent genes, and we have searched genes are indicated with stripes.

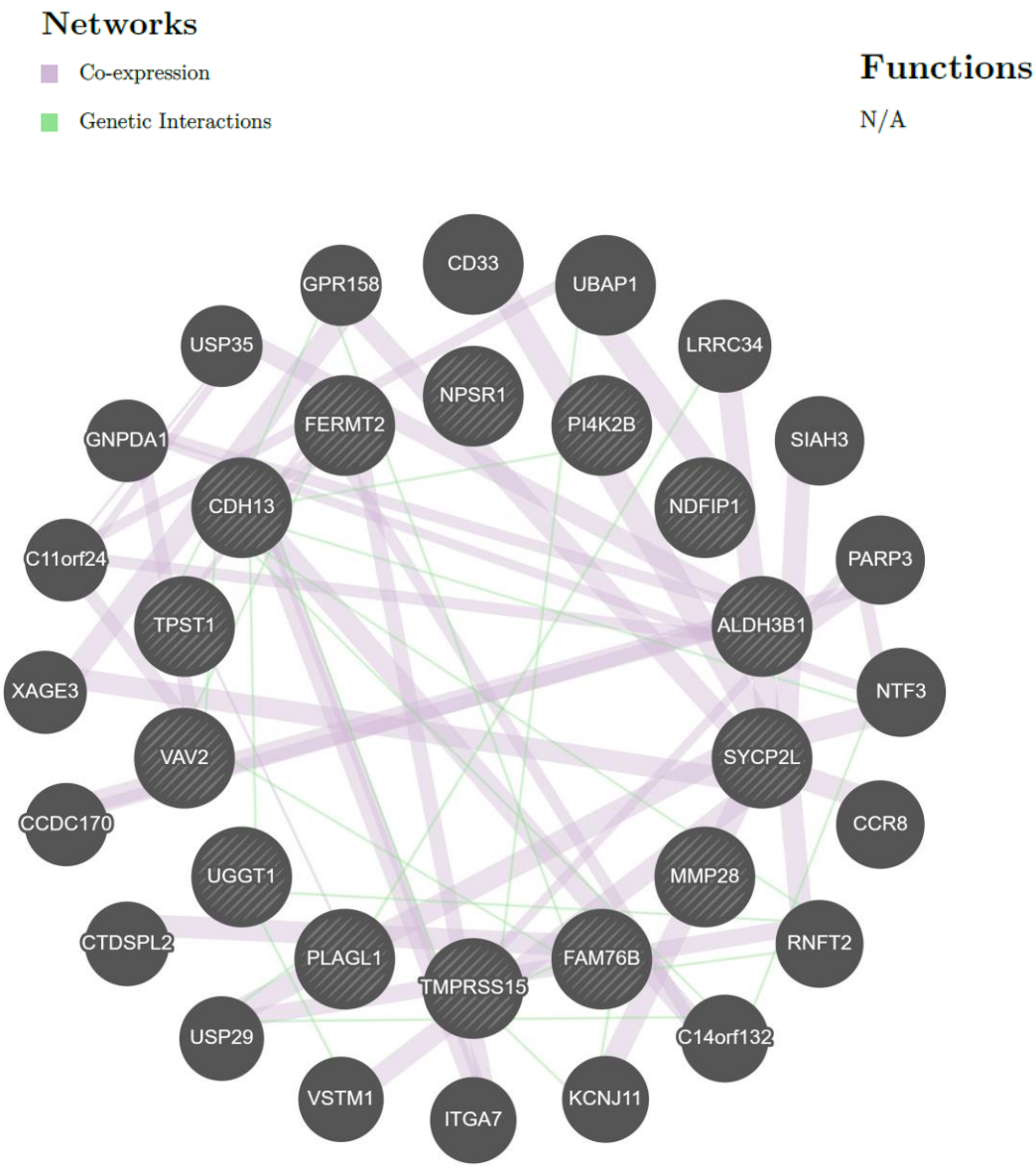

Figure 4.4. Topological relationship between ADNI genes. Purple strings represent co-expression, and green strings represent genetic interactions between nodes. Nodes represent genes, and we have searched genes are indicated with stripes.

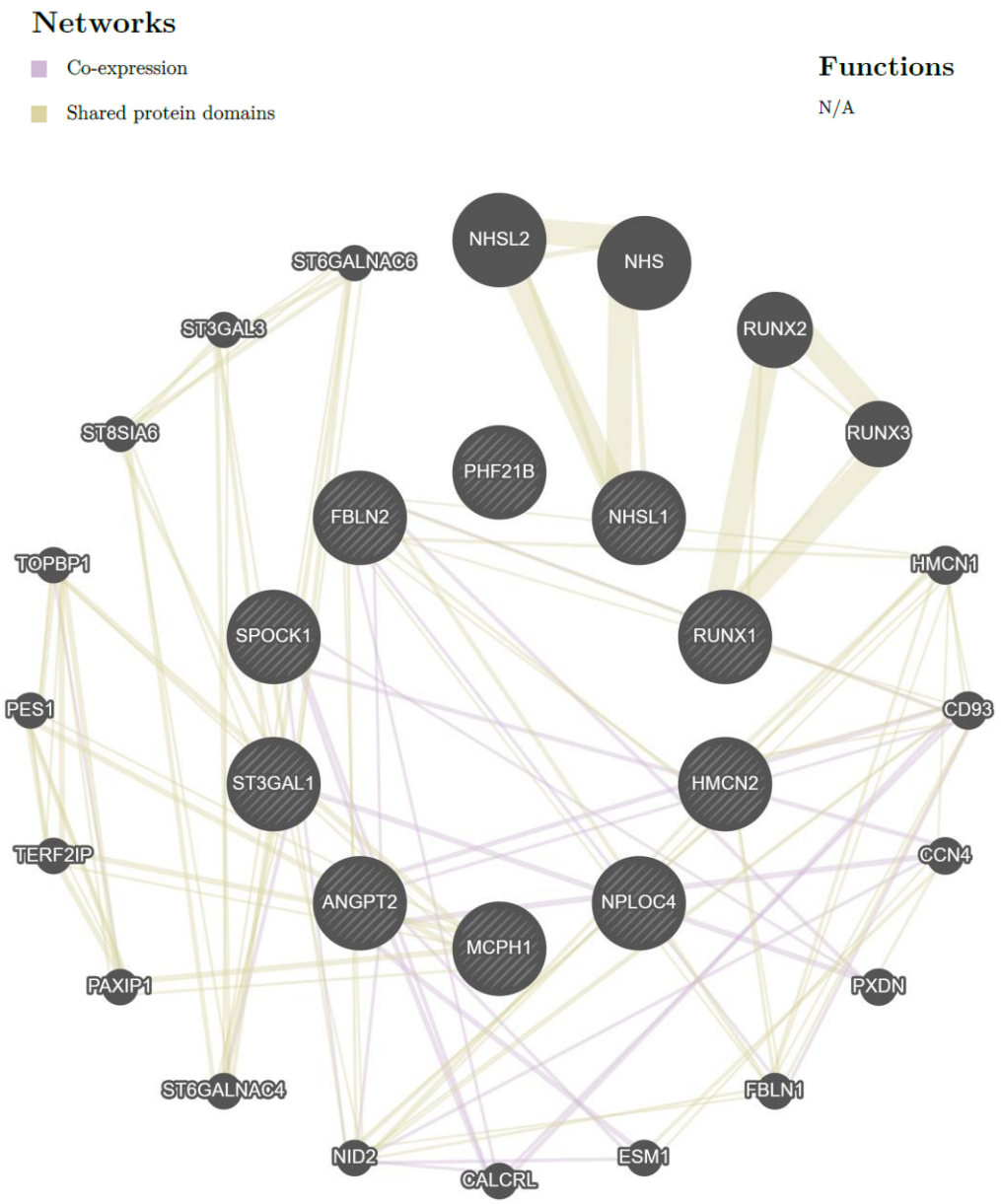

Figure 4.5. Topological relationship for GenADA genes and related genes obtained my GeneMANIA. Purple strings represent co-expression and yellow string represent shared protein domains. Nodes represent genes and genes that we have searched are indicated with stripes.

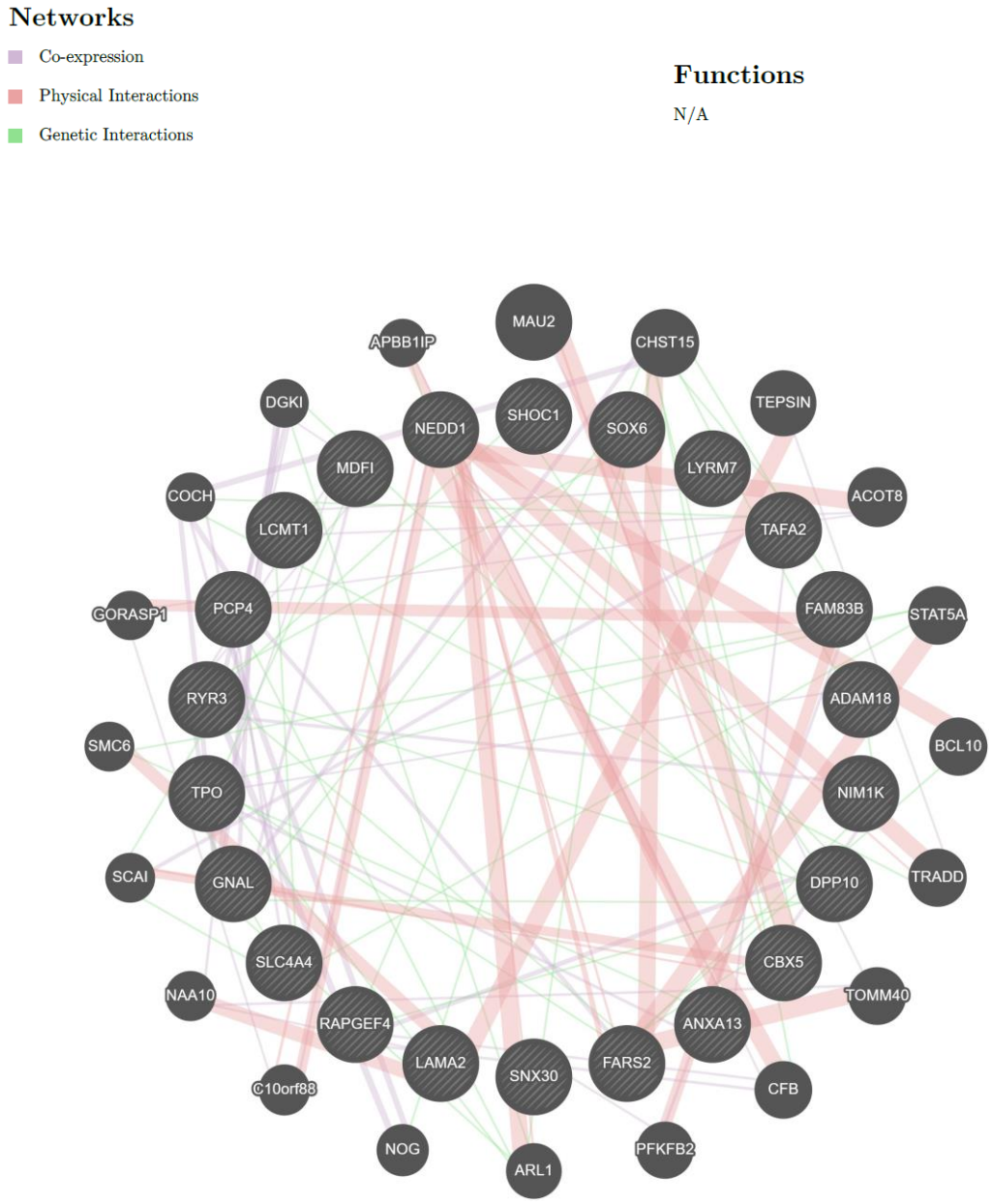

Figure 4.6. Topological relationship for NCRAD genes and related genes obtained by GeneMANIA. Purple strings represent co-expression, green strings represent genetic interactions and pink strings represent physical interactions. Nodes represent genes and genes that we have searched are indicated with stripes.

Other figures obtained by GeneMANIA analysis are given in the **Figure 7-22** in the **Appendix B**.

Table 4.3. Number of total genes and genes not connected to any other genes for Ensemble and Entropy networks.

#genes	Ensemble > 2.31 physical int	Ensemble > 2.31 shared protein domains	Ensemble >2.31 co-loc	Ensemble > 3.21 physical int	Ensemble >3.21 shared protein domains	Ensemble >3.21 co-loc	Entropy physical int	Entropy co-loc
Total genes	218	218	218	70	70	70	66	66
Not connected genes	136	110	105	60	27	57	26	53
% CONNECTED	37.62	49.55	**51.84**	14.29	**61.43**	18.58	**60.61**	19.70

When we look at the number of connected genes from the obtained GeneMANIA figures; for Ensemble >2.31, physical interactions network 110 genes are not connected to any other genes, out of 218 total genes. So, 37.6% of genes are connected. In the Ensemble >2.31 shared protein domains network 49.5% of genes are connected. And for the Ensembl >2.31 co-localization network 51.8% of genes are connected.

On the other hand, for Ensemble >3.21 physical interactions network 14.3% of genes are connected and for shared protein domains network 61.4% of genes are connected. While for Ensemble >3.21 co-localization network 18.6% of genes are connected.

Lastly, for Entropy physical interactions network, 60.6% of genes are connected and for the co-localization network, 19.7% of genes are connected.

4.2. Protein-protein interaction networks by STRING

We constructed protein-protein interaction networks for our gene sets; Ensemble, Entropy, Entropy-GenADA, Entropy-ADNI, and Entropy-NCRAD by STRING. The p-value obtained for Ensemble, Entropy, ADNI, GenADA, and NCRAD is 0.291, 0.968, 1, 1, and 0.723. None of the obtained interaction networks is significant.

4.3. STRING enrichment analysis

The networks constructed for gene sets obtained by STRING enrichment were compared by expanding each network to the point of significantly more interactions than expected.

The networks created for the Ensemble > 2.31, Ensemble >3.21, Entropy, Entropy-ADNI, Entropy-GenADA, and Entropy-NCRAD for only overlapped genes are given in detail information about the number of nodes and edges for the gene sets are given in **Table 4.4.** Moreover, detailed summary statistics for STRING enrichment are given in **Table 4.5**. **Figure 4.23, Figure 4.24, Figure 4.25, Figure 4.26, Figure 4.27** and **Figure 4.28** show the extended networks for each group of proteins.

Table 4.4. Numbers of nodes and edges for gene sets.

	ENSEMBLE >2.31	**ENSEMBLE>3.21**	**ENTROPY**	**Entropy-ADNI**	**Entropy-GenADA**	**Entropy-NCRAD**
Number of nodes	199	51+10	43+10	15+10	11+10	23+10
Number of edges	191	96	74	46	34	49

Table 4.5. Summary statistics for STRING enrichment analysis. The highest values for each row is marked in bold.

SUMMARY STATISTICS	**ENSEMBLE >2.31**	**ENSEMBLE >3.21**	**ENTROPY**	**E-ADNI**	**E-GenADA**	**E-NCRAD**
#of nodes	199	61	53	25	21	33
#of edges	191	96	74	46	34	49
Avg # of neighbors	3.064	**5.875**	5.481	5.75	3.778	5.444
network diameter	**14**	5	5	3	6	5
network radius	**7**	3	2	2	3	3
characteristic path lenght	**5.057**	2.337	2.282	1.808	2.549	2.386
clustering coefficient	0.153	0.443	**0.493**	0.608	0.357	0.48
network density	0.028	0.19	0.211	**0.383**	0.222	0.32
network heterogeneity	0.758	**0.797**	**0.797**	0.462	0.511	0.64
network centralization	0.075	0.28	**0.354**	0.324	0.279	0.301
connected components	**76**	28	27	10	4	16
analysis time	0.086	0.02	0.004	0.001	0.001	0.002
efficiency	0.197746	0.427899	0.438212	0.553097	0.392311	0.419111
PPI enrichment p-value	**1.00E-16**	9.30E-04	3.99E-09	1.35E-11	2.72E-10	1.00E-16
significantly more interactions than expected	Yes	Yes	Yes	Yes	Yes	Yes

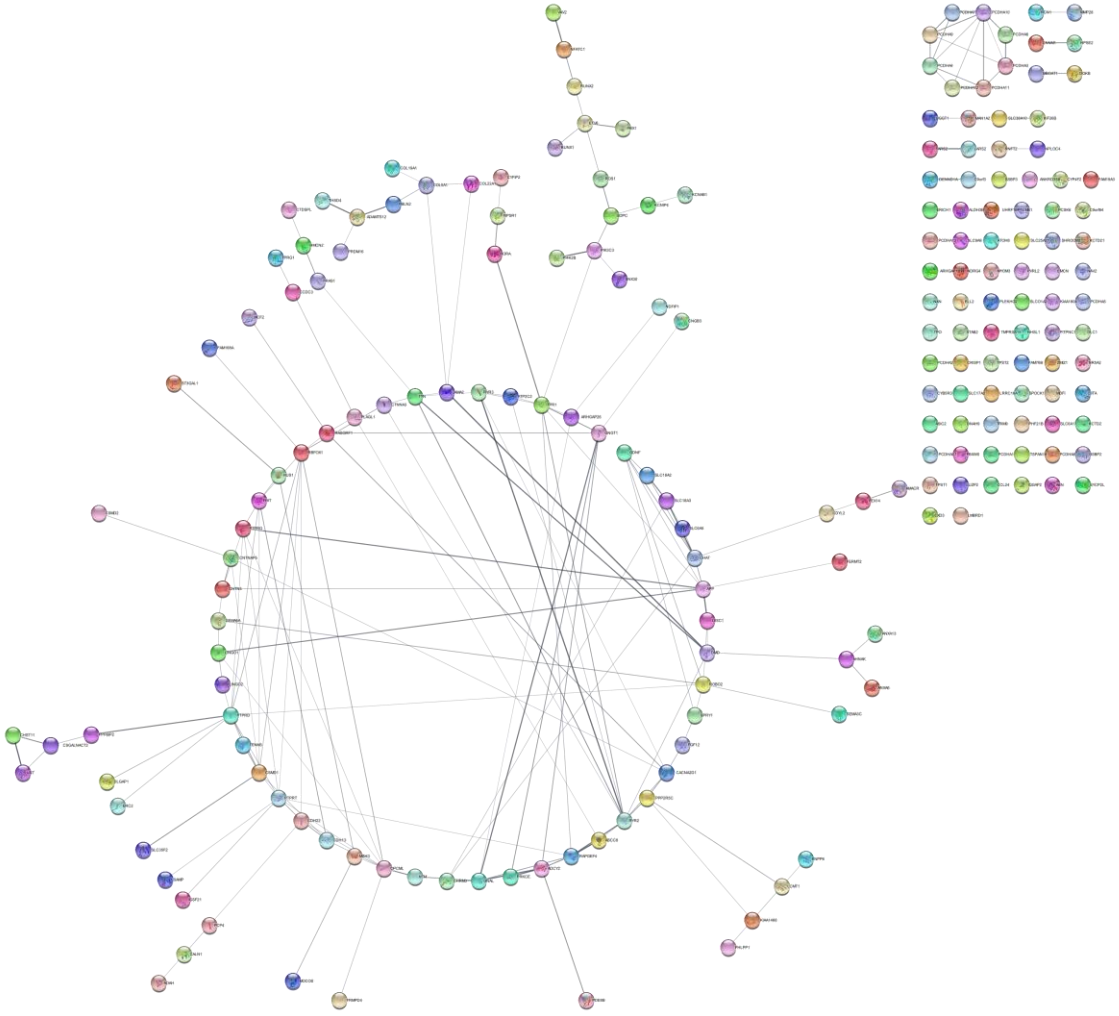

Figure 4.23. Protein-protein interaction network of Ensemble > 2.31 scored only overlapped genes with PPI enrichment p-value 1.00E-16. Nodes represent the genes, while the edges represent the connections between the nodes. The colors of nodes do not have any meaning.

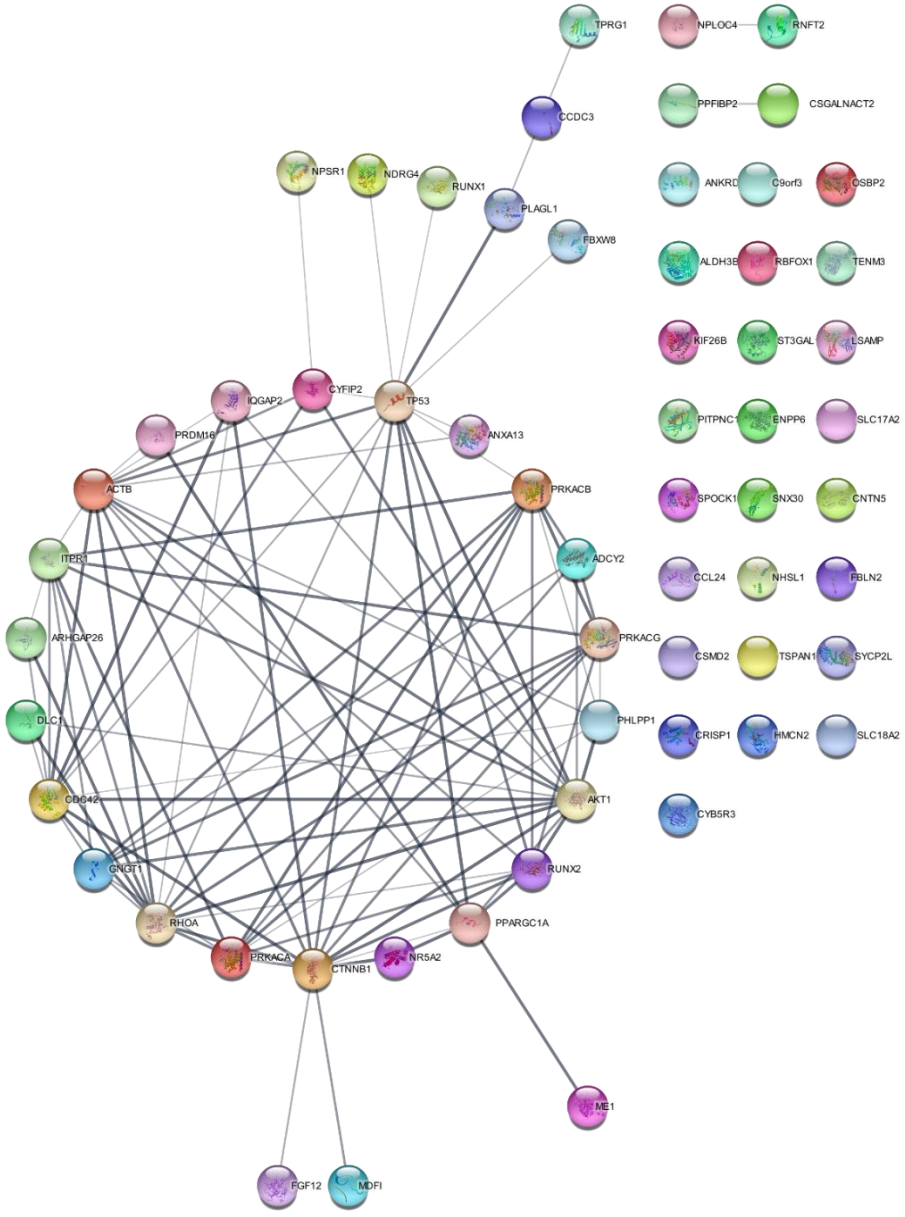

Figure 4.24. Protein-protein interaction network of Ensemble > 3.21 scored for only overlapped genes with PPI enrichment p-value 9.30E-04. Nodes represent the genes, while the edges represent the connections between the nodes. The colors of nodes do not have any meaning.

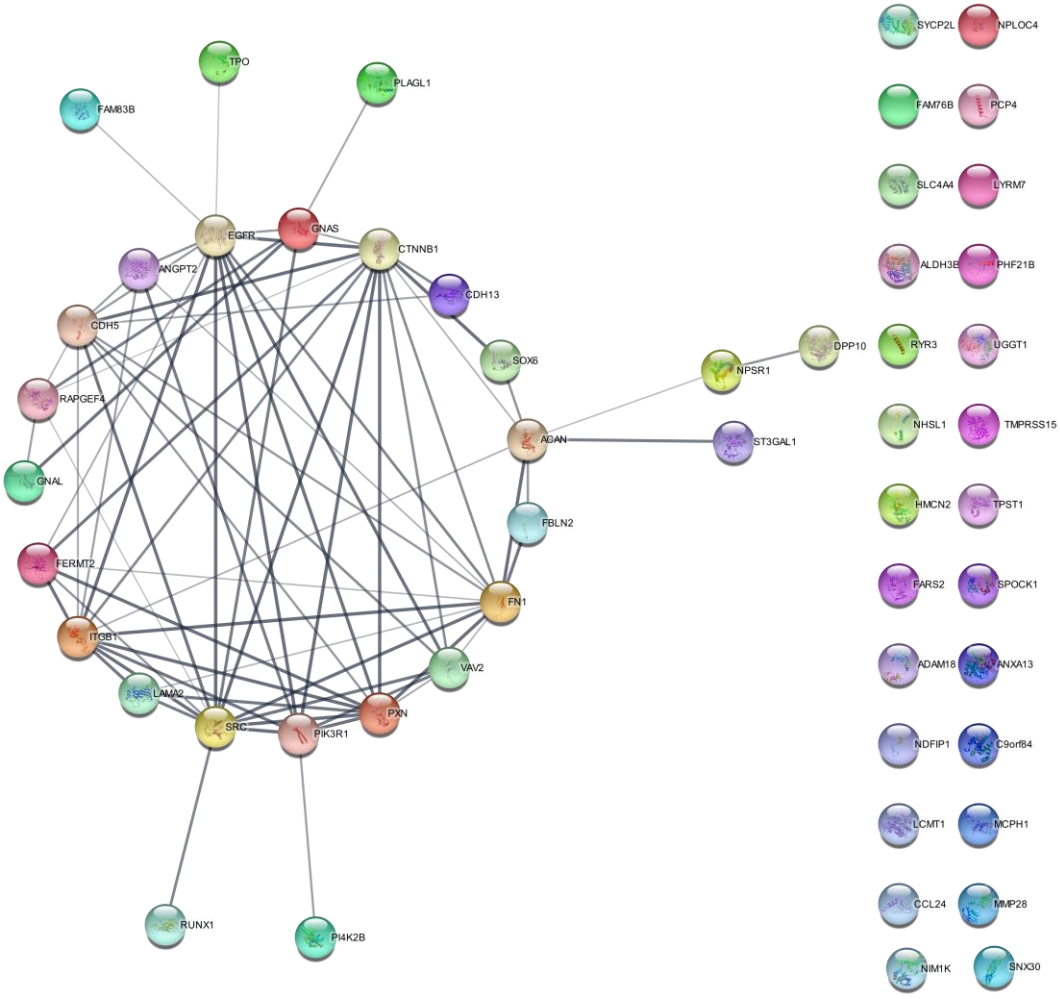

Figure 4.25. Protein-protein interaction network of Entropy for only overlapped genes with the PPI enrichment p-value 3.99E-09. Nodes represent the genes, and the edges represent the connections between the nodes. The colors of nodes do not have any meaning.

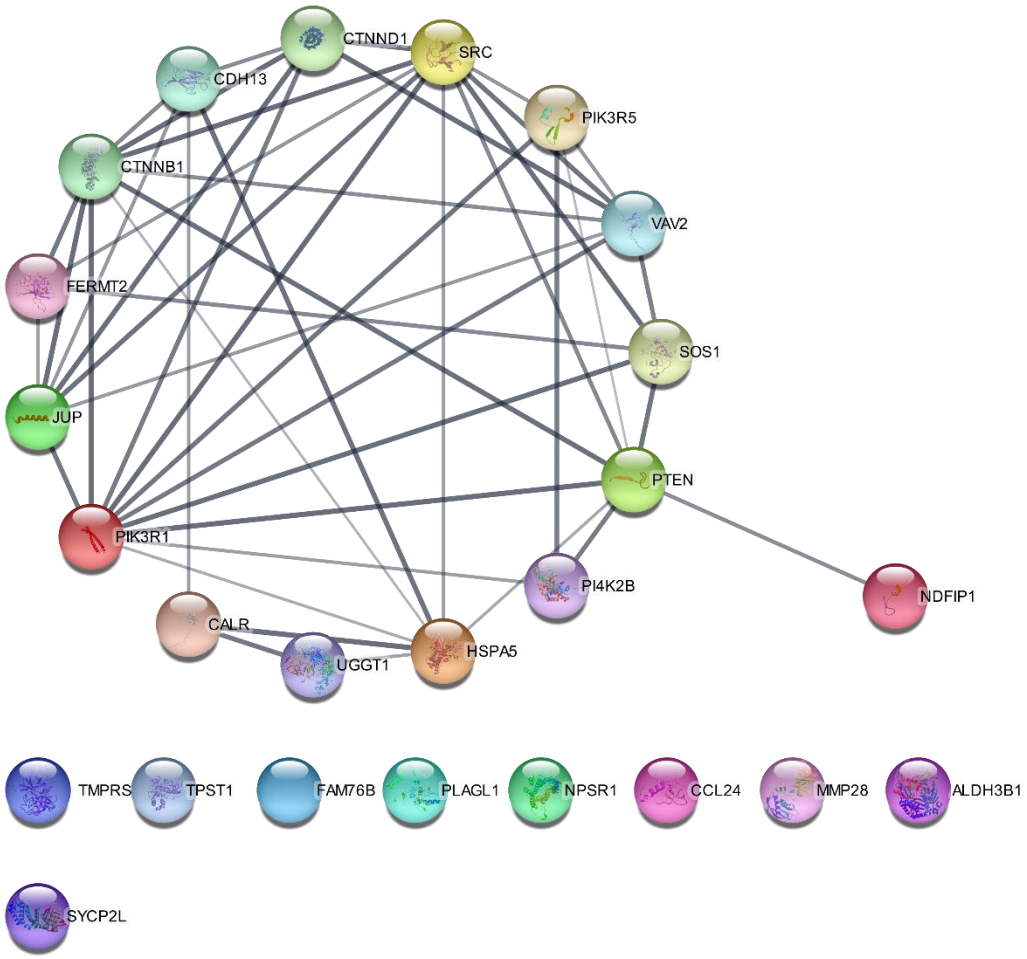

Figure 4.26. Protein-protein interaction network for ADNI only overlapped genes with the PPI enrichment p-value 1.35E-11. Nodes represent the genes, and the edges represent the connections between the nodes. The colors of nodes do not have any meaning.

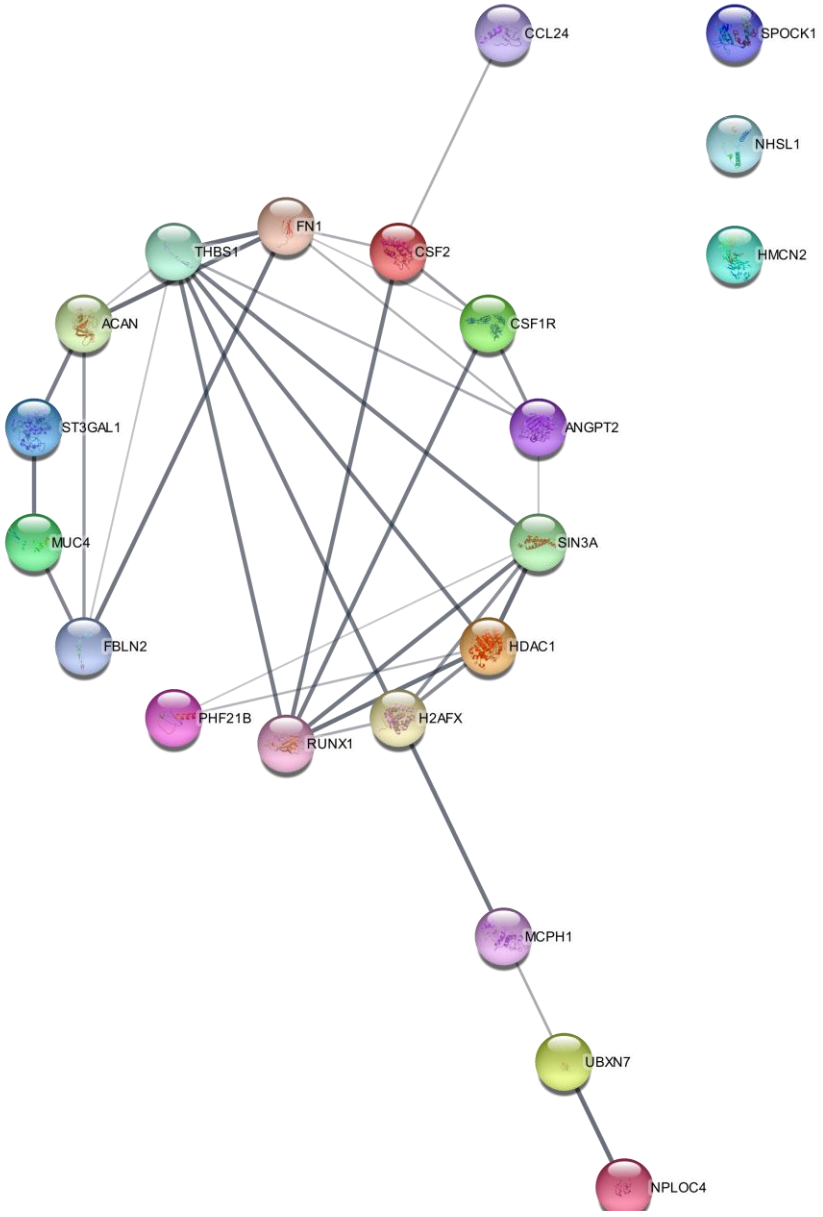

Figure 4.27. Protein-protein interaction network for GenADA only overlapped genes with the PPI enrichment p-value 2.72E-10. Nodes represent the genes, and the edges represent the connections between the nodes. The colors of nodes do not have any meaning.

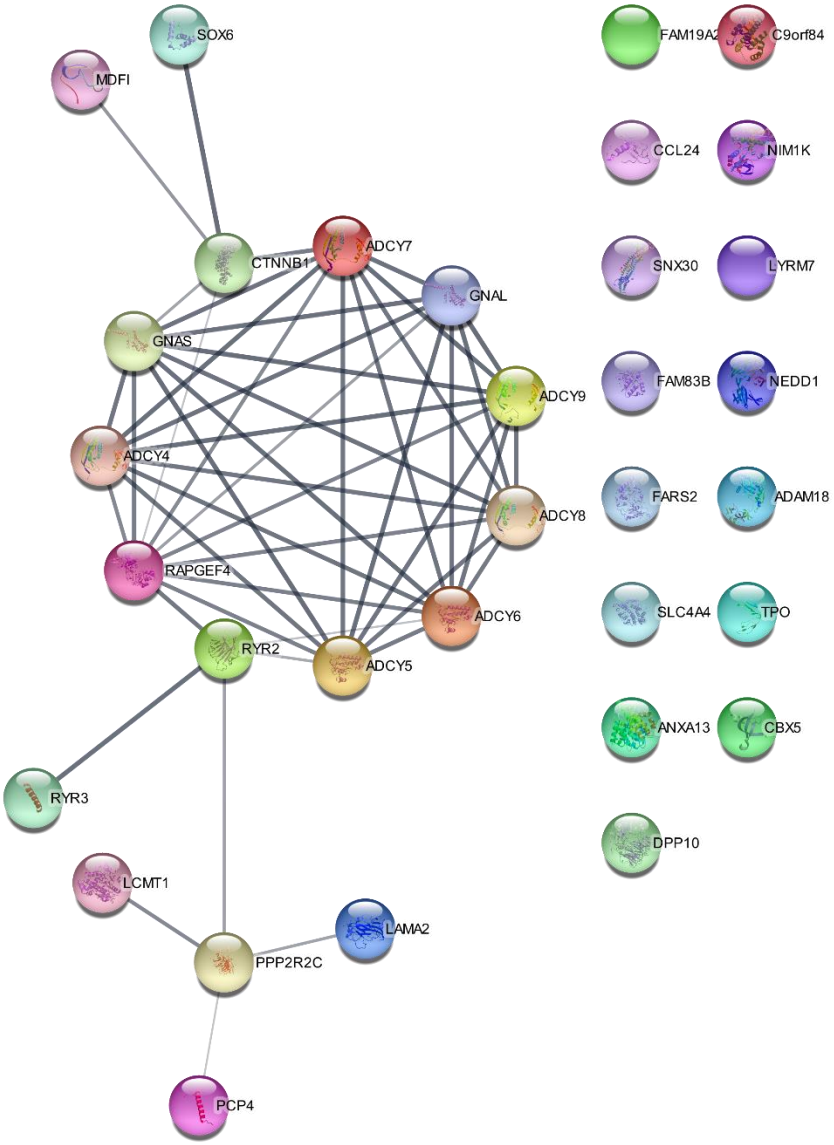

Figure 4.28. Protein-protein interaction network for NCRAD only overlapped genes with the PPI enrichment p-value 1.00E-16. Nodes represent the genes, and the edges represent the connections between the nodes. The colors of nodes do not have any meaning.

4.4. Network genes-candidate diseases networks

Cytoscape plugin HGPEC obtains gene-disease interaction networks. The gene lists consist of extended gene lists with overlapped genes, nearest upstream genes, and nearest downstream genes. Also, for Ensemble >2.31 scored, only overlapped genes are used to compare. The mapped genes are selected as the original gene set, and the diffusion algorithm finds the most relevant nodes and interactions. In the original entropy gene-disease network, 5110 nodes and 56912 edges are observed; 48 nodes and 61 edges are obtained after the Diffusion algorithm. In the original Ensemble gene-disease network, 5105 nodes and 56924 edges were observed; after the Diffusion algorithm, 64 nodes and 128 edges were obtained. Furthermore, the Ensemble >2.31 scored only overlapped genes 5182 nodes and 57441 edges are observed; after the diffusion algorithm, 100 nodes and 40 edges are obtained. The detailed information on the HGPEC analysis results is indicated in **Table 4.6.**

Table 4.6. HGPEC Summary Statistics.

Summary Statistics	ENSEMBLE >2.31	ENSEMBLE >3.21	ENTROPY	Entropy-ADNI	Entropy-GenADA	Entropy-NCRAD
Number of nodes	100	69	51	12	10	51
Number of edges	40	161	98	6	8	80
Average number of neighbors	2.176	5.033	3.951	1.5	2	3.659
Network diameter	5	8	7	2	4	6
Network radius	3	4	4	1	2	3
Characteristic path length	2.66	2.955	2.999	1.5	2.143	2.918
Clustering coefficient	0	0.078	0.063	0	0	0.082
Network density	0.066	0.085	0.099	0.5	0.286	0.091
Network heterogeneity	1.795	0.741	0.772	0.577	0.612	0.606
Network centralization	0.574	0.21	0.212	1	0.381	0.167
Connected components	65	10	11	9	3	6
Analysis time	0.021	0.024	0.028	0.001	0.001	0.004
Efficiency	0.37593985	0.338409475	0.333444481	0.666666667	0.466635558	0.34270048

The networks obtained for Ensemble >2.31, Ensemble>3.21 and Entropy-all are given in the **Figure 4.29**, **Figure 4.30** and **Figure 4.31**, respectively.

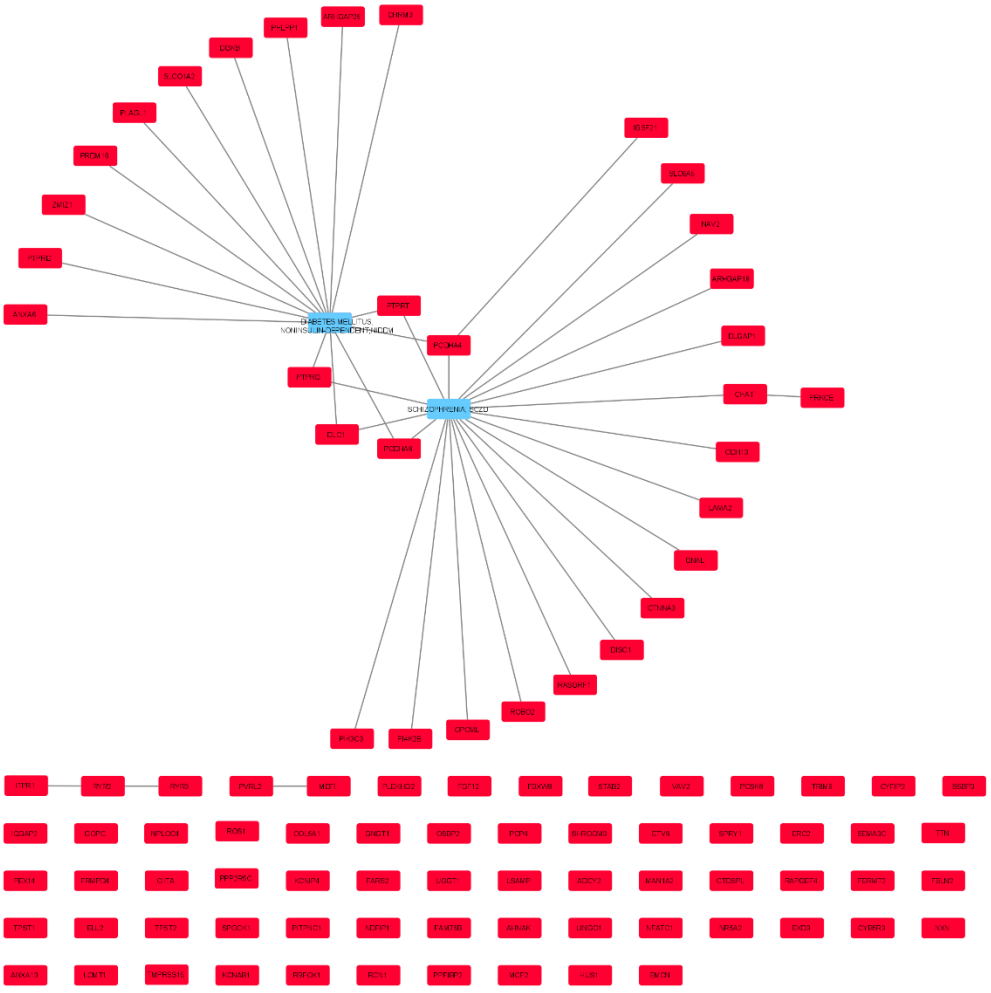

Figure 4.29. Topological relationship between network genes and diseases for Ensemble > 2.31 scored, only overlapped genes. The network obtained by Diffusion rank %20. While red nodes represent network genes, blue nodes represent candidate diseases. Edges represent the relationship between nodes.

The detailed information on genes and diseases in the Ensemble >2.31 network is given in the **Table 4.7 in Appendix C.**

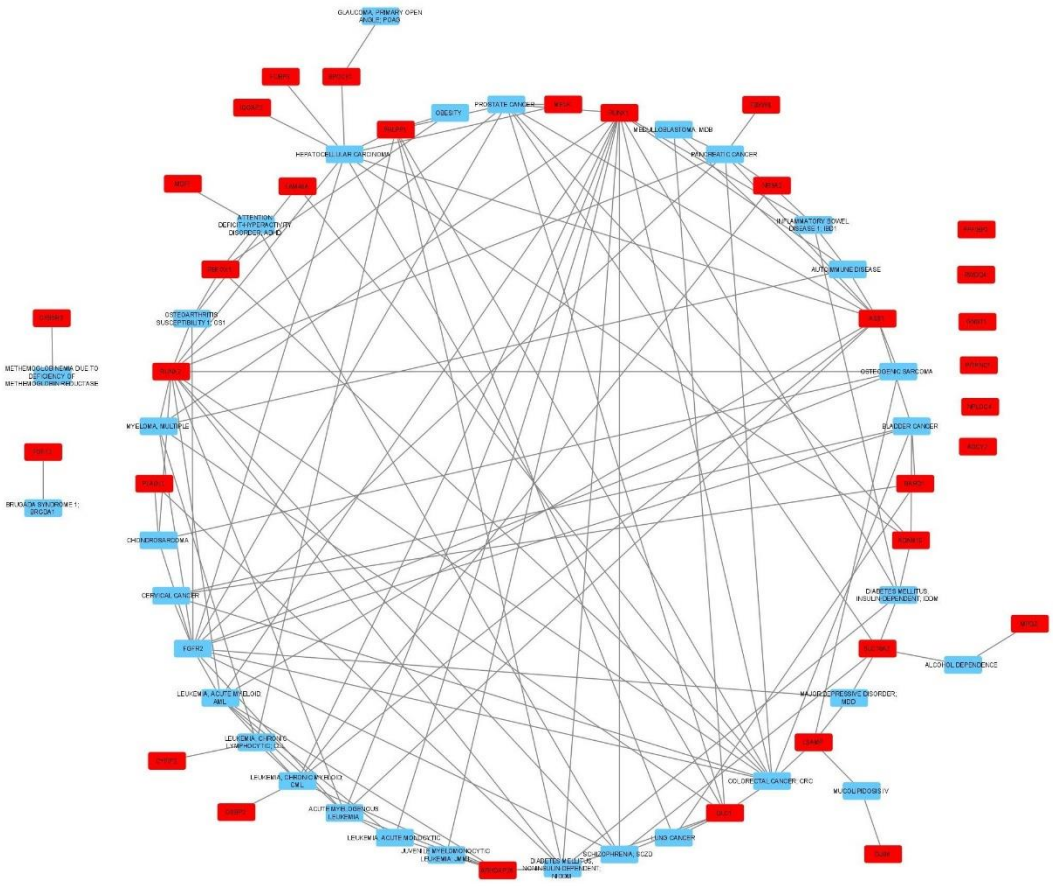

Figure 4.30. Topological relationship between network genes and diseases for Ensemble > 3.21 genes. The network obtained by Diffusion rank %10. While red nodes represent network genes, blue nodes represent candidate diseases. Edges represent the relationship between nodes.

The detailed information on genes and diseases in the Ensemble >3.21 network is given in **Table 4.8.**

Table 4.8. Genes and diseases revealed in **Figure 4.30**.

SHARED NAME	ENTREZ ID	Type	Role	SHARED NAME	ENTREZ ID	Type	Role
SPOCK1	6695	Gene/Protein	Candidate-Gene/Protein	LEUKEMIA, ACUTE MONOCYTIC	MIM151380	Disease	Candidate-Disease
SLC18A2	6571	Gene/Protein	Unknown-Gene/Protein	JUVENILE MYELOMONOCYTIC LEUKEMIA; JMML	MIM607785	Disease	Candidate-Disease
SCHIZOPHRENIA; SCZD	MIM181500	Disease	Candidate-Disease	IQGAP2	10788	Gene/Protein	Unknown-Gene/Protein
RWDD4	201965	Gene/Protein	Unknown-Gene/Protein	INFLAMMATORY BOWEL DISEASE 1; IBD1	MIM266600	Disease	Candidate-Disease
RUNX2	860	Gene/Protein	Unknown-Gene/Protein	HEPATOCELLULAR CARCINOMA	MIM114550	Disease	Candidate-Disease
RUNX1	861	Gene/Protein	Candidate-Gene/Protein	GNGT1	2792	Gene/Protein	Unknown-Gene/Protein
RBFOX1	54715	Gene/Protein	Unknown-Gene/Protein	GLAUCOMA, PRIMARY OPEN ANGLE; POAG	MIM137760	Disease	Candidate-Disease
PROSTATE CANCER	MIM176807	Disease	Candidate-Disease	GJB6	10804	Gene/Protein	Unknown-Gene/Protein
PPFIBP2	8495	Gene/Protein	Unknown-Gene/Protein	FUBP3	8939	Gene/Protein	Unknown-Gene/Protein
PLAGL1	5325	Gene/Protein	Candidate-Gene/Protein	FGFR2	2263	Gene/Protein	Unknown-Gene/Protein
PITPNC1	26207	Gene/Protein	Unknown-Gene/Protein	FGF12	2257	Gene/Protein	Unknown-Gene/Protein
PHLPP1	23239	Gene/Protein	Unknown-Gene/Protein	FBXW8	26259	Gene/Protein	Unknown-Gene/Protein
PANCREATIC CANCER	MIM260350	Disease	Candidate-Disease	FAM46A	55603	Gene/Protein	Unknown-Gene/Protein
OSTEOGENIC SARCOMA	MIM259500	Disease	Candidate-Disease	DLC1	10395	Gene/Protein	Unknown-Gene/Protein
OSTEOARTHRITIS SUSCEPTIBILITY 1; OS1	MIM165720	Disease	Candidate-Disease	DIABETES MELLITUS, NONINSULIN-DEPENDENT; NIDDM	MIM125853	Disease	Candidate-Disease
OSBP2	23762	Gene/Protein	Unknown-Gene/Protein	DIABETES MELLITUS, INSULIN-DEPENDENT; IDDM	MIM222100	Disease	Candidate-Disease
OBESITY	MIM601665	Disease	Candidate-Disease	CYFIP2	26999	Gene/Protein	Unknown-Gene/Protein

NR5A2	2494	Gene/Protein	Unknown-Gene/Protein	CYB5R3	1727	Gene/Protein	Unknown-Gene/Protein
NPLOC4	55666	Gene/Protein	Candidate-Gene/Protein	COLORECTAL CANCER; CRC	MIM114500	Disease	Candidate-Disease
MYELOMA, MULTIPLE	MIM254500	Disease	Candidate-Disease	CHONDROSARCOMA	MIM215300	Disease	Candidate-Disease
MUCOLIPIDOSIS IV	MIM252650	Disease	Candidate-Disease	CERVICAL CANCER	MIM603956	Disease	Candidate-Disease
MPDZ	8777	Gene/Protein	Unknown-Gene/Protein	BRUGADA SYNDROME 1; BRGDA1	MIM601144	Disease	Candidate-Disease
METHEMOGLOBINEMIA DUE TO DEFICIENCY OF METHEMOGLOBIN REDUCTASE	MIM250800	Disease	Candidate-Disease	BLADDER CANCER	MIM109800	Disease	Candidate-Disease
MELK	9833	Gene/Protein	Unknown-Gene/Protein	BARD1	580	Gene/Protein	Unknown-Gene/Protein
MEDULLOBLASTOMA; MDB	MIM155255	Disease	Candidate-Disease	AUTOIMMUNE DISEASE	MIM109100	Disease	Candidate-Disease
MDFI	4188	Gene/Protein	Unknown-Gene/Protein	ATTENTION DEFICIT-HYPERACTIVITY DISORDER; ADHD	MIM143465	Disease	Candidate-Disease
MAJOR DEPRESSIVE DISORDER; MDD	MIM608516	Disease	Candidate-Disease	ASS1	445	Gene/Protein	Unknown-Gene/Protein
LUNG CANCER	MIM211980	Disease	Candidate-Disease	ARHGAP26	23092	Gene/Protein	Unknown-Gene/Protein
LSAMP	4045	Gene/Protein	Unknown-Gene/Protein	ALCOHOL DEPENDENCE	MIM103780	Disease	Candidate-Disease
LEUKEMIA, CHRONIC MYELOID; CML	MIM608232	Disease	Candidate-Disease	ADCY2	108	Gene/Protein	Unknown-Gene/Protein
LEUKEMIA, CHRONIC LYMPHOCYTIC; CLL	MIM151400	Disease	Candidate-Disease	ADAM10	102	Gene/Protein	Unknown-Gene/Protein
LEUKEMIA, ACUTE MYELOID; AML	MIM601626	Disease	Candidate-Disease	ACUTE MYELOGENOUS LEUKEMIA	MIM602439	Disease	Candidate-Disease

The network obtained for the Entropy gene set is given in **Figure 4.13,** and the network obtained for the Ensemble gene set is given in **Figure 4.14.**

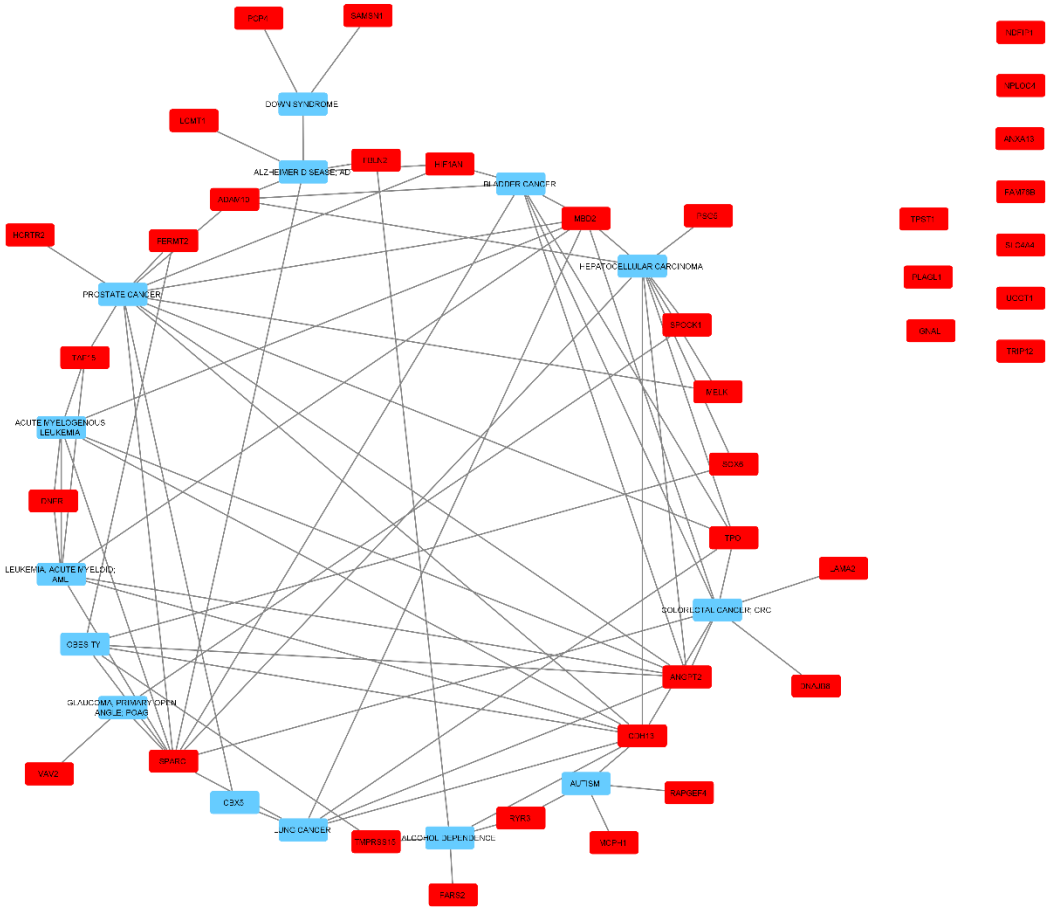

Figure 4.31. Topological relationship between network genes and diseases for all Entropy genes. The network obtained by Diffusion rank %10. While red nodes represent network genes, blue nodes represent candidate diseases. Edges represent the relationship between nodes.

The **Table 4.9.** shows the gene/disease names, Entrez ID's, types and roles of the entities found in the **Figure 4.31**.

Table 4.9. Detailed information on genes and diseases is found in **Figure 4.31**.

SHARED NAME	EntrezID	Type	Role	SHARED NAME	EntrezID	Type	Role
TRIP12	9320	Gene/Protein	Unknown-Gene/Protein	LEUKEMIA, ACUTE MYELOID; AML	MIM601626	Disease	Candidate-Disease
BLADDER CANCER	MIM109800	Disease	Candidate-Disease	NPLOC4	55666	Gene/Protein	Candidate-Gene/Protein
LUNG CANCER	MIM211980	Disease	Candidate-Disease	DNAJB8	165721	Gene/Protein	Unknown-Gene/Protein
SAMSN1	64092	Gene/Protein	Unknown-Gene/Protein	PSG5	5673	Gene/Protein	Unknown-Gene/Protein
FAM76B	143684	Gene/Protein	Candidate-Gene/Protein	TPO	7173	Gene/Protein	Unknown-Gene/Protein
FARS2	10667	Gene/Protein	Unknown-Gene/Protein	AUTISM	MIM209850	Disease	Candidate-Disease
FERMT2	10979	Gene/Protein	Candidate-Gene/Protein	ANXA13	312	Gene/Protein	Unknown-Gene/Protein
ACUTE MYELOGENOUS LEUKEMIA	MIM602439	Disease	Candidate-Disease	VAV2	7410	Gene/Protein	Candidate-Gene/Protein
ADAM10	102	Gene/Protein	Unknown-Gene/Protein	FBLN2	2199	Gene/Protein	Candidate-Gene/Protein
CDH13	1012	Gene/Protein	Candidate-Gene/Protein	ANGPT2	285	Gene/Protein	Candidate-Gene/Protein
UGGT1	56886	Gene/Protein	Candidate-Gene/Protein	DNER	92737	Gene/Protein	Unknown-Gene/Protein
MELK	9833	Gene/Protein	Unknown-Gene/Protein	ALZHEIMER DISEASE; AD	MIM104300	Disease	Training-Disease
SLC4A4	8671	Gene/Protein	Unknown-Gene/Protein	SPARC	6678	Gene/Protein	Unknown-Gene/Protein
PCP4	5121	Gene/Protein	Unknown-Gene/Protein	SOX6	55553	Gene/Protein	Unknown-Gene/Protein
CBX5	23468	Gene/Protein	Unknown-Gene/Protein	MBD2	8932	Gene/Protein	Unknown-Gene/Protein
MCPH1	79648	Gene/Protein	Candidate-Gene/Protein	TPST1	8460	Gene/Protein	Candidate-Gene/Protein
DOWN SYNDROME	MIM190685	Disease	Candidate-Disease	LCMT1	51451	Gene/Protein	Unknown-Gene/Protein
RYR3	6263	Gene/Protein	Unknown-Gene/Protein	HEPATOCELLULAR CARCINOMA	MIM114550	Disease	Candidate-Disease

ALCOHOL DEPENDENCE	MIM103780	Disease	Candidate-Disease	TMPRSS15	5651	Gene/Protein	Candidate-Gene/Protein
GNAL	2774	Gene/Protein	Unknown-Gene/Protein	SPOCK1	6695	Gene/Protein	Candidate-Gene/Protein
LAMA2	3908	Gene/Protein	Unknown-Gene/Protein	NDFIP1	80762	Gene/Protein	Candidate-Gene/Protein
PLAGL1	5325	Gene/Protein	Candidate-Gene/Protein	TAF15	8148	Gene/Protein	Candidate-Gene/Protein
RAPGEF4	11069	Gene/Protein	Unknown-Gene/Protein	GLAUCOMA, PRIMARY OPEN ANGLE; POAG	MIM137760	Disease	Candidate-Disease
HCRTR2	3062	Gene/Protein	Unknown-Gene/Protein	COLORECTAL CANCER; CRC	MIM114500	Disease	Candidate-Disease
HIF1AN	55662	Gene/Protein	Unknown-Gene/Protein	OBESITY	MIM601665	Disease	Candidate-Disease
PROSTATE CANCER	MIM176807	Disease	Candidate-Disease				

4.5. Network genes-Alzheimer's disease networks

Biological networks constructed for extended gene sets by the HGPEC plugin are represented in **Figure 4.32, Figure 4.33, Figure 4.34, Figure 4.35, Figure 4.36, and Figure 4.37.** In addition, the detailed summary statistics of the networks are given in **Table 4.10.**

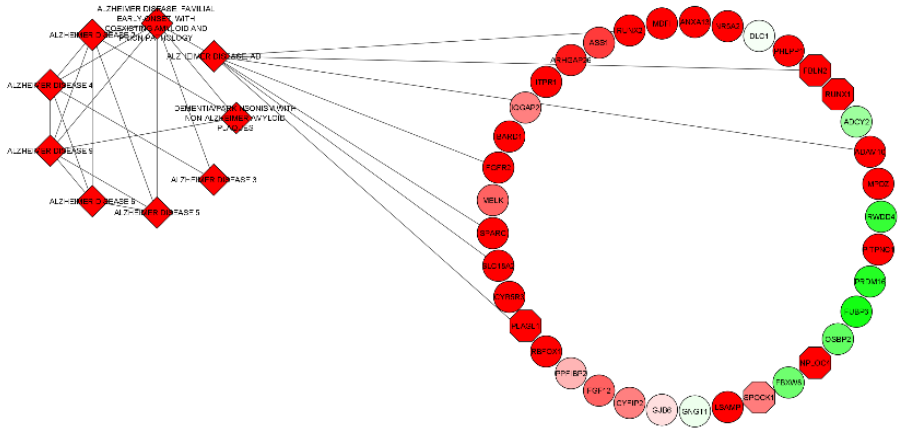

Figure 4.32. Network representation of Ensemble >3.21 genes and Alzheimer's disease. While network genes are represented with an octagon and unknown genes are represented with an ellipse, training diseases are represented with a diamond. The color mapping type is a continuous mapping according to the rank of the nodes, from red to green. Thus, the highest rank is represented with red, and the lowest rank is represented with green. Edges show connections between genes and diseases.

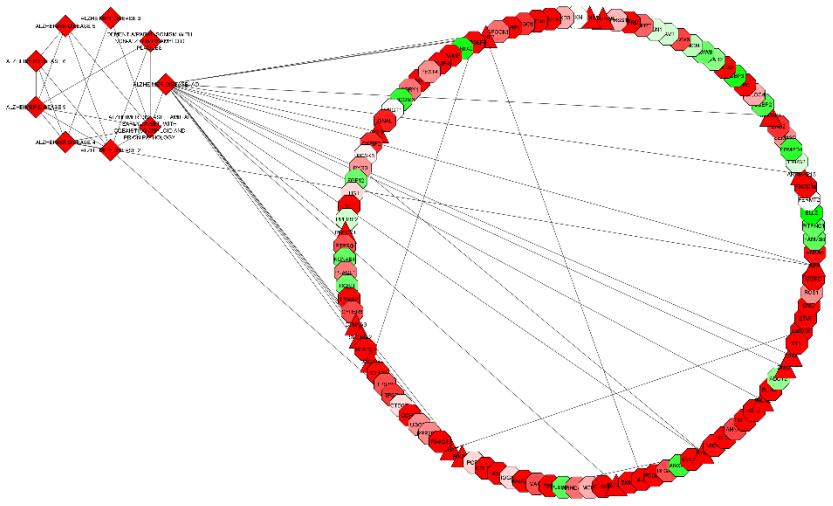

Figure 4.33. Network representation of Ensemble > 2.31 only overlapped genes and Alzheimer's disease. While network genes are represented with an octagon, training genes are represented with a triangle, and unknown genes are represented with an ellipse, training diseases are represented with a diamond. The color mapping type is a continuous mapping according to the rank of the nodes, from red to green. Thus, the highest rank is represented with red, and the lowest rank is represented with green. Edges show connections between genes and diseases.

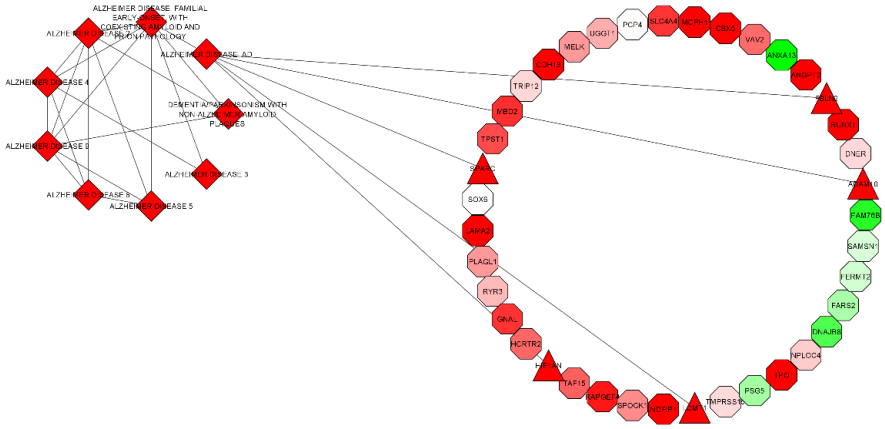

Figure 4.34. Network representation of all Entropy genes and Alzheimer's disease. While network genes are represented with an octagon and unknown genes are represented with an ellipse, training diseases are represented with a diamond. The color mapping type is a continuous mapping according to the rank of the nodes, from red to green. Thus, the highest rank is represented with red, and the lowest rank is represented with green. Edges show connections between genes and diseases.

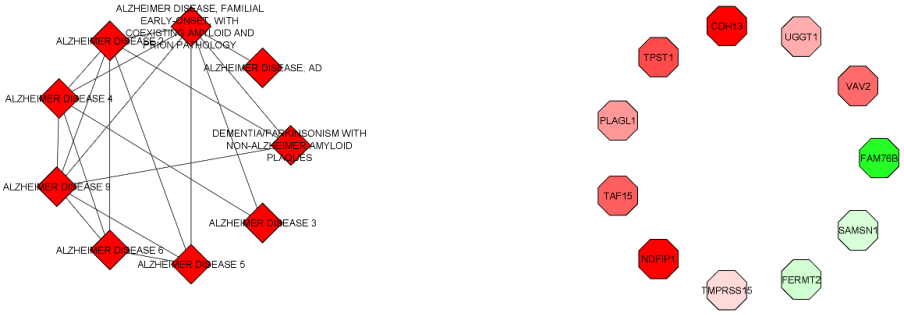

Figure 4.35. Network representation of Entropy-ADNI genes and Alzheimer's disease. While network genes are represented with an octagon and unknown genes are represented with an ellipse, training diseases are represented with a diamond. The color mapping type is a continuous mapping according to the rank of the nodes, from red to green. Thus, the highest rank is represented with red, and the lowest rank is represented with green. Edges show connections between genes and diseases.

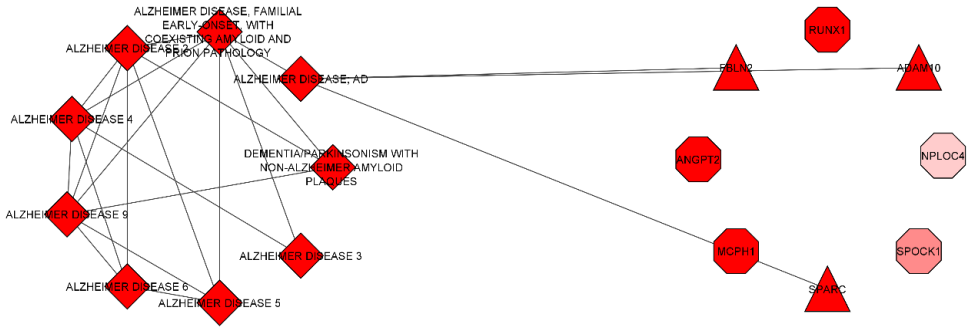

Figure 4.36. Network representation of Entropy-GenADA genes and Alzheimer's disease. While network genes are represented with an octagon and unknown genes are represented with an ellipse, training diseases are represented with a diamond. The color mapping type is a continuous mapping according to the rank of the nodes, from red to green. Thus, the highest rank is represented with red, and the lowest rank is represented with green. Edges show connections between genes and diseases.

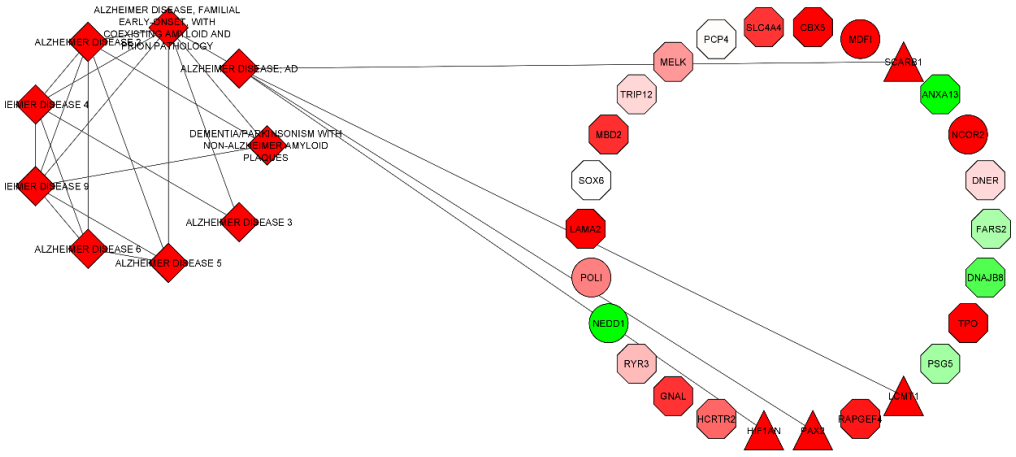

Figure 4.37. Network representation of Entropy-NCRAD genes and Alzheimer's disease. While network genes are represented with an octagon and unknown genes are represented with an ellipse, training diseases are represented with a diamond. The color mapping type is a continuous mapping according to the rank of the nodes, from red to green. Thus, the highest rank is represented with red, and the lowest rank is represented with green. Edges show connections between genes and diseases.

The detailed result of summary statistics of network genes and Alzheimer's Disease networks is given in **Table 4.10**.

Table 4.10. Summary statistics of network genes and Alzheimer's Disease networks.

Summary Statistics	ENSEMBLE >3.21	ENSEMBLE > 2.31	ENTROPY	Entropy-ADNI	Entropy-GenADA	Entropy-NCRAD
Number of nodes	46	120	48	20	17	35
Number of edges	26	45	24	19	22	21
Average number of neighbors	3.25	2.750	3.429	4.222	3.667	3.538
Network diameter	4	5	4	3	4	4
Network radius	2	3	2	2	2	2
Characteristic path length	2.258	2.530	2.187	1.5	2.045	2.128
Clustering coefficient	0.374	0.175	0.427	0.665	0.498	0.46
Network density	0.217	0.089	0.264	0.528	0.333	0.295
Network heterogeneity	0.742	1.256	0.64	0.443	0.549	0.593
Network centralization	0.362	0.559	0.321	0.446	0.364	0.341
Connected components	31	88	35	12	6	23
Analysis time	0.006	0.080	0.037	0.002	0.002	0.007

4.6.Knowledge graphs

Knowledge graphs created by CROssBAR are used to obtain associations of genes, pathways, and diseases. The gene lists consist of extended gene lists which have overlapped genes, nearest upstream genes, and nearest downstream genes. The knowledge graphs created for Ensemble >3.21, Ensemble >2.31, Entropy, ADNI, GenADA, and NCRAD genes with Alzheimer's disease are represented in **Figure 4.38**, **Figure 4.39, Figure 4.40**, **Figure 4.41**, **Figure 4.42, Figure 4.43** and **Figure 4.44**.

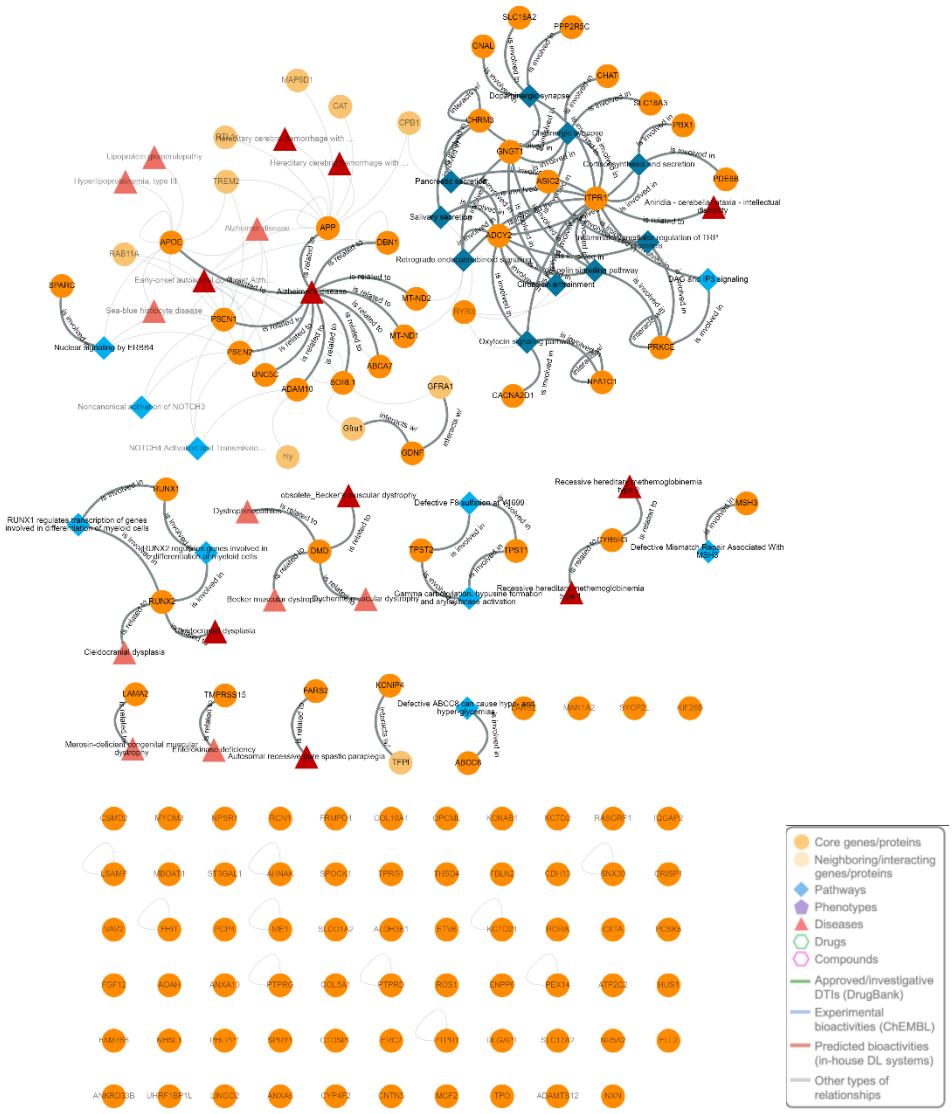

Figure 4.38. Ensemble > 2.31 scored genes knowledge graph created by CROssBAR. Dark orange nodes represent core genes; light orange nodes represent interacting genes/proteins. Blue squares represent pathways, and purple pentagons represent phenotypes. Edges represent the connections between the nodes. Genes related to the "Alzheimer's Disease" component and genes in our Ensemble list are highlighted.

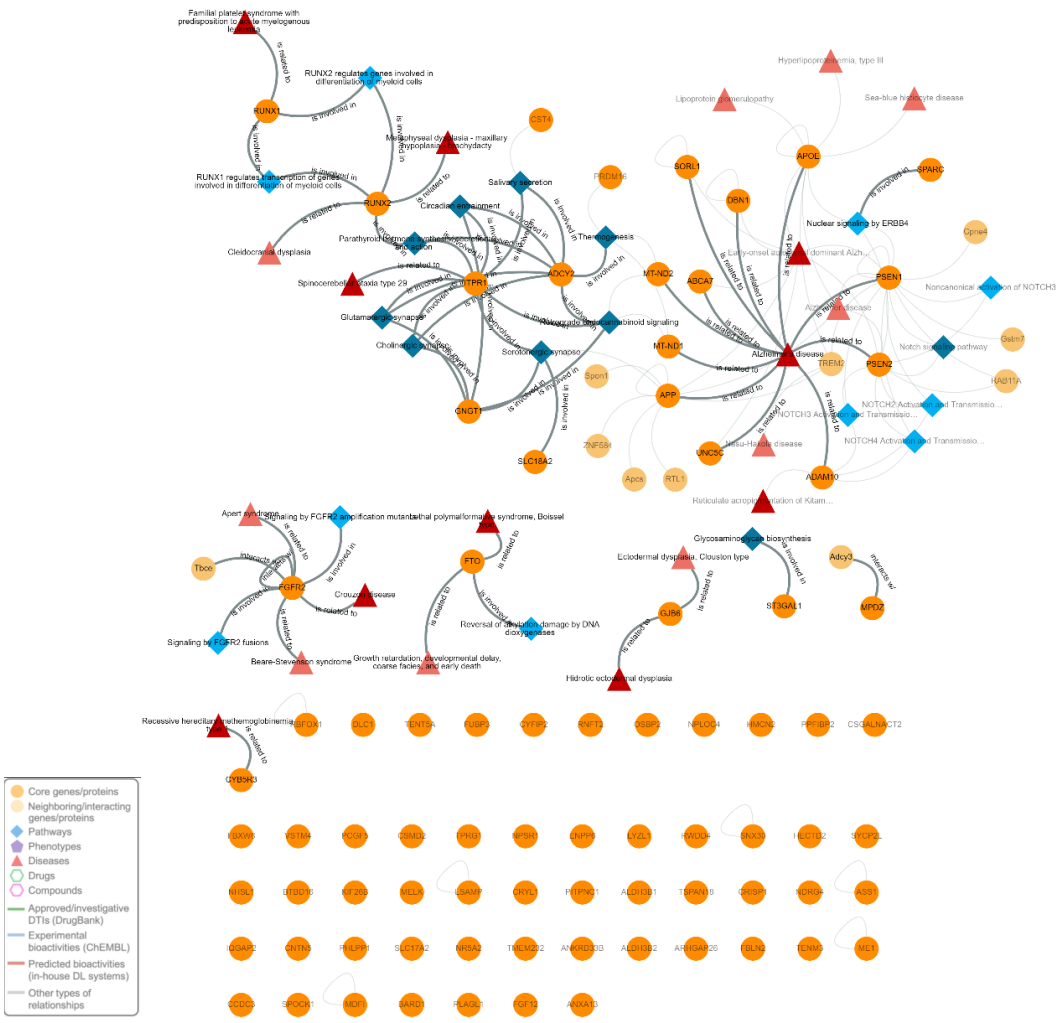

Figure 4.39. Ensemble >3.21 scored genes knowledge graph created by CROssBAR. Dark orange nodes represent core genes; light orange nodes represent interacting genes/proteins. Blue squares represent pathways, and purple pentagons represent phenotypes. Edges represent the connections between the nodes. Genes related to the "Alzheimer's Disease" component and genes in our Ensemble list are highlighted.

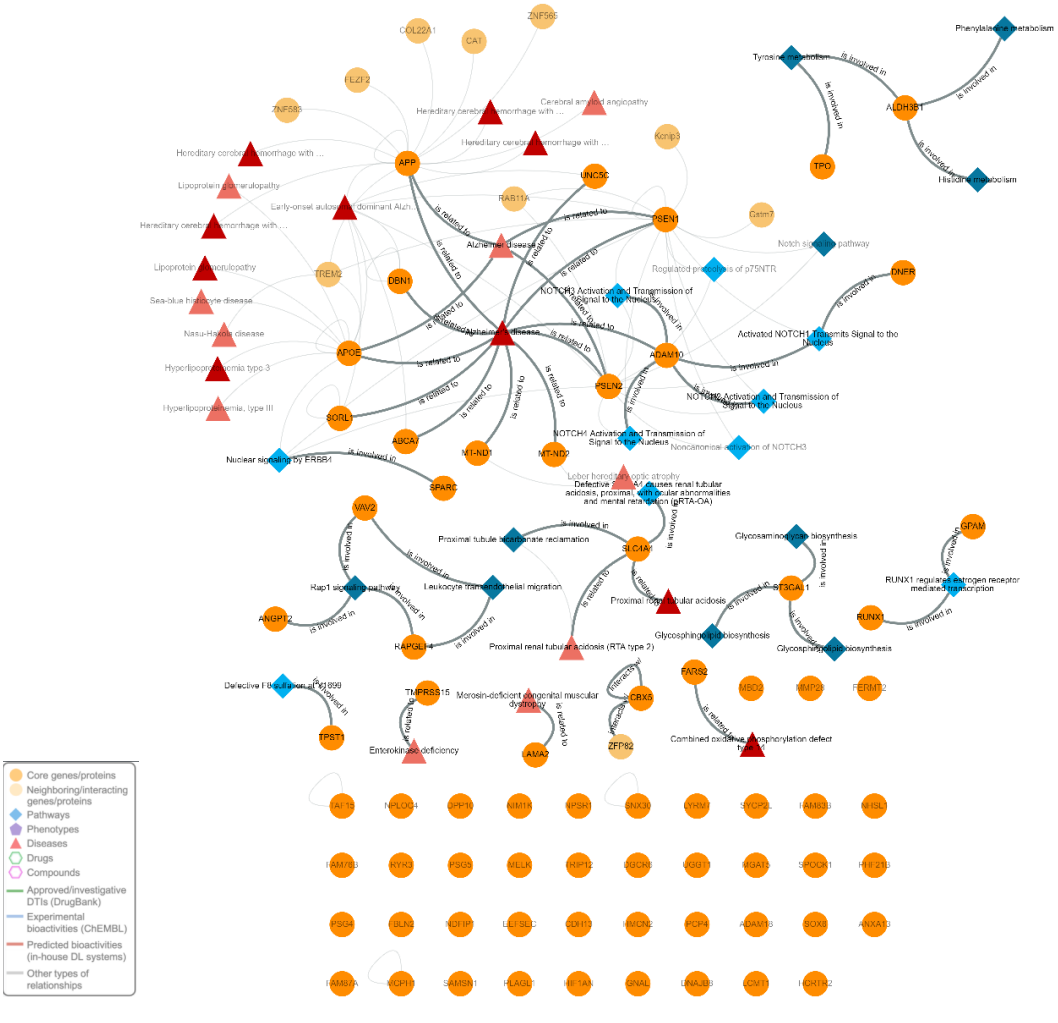

Figure 4.40. Entropy genes knowledge graph created by CROssBAR. Dark orange nodes represent core genes; light orange nodes represent interacting genes/proteins. Blue squares represent pathways, and purple pentagons represent phenotypes. Edges represent the connections between the nodes. Genes related to the “Alzheimer’s Disease” component and genes in our Entropy list are highlighted.

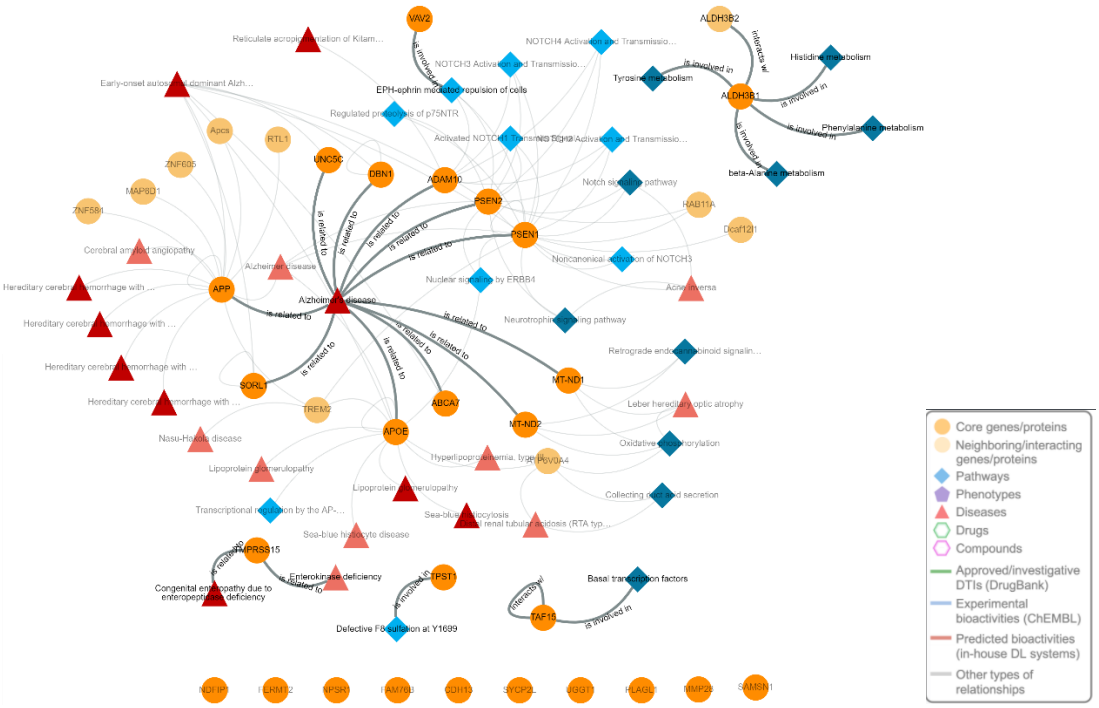

Figure 4.41. Entropy-ADNI genes knowledge graph created by CROssBAR. Dark orange nodes represent core genes; light orange nodes represent interacting genes/proteins. Blue squares represent pathways, and purple pentagons represent phenotypes. Edges represent the connections between the nodes. Genes related to the "Alzheimer's Disease" component and genes in our ADNI list are highlighted.

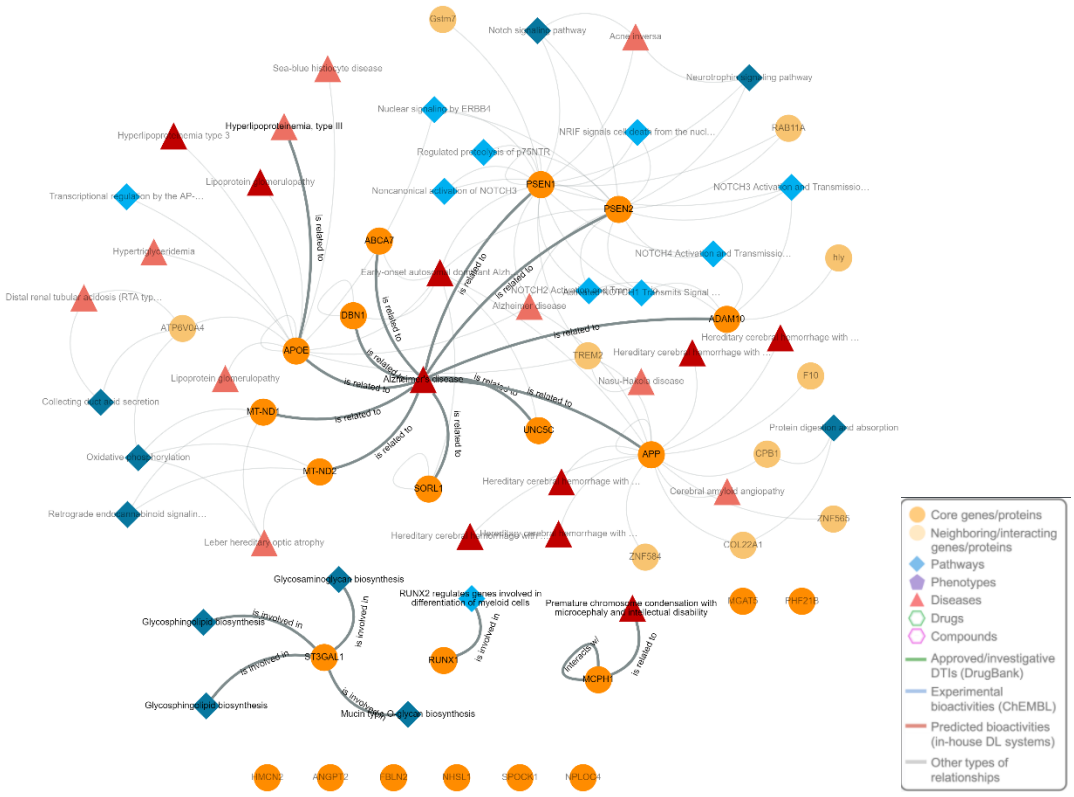

Figure 4.42. Entropy-GenADA genes knowledge graph created by CROssBAR. Dark orange nodes represent core genes; light orange nodes represent interacting genes/proteins. Blue squares represent pathways, and purple pentagons represent phenotypes. Edges represent the connections between the nodes. Genes related to the "Alzheimer's Disease" component and genes in our GenADA list are highlighted.

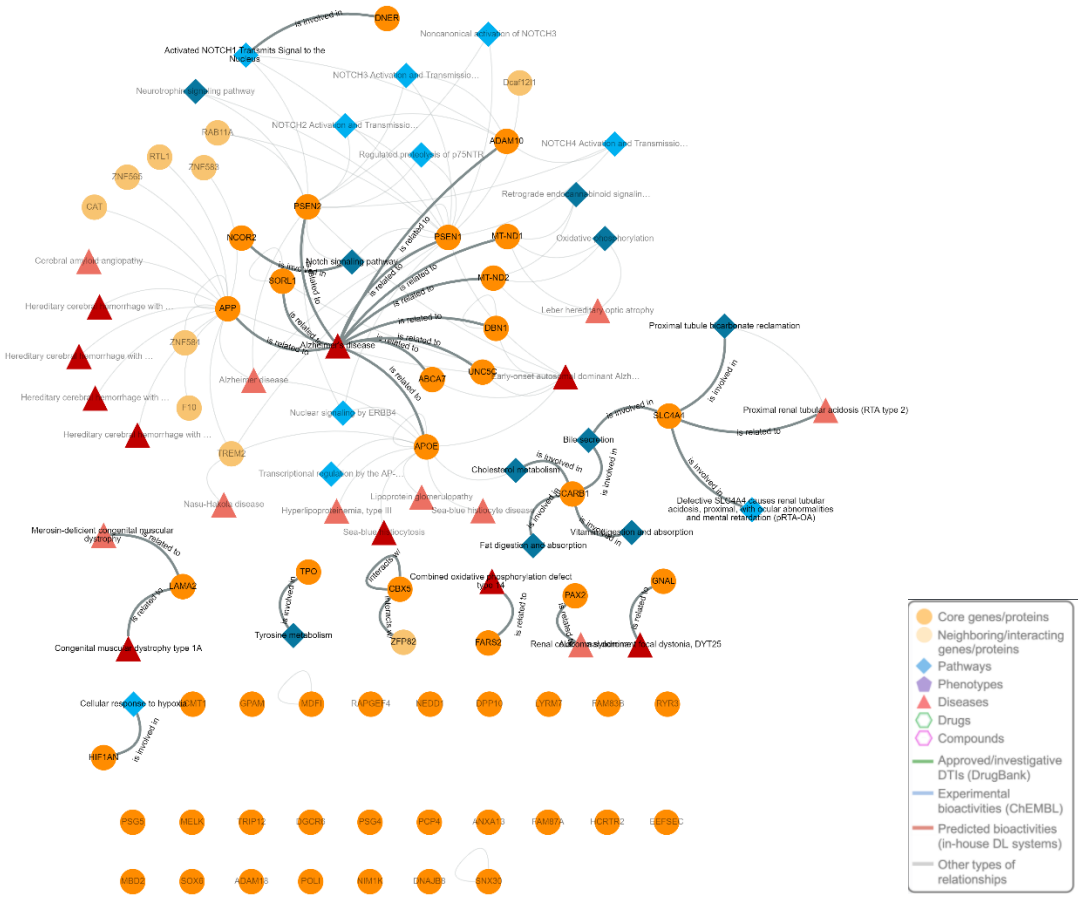

Figure 4.43. Entropy-NCRAD genes knowledge graph created by CROssBAR. Dark orange nodes represent core genes; light orange nodes represent interacting genes/proteins. Blue squares represent pathways, and purple pentagons represent phenotypes. Edges represent the connections between the nodes. Genes related to the "Alzheimer's Disease" component and genes in our NCRAD list are highlighted.

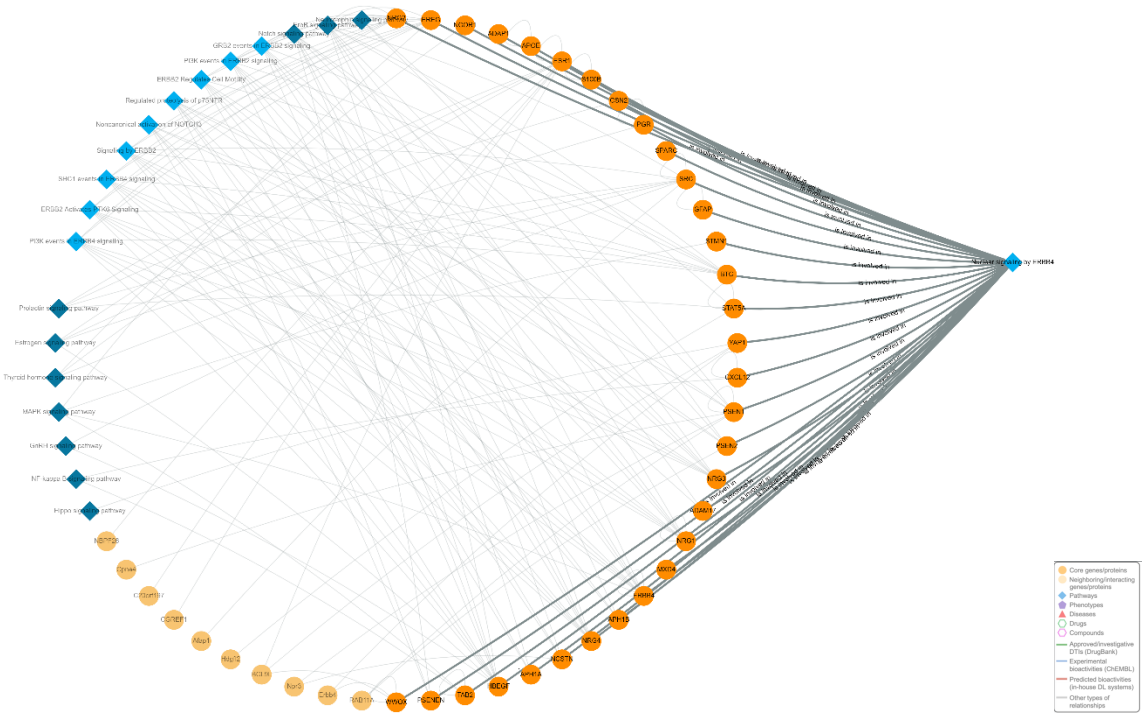

Figure 4.44. CROssBAR knowledge graph created for “Nuclear signaling by ERRB4” pathway. Dark orange nodes represent core genes; light orange nodes represent interacting genes/proteins. Blue squares represent pathways. Edges represent the connections between the nodes.

This figure is obtained to investigate the genes involved in the “Nuclear signaling by ERBB4” pathway coming with CROssBAR analysis. We have added this figure to compare the genes that are related to the “Nuclear signaling by ERBB4” with the genes in our lists and to find out if there are any common genes. As the result of this observation we have found out that SPARC gene takes part both in our gene lists (except entropy-ADNI and entropy-NCRAD) and “Nuclear signaling by ERBB4” pathway.

4.7. Functional enrichment analysis

Functional enrichment analysis of genes of interest is done by the following analysis the g: Profiler, EnrichmentMap, and AutoAnnotate pipeline. The functional enrichment analysis results are obtained for Ensemble only overlapped and Entropy extended gene sets. Ensemble the whole gene list for this analysis, consisting of variants scored 1.11 and over. For Ensemble, only overlapped genes are used, while for Entropy, the extended gene list is used, which consists of overlapped genes, nearest downstream genes, and nearest upstream genes. Clusters obtained for Ensemble and Entropy gene lists are presented in **Figure 4.45** and **Figure 4.46**.

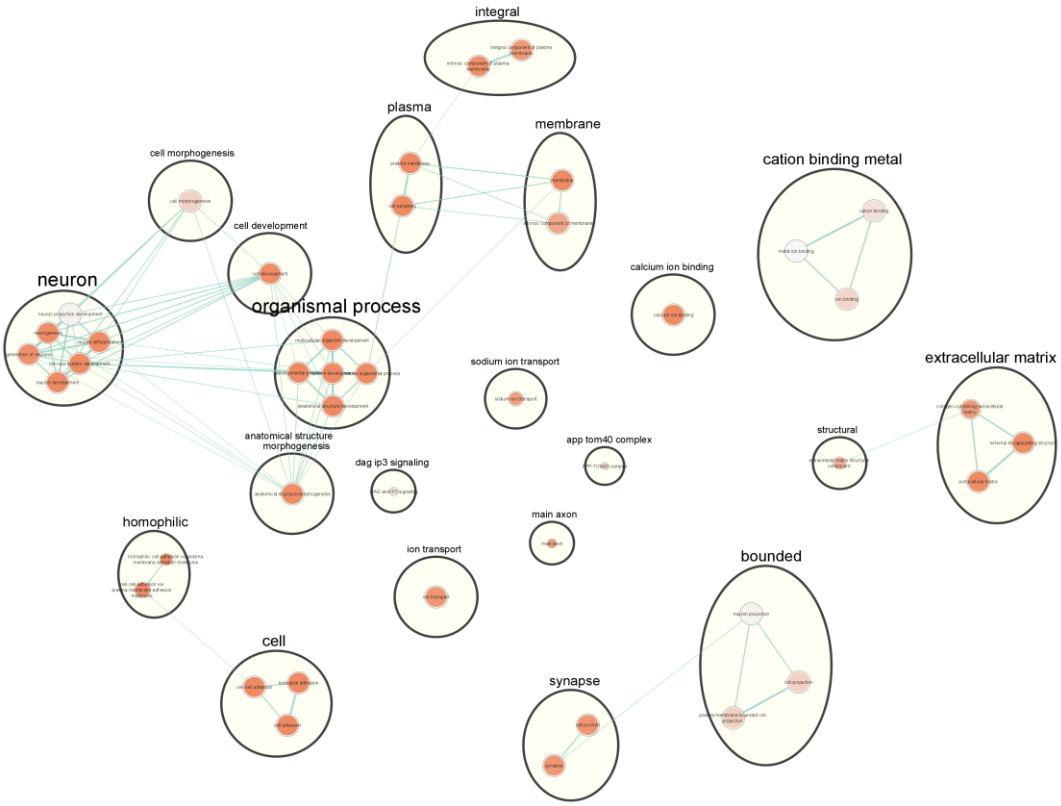

Figure 4.45. Enrichment analysis results for **Ensemble > 1.11** scored gene list. Ensemble only overlapped genes list is used for this analysis.

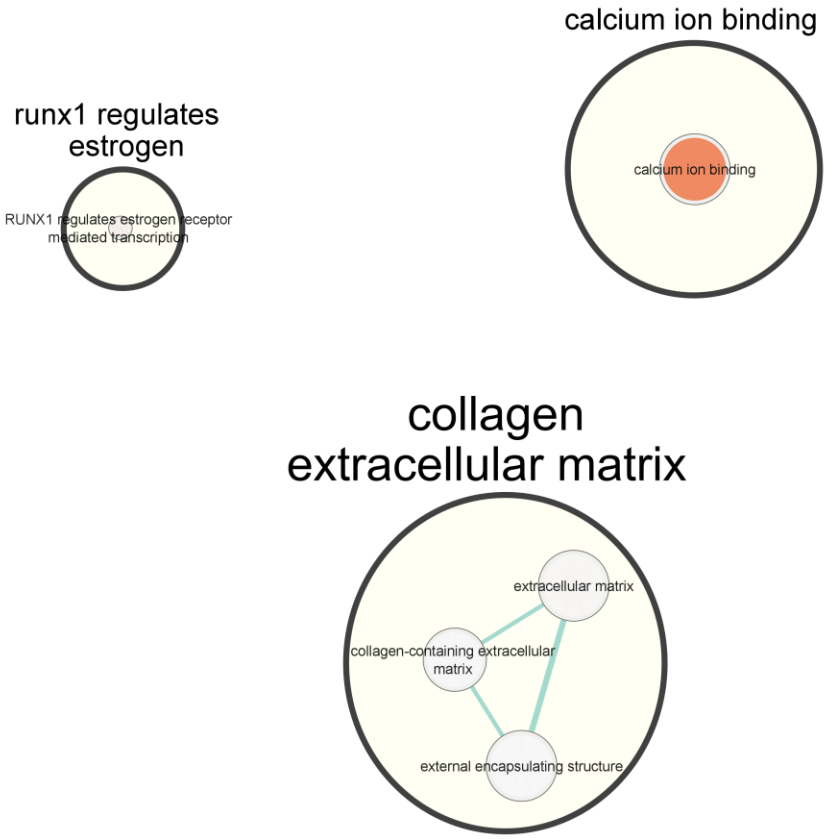

Figure 4.46. Enrichment analysis results for Entropy gene list. For this analysis, Entropy extended gene list is used, which consists of overlapped genes, nearest downstream genes, and nearest upstream genes.

4.8.Genotyping results and comparison with population frequencies

After running the pyrosequencing experiments, obtained runs were analyzed by using Pyromark Q24 software by Qiagen. The obtained pyrograms were interpreted to investigate variations.

The genotyping results of the LOAD and control groups of the variants prioritized with Ensemble scores and Entropy method and whose genotyping was completed are presented in detail. Variants are encoded with the first five digits of the rsIDs. The minor allele frequencies of these variants are summarized in **Table 4.11**.

When the MAFs of the LOAD-TR and Non-Finnish European (NFE) populations were compared (except for rs11749731), it was seen that all variants examined were distinctive in terms of risk and protective. Four of these variants (rs17092948, rs10050568, rs10807701, rs6705017) indicated low risk, while nine variants were seen twice and higher times in the LOAD class.

Table 4.11. Comparison of LOAD-TR and NFE MAF: Comparison of genotyped variants with LOAD-TR and European (NFE) MAF. LOAD-TR and CNT-TR are abbreviations for project participants.

Variant	Gene	LOAD-TR MAF	CNT-TR MAF	LOAD-TR/ CNT-TR	NFE MAF	LOAD TR/NFE
rs100170	PI4K2B	0.103	0.067	1.54	0.047	2.19
rs100505	SPOCK1	0.395	0.4	0.99	0.751	0.53
rs105154	SPOCK1	0.317	0.188	1.691	0.146	2.17
rs105205	TENM3	0.2	0.086	2.333	0.138	1.447
rs108077	TPST1	0.39	0.467	0.84	0.891	0.44
rs111300	ITPR1	0.3	0.324	0.927	0.203	1.481
rs115036	FAM76B	0.917	0.875	1.05	0.492	0.86
rs11749	NDFIP1	0.563	0.308	1.83	0.621	0.91
rs142155	CNTN5	0.61	0.611	0.998	0.344	1.774
rs14571	CSMD1	0.605	0.438	1.382	0.414	1.461
rs170675	NHSL1	0.225	0.25	0.9	0.046	4.93
rs170816	genomic	0.19	0.143	1.33	0.091	2.09
rs170929	DLC1	0.2	0.333	0.6	0.164	1.219
rs176421	LSAMP	0.35	0.143	2.45	0.054	6.504
rs297801	ST3GAL1	0.69	0.75	0.92	0.36	1.92
rs377640	ARHGAP26	0.128	0.057	2.244	0.015	8.633
rs378079	VAV2	0.488	0.563	0.87	0.342	1.43
rs475030	CCDC3	0.476	0.25	1.905	0.238	1.999
rs557098	ALDH3B1	0.033	0	N/A	0.004	8.33
rs557098	UNC93B1	0.024	0	N/A	0.004	6.011
rs605928	ADAM10	0.439	0.455	0.966	0.301	1.46
rs609841	genomic	0.122	0.125	0.98	0.015	8.18
rs670501	UGGT	0.049	0.063	0.78	0.977	0.05
rs715763	FERMT2	0.158	0.25	0.63	0.117	1.35
rs931460	ANGPT2	0.07	0	N/A	0.011	6.34
rs936666	SYCP2L	0.619	0.813	0.76	0.429	1.44

4.8.1 Statistical Power Analysis of Variant Frequency Differences

Power analysis indicates how many participants can reach the required power and error percentage during the planning phase of the studies. While the studies are in progress, the measured effect and the number of participants are used to calculate the strength achieved with the same formulation. Since the minor allele frequencies will be compared, the MAF values listed for the populations found in the gnomAD database were used as the effect mean value in the strength analysis. Based on the distribution of MAF values in these populations, the standard derivation for each variant MAF value was calculated as seen in **Table 4.12**.

Table 4.12. gnomAD population means and standard deviations of the genotyped variants.

LOAD RF RF variant	**gnomAD populations Avg**	**SD**
rs100170	0.071	0.029
rs100505	0.703	0.08
rs105154	0.13	0.06
rs105205	0.144	0.042
rs108077	0.882	0.049
rs111300	0.139	0.059
rs115036	0.523	0.169
rs11749	0.679	0.05
rs142155	0.408	0.169
rs14571	0.383	0.103
rs170675	0.139	0.133
rs170816	0.132	0.054
rs170929	0.171	0.036
rs176421	0.056	0.025
rs297801	0.234	0.122
rs377640	0.085	0.109
rs378079	0.263	0.106
rs475030	0.256	0.085
rs557098	0.032	0.065
rs557098	0.013	0.009
rs605928	0.528	0.209
rs609841	0.061	0.107
rs670501	0.903	0.101
rs715763	0.096	0.068
rs931460	0.079	0.118
rs936666	0.337	0.066

Based on the distribution and standard derivation information calculated in the table above for the genotyped variants, the MAF observed in the LOAD group was compared with the Non-Finnish European (NFE) population and the healthy participant groups formed within the scope of the project (Rosner B., 1982). Calculated force values and significant differences are given in **Table 4.13**.

Table 4.13. Power Analysis of LOAD-RF-RF variants MAF comparisons. When calculating the statistical power for all groups, a type-1 error was accepted as =0.01. The genotyped variants selected with the LOAD-RF-RF model as a result of the strength analysis were found to have comparable reliability with the European population.

Variant	Gene	LOAD-TR/ CNT-TR	power	LOAD TR/NFE	power
rs100170	PI4K2B	1.54	88.00%	2.19	100.00%
rs100505	SPOCK1	0.99	2.00%	0.53	100.00%
rs105154	SPOCK1	1.691	10..0%	2.17	100.00%
rs105205	TENM3	2.333	100.00%	1.447	100.00%
rs108077	TPST1	0.84	98.00%	0.44	100.00%
rs111300	ITPR1	0.927	29.00%	1.481	100.00%
rs115036	FAM76B	1.05	5.00%	0.86	100.00%
rs11749	NDFIP1	1.83	100.00%	0.91	83.00%
rs142155	CNTN5	0.998	1.00%	1.774	100.00%
rs14571	CSMD1	1.382	100.00%	1.461	100.00%
rs170675	NHSL1	0.9	4.00%	4.93	100.00%
rs170816	genomic	1.33	1.00%	2.09	99.00%
rs170929	DLC1	0.6	100.00%	1.219	99.00%
rs176421	LSAMP	2.45	100.00%	6.504	100.00%
rs297801	ST3GAL1	0.92	17.00%	1.92	100.00%
rs377640	ARHGAP26	2.244	70.00%	8.633	99.00%
rs378079	VAV2	0.87	37.00%	1.43	94.00%
rs475030	CCDC3	1.905	100.00%	1.999	100.00%
rs557098	ALDH3B1	N/A	100.00%	8.33	100.00%
rs557098	UNC93B1	N/A	19.00%	6.011	63.00%
rs605928	ADAM10	0.966	2.00%	1.46	77.00%
rs609841	genomic	0.98	57.00%	8.18	100.00%
rs670501	UGGT	0.78	3.00%	0.05	100.00%
rs715763	FERMT2	0.63	93.00%	1.35	68.00%
rs931460	ANGPT2	N/A	26.00%	6.34	50.00%
rs936666	SYCP2L	0.76	100.00%	1.44	100.00%

4.8.2 LOAD Variant Frequencies with Statistical Significance

GenADA, ADNI, and NCRAD datasets have been modeled with high success with machine learning methods (LOAD-RF-RF). After evaluating these models with Ensemble and Entropy analyses, among the prioritized variants, their genotyping in the LOAD-TR and CNT-TR groups was completed.

Genotyping results were evaluated for success in classification and statistical power in the study group and Non-Finnish European populations created in the project. It was observed that three variants with MAF 0.01 and below differed with high risk when compared with both TR and European populations. Although these variants have low statistical power, they were selected for the diagnostic kit considering their low MAF.

The nine variants evaluated were classified as significantly high risk in the LOAD-TR groups. Among these nine variants, except for rs11749731, the risk classification for LOAD in the European population was also statistically significant. It was observed that six variants listed in rows 13 to 18 in **Table 4.14** were in the protective classification for LOAD in the TR group. Out of these variants, rs10807701 and rs6705017 were also evaluated to be protective in the European population.

When the eight variants listed in the 19th and 26th rows in **Table 4.14** were evaluated individually, it was observed that there was no distinctive feature in the TR group. Out of these eight variants, one was found to have a protective (rs10050568) classification power for the European population, and the others for the risk of LOAD.

As a success criterion of the project, it was expected that 75% of the variants prioritized with machine learning models would have a distinctive feature of LOAD risk for the geography of Turkey. As listed in **Table 4.13**, 18 of the 26 variants studied were found to be distinctive in the LOAD-TR and CNT-TR groups in terms of risk and protection. Distinctive ones constitute 70% of the genotyping variants. Considering the Non-Finnish European, which is the population closest to the population frequencies observed in our country, it was observed that only one variant did not report a difference between the LOAD group and NFE, which shows that the variants selected for the NFE population consist of variants with a discrimination potential of 96%.

Table 4.14. Relationship of genotyped variants to LOAD. The success and statistical power of 26 variants Prioritized by Ensemble and Entropy Methods in classification in LOAD-TR and CNT-TR groups and European (NFE) populations.

variant		gene	LOADT-TR maf	CNT-TR maf	LOAD TR/ CNT TR	power	NFE maf	LOAD TR/NFE	power	explanation
1	rs55709	UNC93B1	0.024	0	*	100.00%	0.004	6.011	100.00%	**variants indicating very high risk in both TR and NFE populations**
2	rs93146	ANGPT2/MCPH1	0.07	0	*	26.00%	0.011	6.342	50.00%	
3	rs55709	ALDH3B1	0.03	0	*	19.00%	0.004	8.333	63.00%	
4	rs17642	LSAMP	0.35	0.143	2.45	100.00%	0.054	6.504	100.00%	**LOAD-RF-RF variants that increase risk in TR Variants that also increase risk for NFE, except rs11749731**
5	rs10520	TENM3	0.2	0.086	2.333	100.00%	0.138	1.447	100.00%	
6	rs37764	ARHGAP26	0.256	0.114	2.244	70.00%	0.015	8.633	99.00%	
7	rs47503	CCDC3	0.476	0.25	1.905	100.00%	0.238	1.999	100.00%	
8	rs11749	NDFIP1	0.56	0.31	1.828	100.00%	0.621	0.906	83.00%	
9	rs10515	SPOCK1	0.317	0.188	1.691	100.00%	0.146	2.17	100.00%	
10	rs100117	PI4K2B	0.1	0.07	1.538	88.00%	0.047	2.187	100.00%	
11	rs14571	CSMD1	0.605	0.438	1.382	100.00%	0.414	1.461	100.00%	
12	rs17081	genomic	0.19	0.14	1.333	1.00%	0.091	2.089	99.00%	
13	rs37807	VAV2	0.49	0.56	0.868	37.00%	0.342	1.428	94.00%	**Protective LOAD-RF-RF variants in TR also protective variants for rs10807701 and rs6705017 NFE**
14	rs10807	TPST1	0.39	0.47	0.836	98.00%	0.891	0.428	100.00%	
15	rs67050	UGGT	0.05	0.06	0.78	3.00%	0.977	0.05	100.00%	
16	rs93666	SYCP2L	0.62	0.81	0.762	100.00%	0.429	1.444	100.00%	
17	rs71576	FERMT2	0.16	0.25	0.632	93.00%	0.117	1.351	68.00%	
18	rs17092	DLC1	0.2	0.333	0.6	100.00%	0.164	1.219	99.00%	
19	rs10050	SPOCK1	0.4	0.4	0.988	2.00%	0.751	0.526	100.00%	**Variants with no distinctive features for TR 7 were classified as risk for NFE and 1 as protective for the NFE population**
20	rs60984	genomic	0.12	0.13	0.976	57.00%	0.015	8.185	100.00%	
21	rs17067	NHSL1	0.23	0.25	0.9	4.00%	0.046	4.934	100.00%	
22	rs29780	ST3GAL1	0.69	0.75	0.921	17.00%	0.36	1.92	100.00%	
23	rs11503	FAM76B	0.92	0.88	1.048	5.00%	0.492	1.863	100.00%	
24	rs14215	CNTN5	0.61	0.611	0.998	1.00%	0.344	1.774	100.00%	
25	rs11130	ITPR1	0.3	0.324	0.927	29.00%	0.203	1.481	100.00%	
26	rs60592	ADAM10	0.439	0.455	0.966	2.00%	0.301	1.46	77.00%	

CHAPTER 5

5. DISCUSSION

Diagnosis of Late-Onset Alzheimer's (LOAD) disease is now partially available based on clinical evaluation and imaging, and many patients remain undiagnosed in the early stages of LOAD. Recent advances in genome-wide analysis are accelerating the study of human health and disease. High-dimensional data is produced that can be used for research purposes, including genotyping data of different platforms. High-throughput genotyping technologies such as Affymetrix or Illumina enable association studies. The Discovery of SNV biomarkers at different loci may improve early diagnosis accuracy and early/differential diagnosis processes of these diseases through clinical decision making. SNVs, which can act as biomarkers by causing individual differences and localizing on genes associated with a particular complex disease, will contribute to determining the pathogenetic causation and origin behind the phenotypes of complex genetically based diseases such as LOAD.

Genome-Wide Association Studies (GWAS) allows for the investigation of statistical interactions of individual variants at candidate loci, but univariate analysis checks for interactions between variants. In the current literature, a meta-analysis of LOAD GWAS results is not successful in identifying causal variants. Integrated prioritization of all SNVs associated with post-GWAS or machine learning approaches from the genotyping results produced by different platforms is a fundamental problem in the field. Since SNP genotyping platforms do not have a shared set, finding common biomarkers between individual risk models was challenging. Our group proposed two approaches to address this prioritization problem: Ensemble (Onur E, p.c.,2002) and Entropy (Burcu Yaldız, p.c. 2022). Here we have compared the utility of variants prioritized by the ensemble and entropy-based methods.

Among the variants with an Ensemble score of 2.31 and higher, variants associated with bioinformatics analysis LOAD were selected for network analysis. The variants overlapping with a protein-coding gene and LOAD-related biological pathways were selected for experimental validation based on biological networks, protein-protein interactions, and functional enrichment. Variant triplets are also selected from three databases (ADNI, GenADA, NCRAD) using the information gain approach for the Entropy method in the same manner.

The network analyses reveal the functional relevance of RF-RF model variants associated with LOAD risk as potential causative variants. Here the results of each analysis and discusses the advantages and disadvantages of ensemble and entropy-based approaches are discussed.

The overall statistics of GeneMANIA analysis are summarized as in **Table 4.1**. Ensembl genes with variant score higher than 2.31 showed highest physical interactions (**19.19%**), genetic interactions (**6.00%**), and co-localization (**26.52%**). Ensembl >2.31 list has the highest number of variants compared to the other lists. So these results might be observed due to its large variant size. However, Ensembl >3.21 set showed a higher ratio for co-expression (**59.19%**) and shared protein domains (**10.96%**). When Ensembl and Entropy sets are compared, we observe that the Entropy set outperformed both Ensembl sets regarding co-expression (**85%**).

When we look at the results of the STRING enrichment, we observe that the raw list of genes for Ensembl >2.31 is a more significantly connected network than Ensembl >3.21. For Ensembl > 2.31 we have not added any additional genes (raw list). We have added 10 more genes to the other groups to obtain a significant PPI enrichment p-value.

In **Table 4.5,** we can see that network diameter is 14 for the Ensembl gene list and 5 for the Entropy gene list. Network diameter is the longest shortest path between any two nodes in the network. The characteristic path length is defined as the average shortest path length between all pairs of nodes in the network. When we compare characteristic path length for Ensembl 2.31 and higher, it is 5.057, and for Entropy, it is 2.282. Network density is 0.028 for Ensembl and 0.211 for Entropy; network efficiency is 0.20 for Ensembl gene set and 0.43 for Entropy gene set. These observations indicate that the Entropy gene set has a more connected network.

When we compared network efficiency and network density at the point where all networks had significantly more interactions than expected. Based on these observations, Ensembl >3.21 genes observed to be more comparable to all genes selected with Entropy.

The results of this comparison are listed below;

Rank by Network_density:

1- Entropy-ADNI
2- Entropy-NCRAD
3- Entropy-GenADA
4- **Entropy-all**
5- **Ensembl > 3.21**

6- Ensembl >2.31

Rank by Network_efficiency:

1- Entropy-ADNI
2- Entropy
3- Ensembl >3.21
4- Entropy-NCRAD
5- Entropy-GenADA
6- Ensembl >2.31

The entropy gene set has a higher network density and efficiency than the Ensembl gene set, which indicates that the Entropy gene set is more connected than the Ensembl gene set regarding gene networks. On the other hand, the PPI enrichment p-value is 1.00E-16 for the Ensembl network and 3.99E-9 for the Entropy network, which shows that the Ensembl gene set produced a more connected and meaningful protein-protein interaction network. Hence in the Ensembl scoring method, the variants clustered on neighboring genomic loci are prioritized, probably enriching genes sharing a role in similar processes, supported by highly connected protein-level interactions.

We constructed network genes candidate disease networks to obtain information for our genes from the disease database DisGeNET. When we look at the network obtained for Ensembl >2.31, the diseases obtained in the network are Schizophrenia and Diabetes Mellitus, which did not cover any LOAD-related events. On the other hand, for the Entropy gene set, diseases associated with our genes of interest are; Alzheimer's Disease, Down Syndrome, Prostate Cancer, Acute Myelogenous Leukemia, Obesity, Glaucoma, Lung Cancer, Alcohol Dependence, Autism, Colorectal Cancer, Hepatocellular Carcinoma, and Bladder Cancer.

However, when we obtained genes- Alzheimer's Disease networks from DisGeNET to gain information about our genes and Alzheimer's Disease from the literature, the LOAD-related events were more observable for the ensemble selected gene list. Looking at the networks obtained for Ensembl >2.31 and Entropy, we see many more interactions between Ensembl >2.31 and Alzheimer's Disease clusters than the Entropy gene set and Alzheimer's Disease. More genes are associated with AD included within the genes prioritized by the Ensembl scoring method.

Knowledge graphs created by CROssBAR are used to obtain associations of genes, pathways, and diseases. We have used obtained knowledge graphs to investigate relations of our gene sets with Alzheimer's Disease. We have used extended gene lists for Ensembl

and Entropy to gain as much information as possible from genes, and we observed pathway, gene, and Alzheimer's Disease relationships of our gene sets. When we look at the knowledge graphs obtained for Ensembl >2.31, the genes connected to Alzheimer's Disease node were; ADAM10, PSEN1, PSEN2, UNC5C, MT-ND1, MT-ND2, DBN1, SORL1, ABCA7, APP, and APOE. For Entropy knowledge graph; APP, UNC5C, PSEN1, ADAM10, PSEN2, MT-ND2, MT-ND1, ABCA7, SORL1, APOE, and DBN1.

When we look at the Figure 4.31; knowledge graph obtained for Ensemble >2.31 and higher scored variants, genes that are related to Alzheimer's Disease node are APP, DBN1, MT-ND2, MT-ND1, ABCA7, SORL1, ADAM10, UNC5C, PSEN2, PSEN1 and APOE. We know from the literature that ADAM10 and APOE are the genes that are related to Alzheimer's Disease. At the same time ADAM10 is one of the genes that is present in our Ensembl gene list. Pathways that are found to be associated with these genes are NOTCH activation and transmission, Noncanonical activation of NOTCH3 and Nuclear signaling by ERBB4. When we look at the Figure 4.32; knowledge graph obtained for Ensemble >3.21 and higher scored variants, genes related to Alzheimer's Disease node are PSEN1, PSEN2, ADAM10, UNC5C, APP, MT-ND1, MT-ND2, ABCA7, SORL1, DBN1 and APOE. Although the number of genes is fewer in the Ensembl 3.21> list we still observe the same pathways; noncanonical activation of NOTCH3, NOTCH2 activation and transmission, NOTCH3 activation and transmission, NOTCH4 activation and transmission and Nuclear signaling by ERBB4. When we look at the connection of NOTCH 2-3-4 activation and transmission pathways with the genes, it is observed that ADAM10 is involved in the NOTCH pathways. ADAM10 is associated with Alzheimer's Disease pathology from the literature. According to Yuen I. et al., 2017, increasing amounts of data shows that ADAM10 not only inhibits the formation of Aβ, but it may also alter the pathology of Alzheimer's disease through mechanisms such as lowering tau pathology, maintaining normal synaptic functioning, and encouraging hippocampus neurogenesis and neural network homeostasis.

On the other hand, Nuclear signaling by ERBB4 has been shown to take part in AD pathogenesis, and it is known to regulate SPOCK1 and APOE at gene expression level in the nucleus. APOE is one of the major players in AD etiology, responsible for %25 of hereditary AD cases. APOE, which is found in both the CNS and the periphery, is a crucial link between these two compartments and playing a role in the pathogenesis of Alzheimer's disease (AD) by breaking the blood–brain barrier (BBB) on both sides (Chernick et al., 2019). But there are no AD-related studies for the SPOCK1 gene, which is a gene from the SPARC protein family. It is known that SPOCK1 is regulated by ERBB4 signaling. There are studies on the role of ERRB4 regulation in the pathology of

AD, but so far, there aren't any studies showing that SPOCK1 is related to AD through ERBB4 or causes a pathology through nerve cells.

The SPARC gene is shared between Ensemble and Entropy gene lists and genes related to "Nuclear signaling by ERBB4 pathway". There is a lack of studies on the role of SPOCK1 (SPARC) in the brain and LOAD. These observations lead us to propose a new hypothesis that ***SPOCK1 (SPARC) may be regulated by ERBB4 signaling in brain cells, thus having a causative role in AD pathogenesis.***

The functional enrichment analysis results revealed a higher number of clusters for Ensemble >2.31 set than the Entropy set when we look at the results. The clusters obtained for the Entropy method are at the same time part of the Ensemble>2.31 network. Enrichment and clustering analyses showed that neuron and cell development, extracellular matrix, and cell membrane functions occur among genes with ensemble scores of 2.31 and higher.

As the result of the genotyping, risk variants that are significantly different in case and control groups are; UNC93B1, ANGPT2/MCPH1, ALDH3B1, LSAMP, TENM3, ARHGAP26, CCDC3, NDFIP1, SPOCK1, PI4K2B, CSDMD and a genomic variant. When we look at the SPOCK1 gene, it is observed with a much greater frequency in the LOAD group. As a success criterion of the proposed TUBİTAK ARDEB 1003 project, it was expected that 75% of the variants prioritized with machine learning models would have a distinctive feature of LOAD risk for the geography of Turkey. As listed in **Table 4.14**, 18 of the 26 variants studied were found to be distinctive in the LOAD-TR and CNT-TR groups in terms of risk and protection. Distinctive ones constitute 70% of the genotyping variants. Considering the Non-Finnish European, which is the population closest to the population frequencies observed in our country, it was observed that only one variant did not report a difference between the LOAD group and NFE, which shows that the variants selected for the NFE population consist of variants with a discrimination potential of 96%.

CHAPTER 6

6. CONCLUSIONS

The Late-Onset Alzheimer's Disease's (LOAD) complex genetic etiology is still unclear, which prevents the early and differential diagnosis of LOAD. Genome-Wide Association Studies (GWAS) allows exploration of the statistical interactions of individual variants, but the univariate analysis oversees these interactions between variants. The machine learning algorithms can capture hidden, novel, and significant patterns considering nonlinear interactions between variants to understand the genetic predisposition for complex genetic disorders, where multiple variants determine the risk.

In this study, the variants produced as the results of the machine learning algorithms developed by our group are compared by network-based approaches. Firstly, we constructed biological networks for our genes of interest to compare the connectivity of networks obtained by different methods. Then we did functional enrichment analysis to discover pathways related to Alzheimer's Disease. Lastly, we observed the number of shared pathway clusters obtained by Ensemble and Entropy gene sets. Among the selected variants after prioritization associated with LOAD, the pyrosequencing primer design was completed for a total of 32 variants, and the sequencing primers were optimized. In addition, within the project's scope, a case-control group consisting of 43 LOAD diagnosed and 38 healthy participants were formed, and genotyping for the prioritized variants was completed.

The result obtained by the Ensemble method included interacting proteins, while the result obtained by entropy showed that the genes on the same genetic network were more concentrated. We showed that, both of the meta-analysis methods helped us to understand the relation of associated variants prioritized with LOAD-RF-RF and LOAD. Genotyping studies have shown that the proposed LOAD-RF-RF model variants will be successful in the early and differential diagnosis of patients, especially in NFE populations. These results should be reconfirmed with clinical studies to be carried out with expanded patient groups, and confirmation should be made about the use for diagnostic purposes in the clinic.

In addition, we think that the rate of choosing causative variants among the associated variants has increased with both the application of machine learning methods applied after GWAS and the meta-analysis of the results obtained from different data sets with ensmebl

and entropy methods. In this study, the strongest candidate variants with these features were on the SPOCK1 gene, which is regulated by ERBB4 signal.

ERBB4 gene is one of the four members in the EGFR subfamily of receptor tyrosine kinases. Ligands include EGF, epiregulin, betacellulin and the neuregulins. (Sundvall et.al.) ERBB4 is a key neuregulin-1 receptor, it is expressed in multiple regions in the adult animal brain, but there is a little information about its localization in AD brains. In a 2010 study by Woo et al., in AD brains, ERBB4 immunoreactivity was demonstrated to colocalize with the apoptotic signal Bax in apoptotic hippocampal pyramidal neurons. These results suggest that up-regulation of ERBB4 immunoreactivity in the apoptotic neuron may involve the progression of AD pathology. The neuregulins (NRG1) are a complex family of factors that perform many functions during neural development. At interneuronal synapses, neuregulin ERBB receptors associate with PDZ-domain proteins at postsynaptic densities, where they can modulate synaptic plasticity. (Buonanno and Fischbach, 2001). According to the study Sardi et al. 2006, presenilin-dependent ERBB4 Nuclear signaling regulates the timing of astrogenesis in the developing brain. Relying on these studies, astrogenesis occurs precociously in ERBB4 knockout mice. Corresponding to the 2011 study by Woo et al., up-regulating of ERBB4 immunoreactivity may be involved in the progression of pathology of AD. Neuregulin 4 (NRG4) +/+ mice embryonic cortical pyramidal neurons co-express NRG4 and its receptor ERBB4, according to a study by Paramo et al., 2018.

When we look at the recent 2021 study by Ou et al., it is stated that NRG4 has a crucial function in the developing brain. Thus, it will be interesting to ascertain how the putative NRG4/ERBB4 autocrine loop is regulated in pyramidal neurons and investigate how NRG4 contributes to the pathogenesis of particular neurodegenerative diseases such as AD. On the other hand, in a 2016 study by Chang et al., it was stated that increased plasma soluble neuregulin-1 levels might be a novel and reliable biological marker for the early diagnosis of AD. Furthermore, Xu et al., 2016 stated that soluble ectodomains of type I and type III Neuregulin 1 significantly increased the expression of Aβ-degrading enzyme neprilysin (NEP) in primary neuronal cultures. A 2020 review by Guma et al. investigated the impact of the growth factor neuregulin and its ERBB receptors on mitochondria. While the APOE4 gene, which is associated with AD etiology, also has a detrimental role in regulating fatty acid metabolism across neurons and astrocytes in tandem with their distinctive mitochondrial phenotypes. The literature supports the role of ERBB4 signaling in AD etiology. Moreover, the gene SPOCK1 (SPARC) present in our gene lists is related to the ERBB4 nuclear signaling pathway. Upon this, we proposed that the role of SPOCK1 in brain cells and Alzheimer's Disease should be further investigated.

REFERENCES

Amberger, J., Bocchini, C., & Hamosh, A. (2011). A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Human mutation*, *32*(5), 564-567.

Assenov, Y., Ramírez, F., Schelhorn, S. E., Lengauer, T., & Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, *24*(2), 282-284.

Barreto, G., E. White, R., Ouyang, Y., Xu, L., & G. Giffard, . (2011). Astrocytes: targets for neuroprotection in stroke. *Central Nervous System Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Central Nervous System Agents)*, *11*(2), 164-173

Bodily, P. M., Fujimoto, M. S., Page, J. T., Clement, M. J., Ebbert, M. T., & Ridge, P. G. (2016). A novel approach for multi-SNP GWAS and its application in Alzheimer's disease. *BMC bioinformatics*, *17*(7), 455-463.

Buonanno, A., & Fischbach, G. D. (2001). Neuregulin and ErbB receptor signaling pathways in the nervous system. *Current opinion in neurobiology*, *11*(3), 287–296. https://doi.org/10.1016/s0959-4388(00)00210-5

Cabezas, R., El-Bachá, R. S., González, J., & Barreto, G. E. (2012). Mitochondrial functions in astrocytes: neuroprotective implications from oxidative damage by rotenone. *Neuroscience research*, *74*(2), 80-90.

Chang, K. A., Shin, K. Y., Nam, E., Lee, Y. B., Moon, C., Suh, Y. H., & Lee, S. H. (2016). Plasma soluble neuregulin-1 as a diagnostic biomarker for Alzheimer's disease. *Neurochemistry International*, *97*, 1-7.

Chelala, C., Khan, A., & Lemoine, N. R. (2009). SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, *25*(5), 655-661.

Chernick, D., Ortiz-Valle, S., Jeong, A., Qu, W., & Li, L. (2019). Peripheral versus central nervous system APOE in Alzheimer's disease: Interplay across the blood-brain barrier. *Neuroscience letters*, *708*, 134306.

Chouraki, V., & Seshadri, S. (2014). Genetics of Alzheimer's disease. *Advances in genetics*, *87*, 245-294.

Coulter, D. A., & Steinhäuser, C. (2015). Role of astrocytes in epilepsy. *Cold Spring Harbor perspectives in medicine*, *5*(3), a022434.Cover, T. M., & Thomas, J. A. (1991). Entropy, relative entropy and mutual information. *Elements of information theory*, *2*(1), 12-13.

Cuyvers, E., & Sleegers, K. (2016). Genetic variations underlying Alzheimer's disease: evidence from genome-wide association studies and beyond. *The Lancet Neurology*, *15*(8), 857-868.

Dayem Ullah, A. Z., Lemoine, N. R., & Chelala, C. (2012). SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic acids research*, *40*(W1), W65-W70.

Dayem Ullah, A. Z., Lemoine, N. R., & Chelala, C. (2013). A practical guide for the functional annotation of genetic variations using SNPnexus. *Briefings in bioinformatics*, *14*(4), 437-447.

Dayem Ullah, A. Z., Oscanoa, J., Wang, J., Nagano, A., Lemoine, N. R., & Chelala, C. (2018). SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic acids research*, *46*(W1), W109-W113.

de Rojas, I., Moreno-Grau, S., Tesi, N., Grenier-Boley, B., Andrade, V., Jansen, I. E., ... & Sleegers, K. (2021). Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. *Nature communications*, *12*(1), 1-16.

Desikan, R. S., Fan, C. C., Wang, Y., Schork, A. J., Cabral, H. J., Cupples, L. A., ... & Dale, A. M. (2017). Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS medicine*, *14*(3), e1002258.

Doğan, T., Atas, H., Joshi, V., Atakan, A., Rifaioglu, A.S., Nalbat, E., Nightingale, A., Saidi, R., Volynkin, V., Zellner, H. Cetin-Atalay, R., Martin, M. J. & Atalay, V. (2021). CROssBAR: Comprehensive Resource of Biomedical Relations with Knowledge Graph Representations. Nucleic Acids Research, 49(16), e96-e96.

Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, *48*(1-4), 2.

Emahazion, T., Feuk, L., Jobs, M., Sawyer, S. L., Fredman, D., St Clair, D., ... & Brookes, A. J. (2001). SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. *TRENDS in Genetics*, *17*(7), 407-413.

Escott-Price, V., Bellenguez, C., Wang, L. S., Choi, S. H., Harold, D., Jones, L., ... & Montine, T. J. (2014). Gene-wide analysis detects two new susceptibility genes for Alzheimer's disease. *PloS one*, *9*(6), e94661.

Filippini, N., Rao, A., Wetten, S., Gibson, R. A., Borrie, M., Guzman, D., ... & Matthews, P. M. (2009). Anatomically-distinct genetic associations of APOE ε4 allele load with regional cortical atrophy in Alzheimer's disease. *Neuroimage*, *44*(3), 724-728.

Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., ... & Pedersen, N. L. (2006). Role of genes and environments for explaining Alzheimer disease. *Archives of general psychiatry*, *63*(2), 168-174.

Gumà, A., Díaz-Sáez, F., Camps, M., & Zorzano, A. (2020). Neuregulin, an Effector on Mitochondria Metabolism That Preserves Insulin Sensitivity. *Frontiers in physiology*, *11*, 696. https://doi.org/10.3389/fphys.2020.00696

Hamby, M. E., & Sofroniew, M. V. (2010). Reactive astrocytes as therapeutic targets for CNS d disorders. *Neurotherapeutics*, *7*(4), 494-506.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., ... & Williams, J. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature genetics*, *41*(10), 1088-1093.

Isserlin, R., Merico, D., Voisin, V., & Bader, G. D. (2014). Enrichment Map–a Cytoscape app to visualize and explore OMICs pathway enrichment results. F1000Research, 3.

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., ... & Posthuma, D. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature genetics*, *51*(3), 404-413.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., & von Mering, C. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, *37*(Database issue), D412–D416. https://doi.org/10.1093/nar/gkn760

Junker, B. H., & Schreiber, F. (2011). *Analysis of biological networks* (Vol. 2). John Wiley & Sons.

Kamboh, M. I., Demirci, F. Y., Wang, X., Minster, R. L., Carrasquillo, M. M., Pankratz, V. S., ... & Barmada, M. M. (2012). Genome-wide association study of Alzheimer's disease. *Translational psychiatry*, *2*(5), e117-e117.

Kamboh, M. I. (2018). A brief synopsis on the genetics of Alzheimer's disease. *Current genetic medicine reports*, *6*(4), 133-135.

Karagiannis, G. S., Berk, A., Dimitromanolakis, A., & Diamandis, E. P. (2013). Enrichment map profiling of the cancer invasion front suggests regulation of colorectal cancer progression by the bone morphogenetic protein antagonist, gremlin-1. *Molecular oncology*, *7*(4), 826-839.

Kent, W. J. (2002). BLAT---The BLAST-Like Alignment Tool. *Genome research, 12* (4), s. 656–664. doi:10.1101/gr.229202

Kimelberg, H. K., & Nedergaard, M. (2010). Functions of astrocytes and their potential as therapeutic targets. *Neurotherapeutics*, *7*(4), 338-353.

Kucera, M., Isserlin, R., Arkhangorodsky, A., & Bader, G. D. (2016). AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. *F1000Research*, *5*, 1717. https://doi.org/10.12688/f1000research.9090.1

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., ... & Nalls, M. A. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics*, *45*(12), 1452-1458.

Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature*, *470*(7333), 187-197.

Latora, V., & Marchiori, M. (2001). Efficient behavior of small-world networks. *Physical review letters*, *87*(19), 198701.

Le, D. H., & Pham, V. H. (2017). HGPEC: a Cytoscape app for prediction of novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network. *BMC systems biology*, *11*(1), 61. https://doi.org/10.1186/s12918-017-0437-x

Li, H., Wetten, S., Li, L., Jean, P. L. S., Upmanyu, R., Surh, L., ... & Roses, A. D. (2008). Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Archives of neurology*, *65*(1), 45-53.

Maragakis, N. J., & Rothstein, J. D. (2006). Mechanisms of disease: astrocytes in neurodegenerative disease. *Nature clinical practice Neurology*, *2*(12), 679-689.

Meng, X., Li, J., Zhang, Q., Chen, F., Bian, C., Yao, X., ... & Shen, L. (2020). Multivariate genome wide association and network analysis of subcortical imaging phenotypes in Alzheimer's disease. *BMC genomics*, *21*(11), 1-12.

Merico, D., Isserlin, R., & Bader, G. D. (2011). Visualizing gene-set enrichment results using the Cytoscape plugin enrichment map. In Network biology (pp. 257-277). Humana Press.

Moradifard, S., Hoseinbeyki, M., Ganji, S. M., & Minuchehr, Z. (2018). Analysis of microRNA and gene expression profiles in Alzheimer's disease: a meta-analysis approach. *Scientific reports*, *8*(1), 1-17.

Mukherjee, S., Seal, M., & Dey, S. G. (2014). Kinetics of serotonin oxidation by heme–Aβ relevant to Alzheimer's disease. *JBIC Journal of Biological Inorganic Chemistry*, *19*(8), 1355-1365.

Musani, S. K., Shriner, D., Liu, N., Feng, R., Coffey, C. S., Yi, N., ... & Allison, D. B. (2007). Detection of gene× gene interactions in genome-wide association studies of human population data. *Human heredity*, *63*(2), 67-84.

Parpura, V., Grubišić, V., & Verkhratsky, A. (2011). Ca2+ sources for the exocytotic release of glutamate from astrocytes. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, *1813*(5), 984-991.

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., ... & Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData mining*, *4*(1), 1-27.

Perea, G., Navarrete, M., & Araque, A. (2009). Tripartite synapses: astrocytes process and control synaptic information. *Trends in neurosciences*, *32*(8), 421-431.

Perez-Alvarez, A., & Araque, A. (2013). Astrocyte-neuron interaction at tripartite synapses. *Current drug targets*, *14*(11), 1220-1224.

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, *48*(D1), D845-D855.

Posada-Duque, R. A., Barreto, G. E., & Cardona-Gomez, G. P. (2014). Protection after stroke: cellular effectors of neurovascular unit integrity. *Frontiers in cellular neuroscience*, *8*, 231.

Qi, G., Mi, Y., Shi, X., Gu, H., Brinton, R. D., & Yin, F. (2021). ApoE4 Impairs Neuron-Astrocyte Coupling of Fatty Acid Metabolism. *Cell reports*, *34*(1), 108572. https://doi.org/10.1016/j.celrep.2020.108572

Prokopenko, D., Morgan, S. L., Mullin, K., Hofmann, O., Chapman, B., Kirchner, R., ... & Tanzi, R. E. (2021). Whole-genome sequencing reveals new Alzheimer's disease–associated rare variants in loci related to synaptic function and neuronal development. *Alzheimer's & Dementia*.

Rahman, M. R., Islam, T., Turanli, B., Zaman, T., Faruquee, H. M., Rahman, M. M., ... & Moni, M. A. (2019). Network-based approach to identify molecular signatures and therapeutic agents in Alzheimer's disease. *Computational biology and chemistry*, *78*, 431-439.

Raman, K. (2010). Construction and analysis of protein–protein interaction networks. *Automated experimentation*, *2*(1), 1-11.

Rao, V. S., Srinivas, K., Kumar, G. S., & Sujin, G. N. (2013). Protein interaction network for Alzheimer's disease using computational approach. *Bioinformation*, *9*(19), 968.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, *47*(W1), W191-W198.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., & Bader, G. D. (2019). Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature protocols*, *14*(2), 482-517.

Ruiz, A., Heilmann, S., Becker, T., Hernández, I., Wagner, H., Thelen, M., ... & Ramírez, A. (2014). Follow-up of loci from the International Genomics of Alzheimer's Disease Project identifies TRIP4 as a novel susceptibility gene. *Translational psychiatry*, *4*(2), e358-e358.

Oscanoa, J., Sivapalan, L., Gadaleta, E., Dayem Ullah, A. Z., Lemoine, N. R., & Chelala, C. (2020). SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic acids research*, *48*(W1), W185-W192.

Ou, G. Y., Lin, W. W., & Zhao, W. J. (2021). Neuregulins in Neurodegenerative Diseases. *Frontiers in aging neuroscience*, *13*, 170.

Sardi, S. P., Murtie, J., Koirala, S., Patten, B. A., & Corfas, G. (2006). Presenilin-dependent ErbB4 nuclear signaling regulates the timing of astrogenesis in the developing brain. *Cell*, *127*(1), 185-197.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504.

Sherva, R., Tripodis, Y., Bennett, D. A., Chibnik, L. B., Crane, P. K., de Jager, P. L., ... & GENAROAD Consortium, The Alzheimer's Disease Neuroimaging Initiative, and The Alzheimer's Disease Genetics Consortium. (2014). Genome-wide association study of the rate of cognitive decline in Alzheimer's disease. *Alzheimer's & Dementia*, *10*(1), 45-52.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering,

C. V. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, *47*(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

Sundvall, M., Iljin, K., Kilpinen, S., Sara, H., Kallioniemi, O. P., & Elenius, K. (2008). Role of ErbB4 in breast cancer. *Journal of mammary gland biology and neoplasia*, *13*(2), 259-268.

Talwar, P., Silla, Y., Grover, S., Gupta, M., Agarwal, R., Kushwaha, S., & Kukreti, R. (2014). Genomic convergence and network analysis approach to identify candidate genes in Alzheimer's disease. *BMC genomics*, *15*(1), 1-16.

Tan, E. Y., Köhler, S., Hamel, R. E., Muñoz-Sánchez, J. L., Verhey, F. R., & Ramakers, I. H. (2019). Depressive symptoms in mild cognitive impairment and the risk of dementia: a systematic review and comparative meta-analysis of clinical and community-based studies. *Journal of Alzheimer's Disease*, *67*(4), 1319-1329.

Tosto, G., & Reitz, C. (2013). Genome-wide association studies in Alzheimer's disease: a review. *Current neurology and neuroscience reports*, *13*(10), 381.

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N. & Baudot, A. (2019). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, *35*(3), 497-505.

Van Wijk, B. C., Stam, C. J., & Daffertshofer, A. (2010). Comparing brain networks of different size and connectivity density using graph theory. *PloS one*, *5*(10), e13701.

Volterra, A., & Meldolesi, J. (2005). Astrocytes, from brain glue to communication elements: the revolution continues. *Nature Reviews Neuroscience*, *6*(8), 626-640.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., & Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, *38*(Web Server issue), W214–W220. https://doi.org/10.1093/nar/gkq537

Woo, R. S., Lee, J. H., Yu, H. N., Song, D. Y., & Baik, T. K. (2010). Expression of ErbB4 in the apoptotic neurons of Alzheimer's disease brain. *Anatomy & Cell Biology*, *43*(4), 332-339.

Woo, R. S., Lee, J. H., Yu, H. N., Song, D. Y., & Baik, T. K. (2011). Expression of ErbB4 in the neurons of Alzheimer's disease brain and APP/PS1 mice, a model of Alzheimer's disease. *Anatomy & cell biology*, *44*(2), 116-127.

Xu, J., de Winter, F., Farrokhi, C., Rockenstein, E., Mante, M., Adame, A., ... & Lee, K. F. (2016). Neuregulin 1 improves cognitive deficits and neuropathology in an Alzheimer's disease model. *Scientific reports*, *6*(1), 1-9.

Yuan, X. Z., Sun, S., Tan, C. C., Yu, J. T., & Tan, L. (2017). The Role of ADAM10 in Alzheimer's Disease. *Journal of Alzheimer's disease : JAD*, *58*(2), 303–322. https://doi.org/10.3233/JAD-170061

Zhang, F., Kang, Z., Li, W., Xiao, Z., & Zhou, X. (2012). Roles of brain-derived neurotrophic factor/tropomyosin-related kinase B (BDNF/TrkB) signalling in Alzheimer's disease. *Journal of Clinical Neuroscience*, *19*(7), 946-949.

APPENDICES

APPENDIX A

Materials and Methods

Table 3.5. Detailed parameters of CROssBAR query.

Ensemble	Entropy	ADNI	GenADA	NCRAD	Disease	#nodes	Interacting proteins	Only reviewed genes/proteins	ChEMBL compounds	Predicted compounds	Included tax IDs
AADACL2-AS1,AC000058.1,AC002460.1	AADACL2-AS1,AC009117.1,AC009509.1	AC026396.1,AC026884.1,AC091895.1	AP003398.2,AC079362.1,AC024909.1	HIF1AN,RF00017,AC087379.1	AD	10	Yes	Yes	No	No	9606
AC007907.1,AC008676.3,AC008696.2	,AC009869.1,AC010255.3,AC022784.1	AC092100.1,AC093840.1,AC104662.2	AL161716.1,RF00017,ADAM10	AC009869.1,SOX6,AC009509.1							
AC021683.2,AC021683.3,AC024909.1	AC024909.1,AC025252.1,AC026396.1,	AC106744.1,AF127936.2,AL024498.2	AC090515.3,RN7SKP190,AC009117.1	NEDD1,RF00019,AC073592.8							
AC025254.1,AC025281.1,AC026396.1	AC026884.1,AC034268.2,AC068787.1	AL138885.1,AL139317.5,AL353595.1	AC034268.2,NPLOC4,RN7SKP93	AC073592.2,AC068787.1,AC025252.1							
AC034195.1,AC068050.1,AC078845.1	AC073592.2,AC073592.8,AC079362.1,	AL713851.1,ALDH3B1,APOOP2	MGAT5,RPL12P4,LINC01440	LINC00430,UBBP5,AL137230.2,							
AC079760.1,AC090049.2,AC090515.3	,AC079380.1,AC087379.1,AC090515.3	CDH13,FAM76B,FERMT2	RUNX1,PHF21B,FBLN2	RYR3,LCMT1,GNAL							
AC091895.1,AC093523.1,AC093648.1	AC091895.1,AC092100.1,AC092364.1,	HSPD1P15,LINC00507,LINC01376	SPOCK1,NHSL1,MCPH1	MBD2,AC093462.1,AC092364.1,							
AC106744.1,AC107214.2,AC108025.1	AC093277.1,AC093462.1,AC093840.1	LINC02199,MMP28,MRPS35P3	ANGPT2,ST3GAL1,HMCN2	PSG5,PSG4,AC105450.1							

AC122136.1,AC138623.1,ADAM10	AC096992.1,AC104662.2,AC105252.1,	NDFIP1,NPSR1,NPSR1-AS1		TPO,DPP10,RAPGEF4,							
ADCY2,AL024498.2,AL049646.1	AC105450.1,AC106744.1,AC106800.1	PLAGL1,PSMC1P6,RNU6ATAC21P		DNER,TRIP12,PCP4							
AL096828.2,AL137783.1,AL138749.1	AC117382.2,AC123767.1,ADAM10	SAMSN1,SEPSECS-AS1,SYCP2L		FAM230F,DGCR6,EEFSEC							
AL161449.2,AL353595.1,AL391416.1	ADAM18,AF127936.2,AL024498.2,	TAF15,TEMN3-AS1,TMPRSS15		DNAJB8,AC117382.2,AC096992.1							
AL590824.1,AL713851.1,ALDH3B1	AL137230.2,AL138885.1,AL139317.5	TPST1,UGGT1,VAV2		AADACL2-AS1,SLC4A4,AC105252.1							
ALDH3B2,ANKRD33B,ANXA13	AL161716.1,AL353595.1,AL589740.1			AC079380.1,RNU1-150P,AC093277.1							
AP003385.1,ARHGAP26,ASS1	AL589826.1,AL713851.1,ALDH3B1			NIM1K,AC106800.1,AC010255.3							
BARD1,BTBD16,C9orf3	ANGPT2,ANXA13,AP003398.2			LYRM7,FARS2,FAM83B,							
CCDC3,CNTN5,CRISP1	APOOP2,BX119904.1,C9orf84			HCRTR2,AL589740.1,AL589826.1							
CRYL1,CSGALNACT2,CSMD2	CDH13,DGCR6,DNAJB8,			AL589740.1,AL589826.1,LAMA2							
CST2P1,CST4,CYB5R3	DNER,DPP10,EEFSEC			AC022784.1,AC123767.1,ADAM18							
CYFIP2,DHFRP3,DLC1	FAM230F,FAM76B,FAM83B			ANXA13, MELK, MIR4475							
ENPP6,ENSAP3,FAM170B-AS1	FARS2,FBLN2,FERMT2			C9orf84,SNX30,BX119904.1							
FBLN2,FBXW8,FGF12	GNAL,HCRTR2,HIF1AN			RNU6-1323P							
FGFR2,FTO,FUBP3	HMCN2,HSPD1P15,LAMA2										
GJB6,GNGT1,HAR1A	LCMT1,LINC00430,LINC00507										
HECTD2,HMCN2,HSPD1P15	LINC01376,LINC01440,LINC02199										
IQGAP2,ITPR1,KIF26B	LYRM7,MBD2,MCPH1										
LINC00607,LINC01248,LINC01517	MELK,MGAT5,MIR4475										

LINC02100,LINC02199,LSAMP	MMP28,MRPS35P3,NDFIP1,										
LYZL1,MDFI,ME1	NHSL1,NIM1K,NPLOC4										
MELK,MIR4475,MPDZ	NPSR1,NPSR1-AS1,PCP4										
NDRG4,NHSL1,NPLOC4	PHF21B,PLAGL1,PSG4										
NPSR1,NPSR1-AS1,NR5A2	PSG5,PSMC1P6,RAPGEF4										
OSBP2,PCAT5,PCGF5	RF00017,RF00019,RN7SKP190										
PGAM5P1,PHLPP1,PITPNC1	RN7SKP93,RNU1-150P,RNU6-1323P										
PLAGL1,PPFIBP2,PRDM16	RNU6ATAC21P,RPL12P4,RUNX1										
RBFOX1,RN7SKP167,RN7SL87P	RYR3,SAMSN1,SEPSECS-AS1										
RNA5SP428,RNFT2,RNU6-728P	SLC4A4,SNX30,SOX6										
RPL23AP46,RPL7P37,RPS15AP27	SPARC,SPOCK1,ST3GAL1										
RPS20P25,RUNX1,RUNX2	SYCP2L,TAF15,TEMN3-AS1										
RWDD4,SLC17A2,SLC18A2	TMPRSS15,TPO,TPST1										
SNX30,SPARC,SPOCK1	TRIP12,UBBP5,UGGT1,VAV2										
ST3GAL1,SYCP2L,TENM3	AADACL2-AS1,AC009117.1,AC009509.1										
TENT5A,TMEM232,TPRG1	,AC009869.1,AC010255.3,AC022784.1										
TSPAN18,UBE2V1P12,VSTM4	AC024909.1,AC025252.1,AC026396.1,										

APPENDIX B

Results

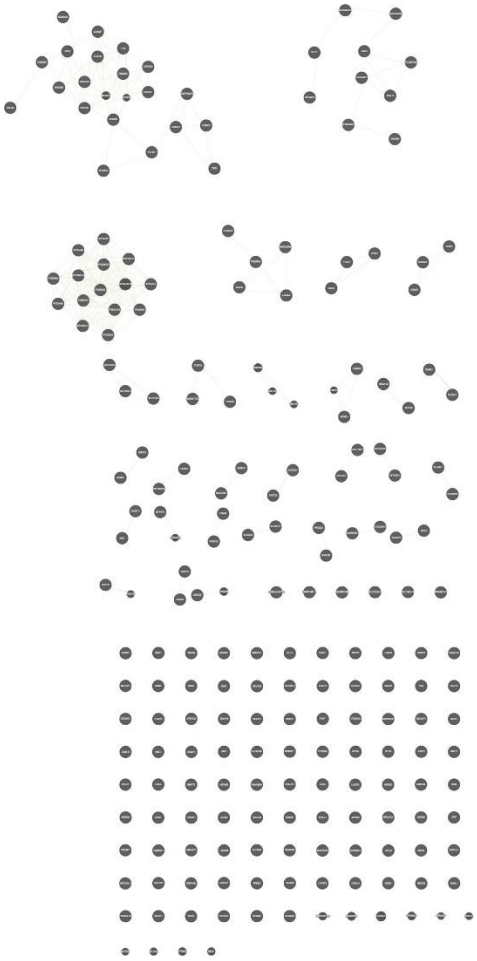

Figure 4.7. Topological relationship for Ensemble > 2.31 scored, only overlapped genes for shared protein domains attribute. Nodes represent genes, genes that we have searched are indicated with stripes and strings represent relation between genes. 110 genes are not connected.

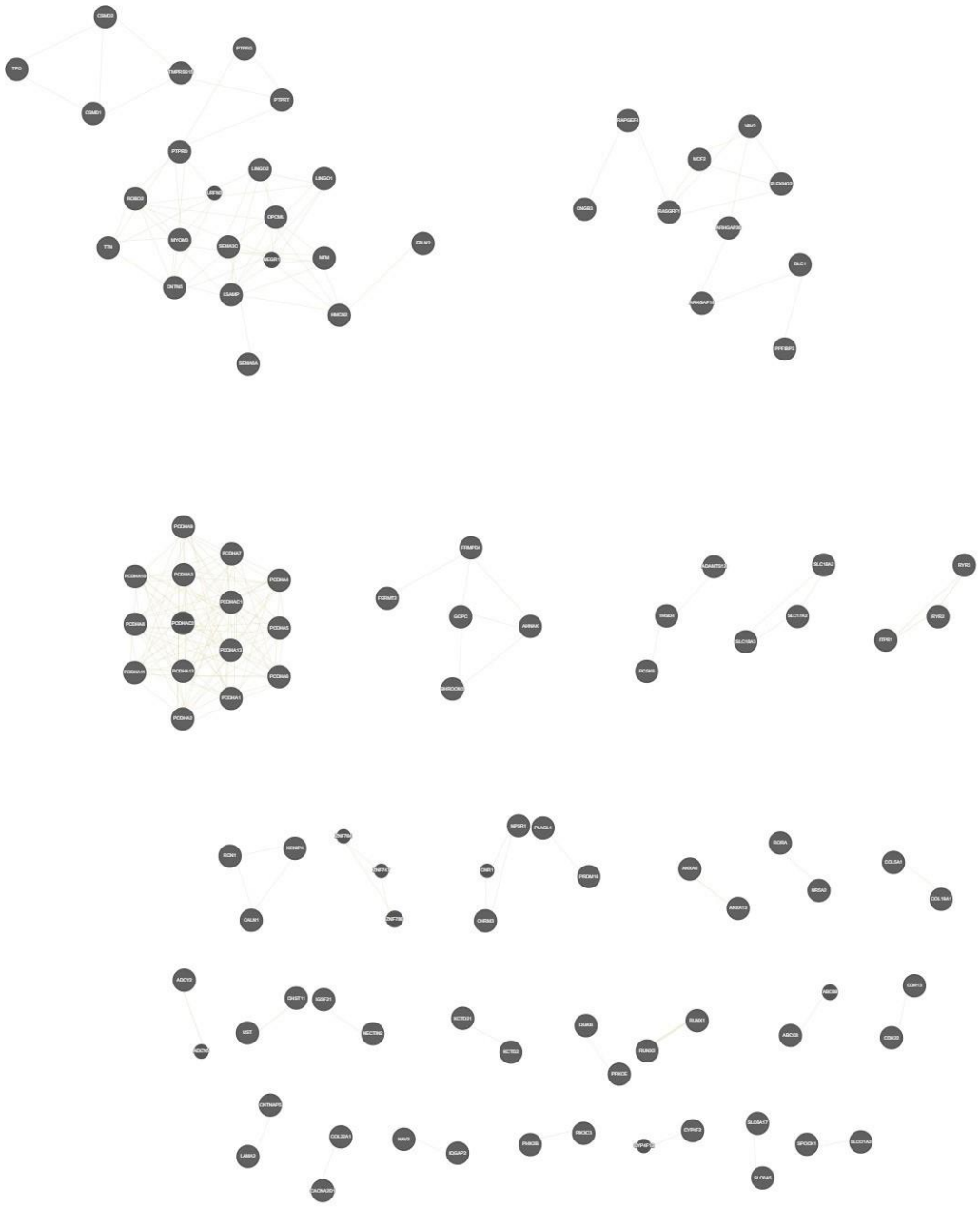

Figure 4.8. Topological relationship for Ensemble 3.31, 3.21 and 2.31 scored, only overlapped genes for shared protein domains attribute. Nodes represent genes, genes that we have searched are indicated with stripes and strings represent relation between genes. Only connected gene are showed.

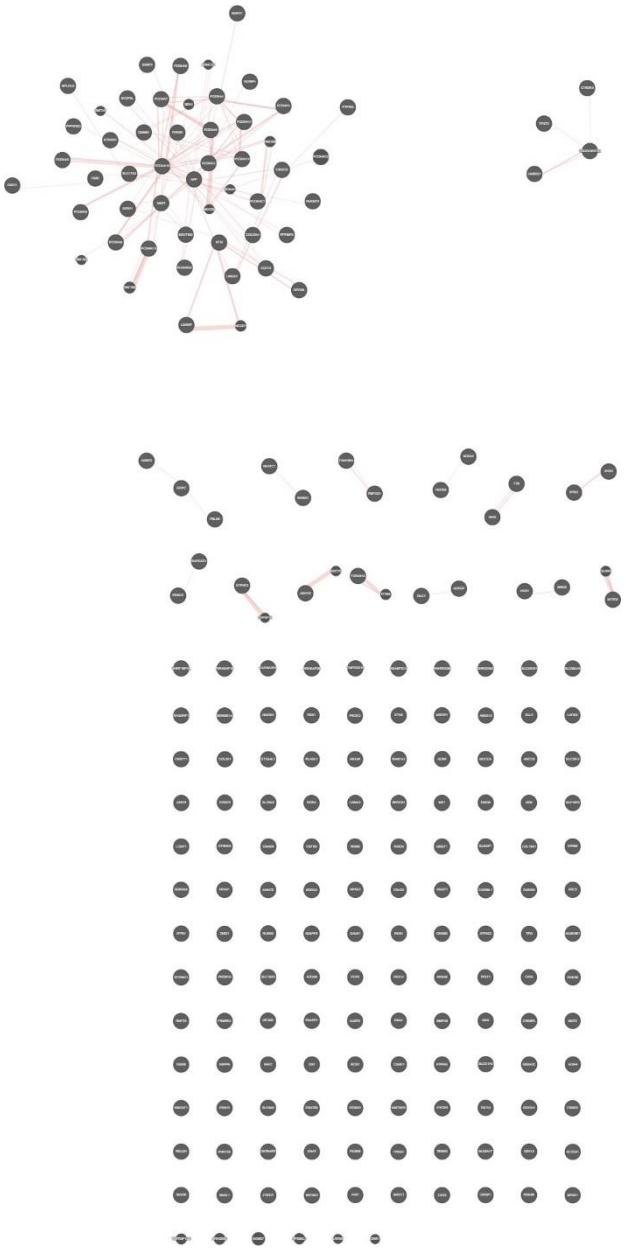

Figure 4.9. Topological relationship for Ensemble > 2.31 scored, only overlapped genes for physical interactions attribute. Nodes represent genes, genes that we have searched are indicated with stripes and strings represent relation between genes. 136 genes are not connected.

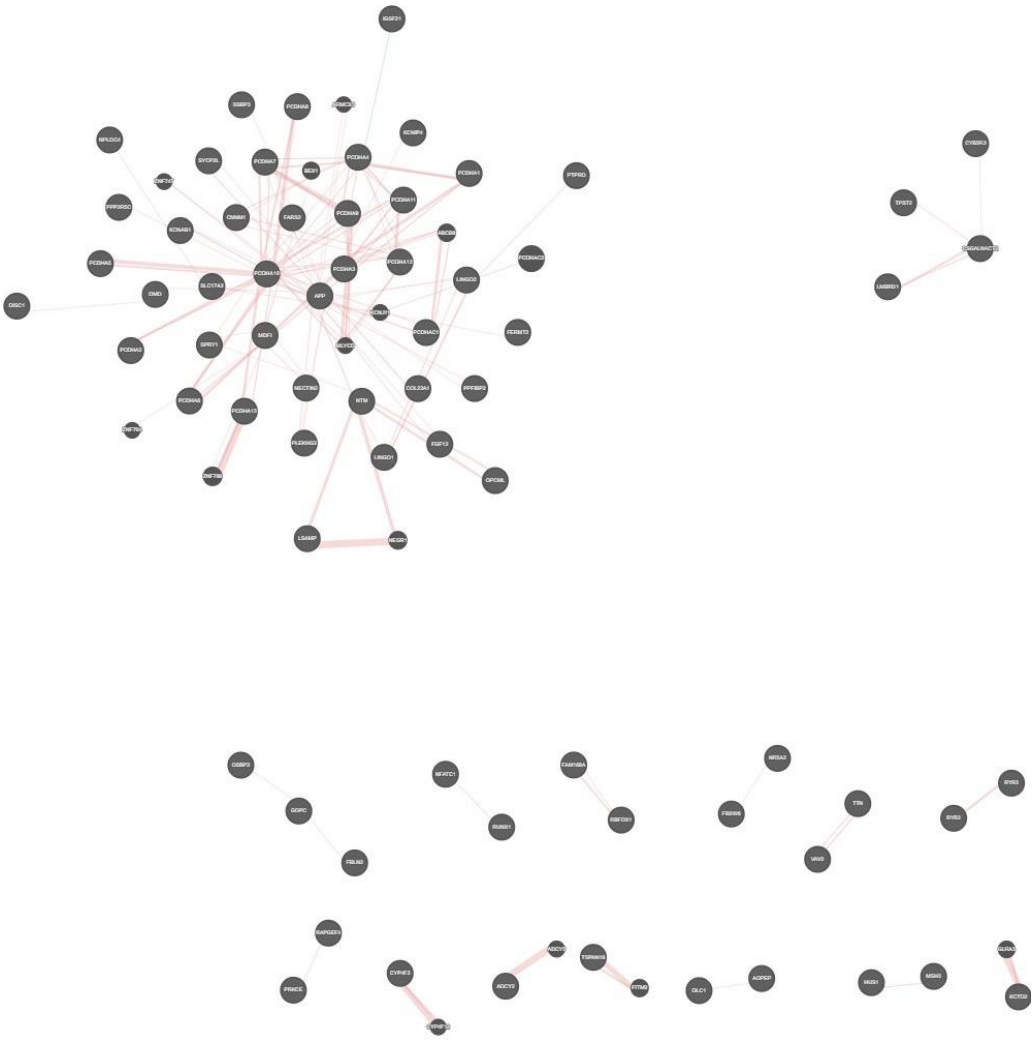

Figure 4.10. Topological relationship for Ensemble > 2.31 scored, only overlapped genes for physical interactions attribute. Nodes represent genes, genes that we have searched are indicated with stripes, and strings represent the relation between genes. Only connected genes are shown.

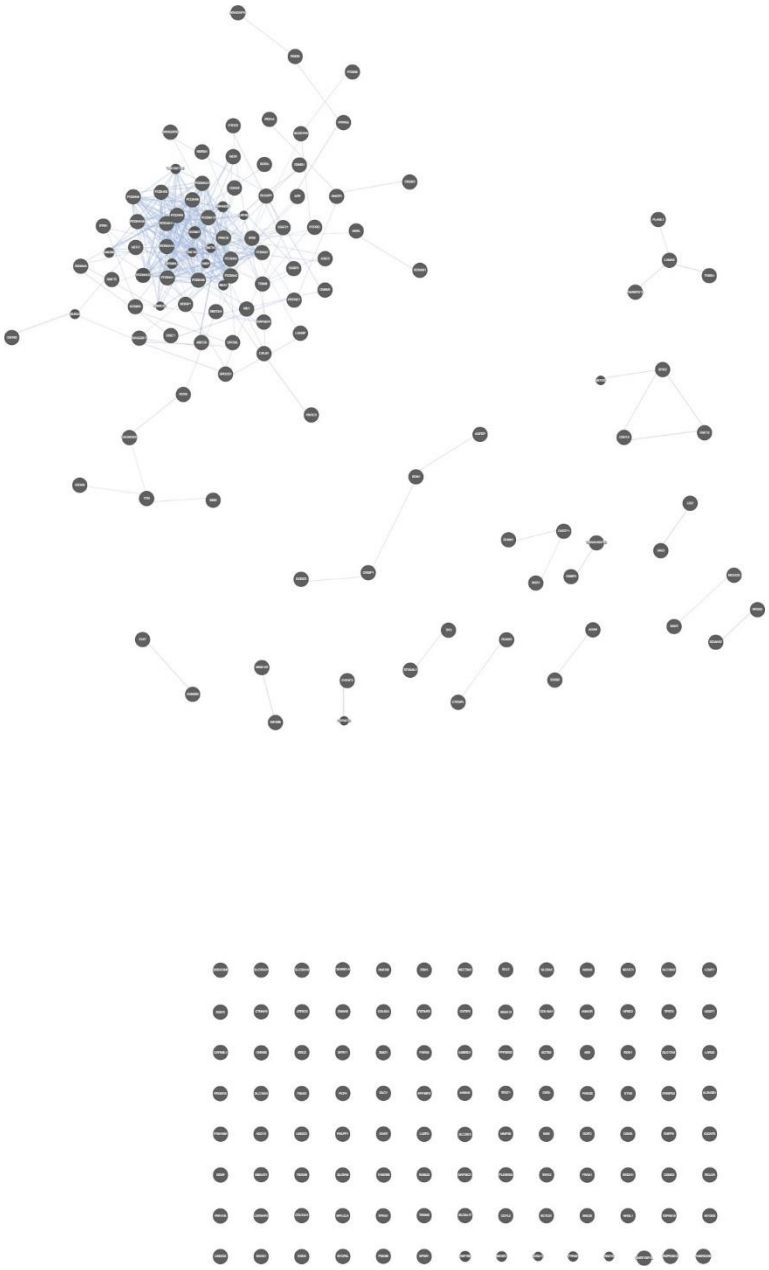

Figure 4.11. Topological relationship for Ensemble >2.31 scored, only overlapped genes for co-localization attribute. Nodes represent genes, genes that we have searched are indicated with stripes and strings represent relation between genes. 105 genes are not connected.

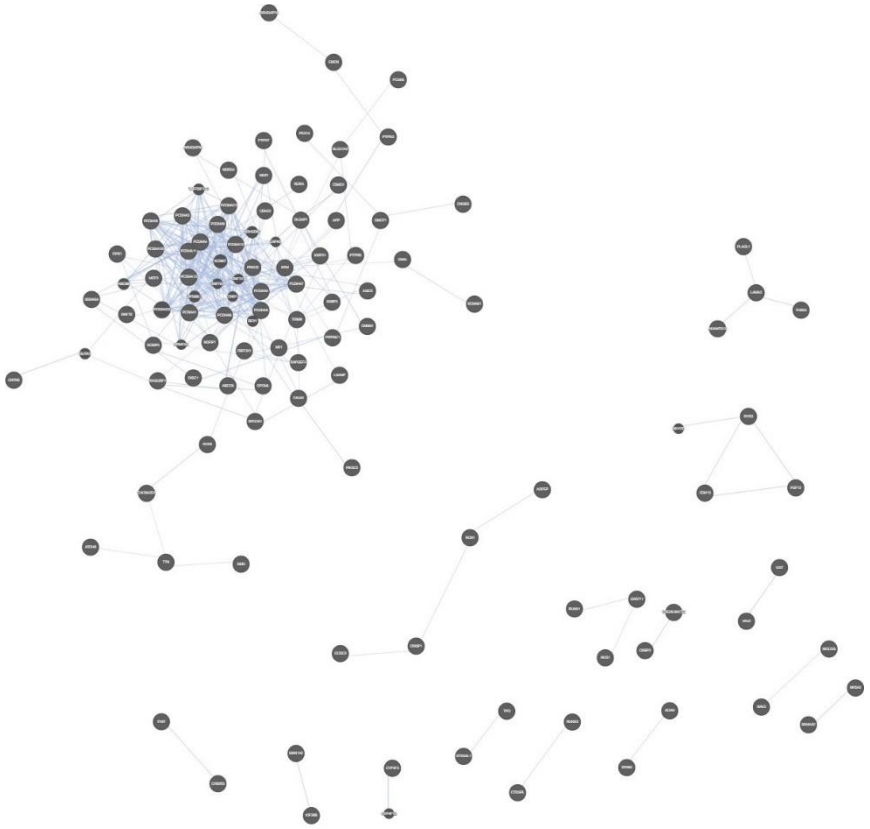

Figure 4.12. Topological relationship for Ensemble > 2.31 scored, only overlapped genes for co-localization. Nodes represent genes, genes that we have searched are indicated with stripes and strings represent relation between genes. Only connected genes are showed.

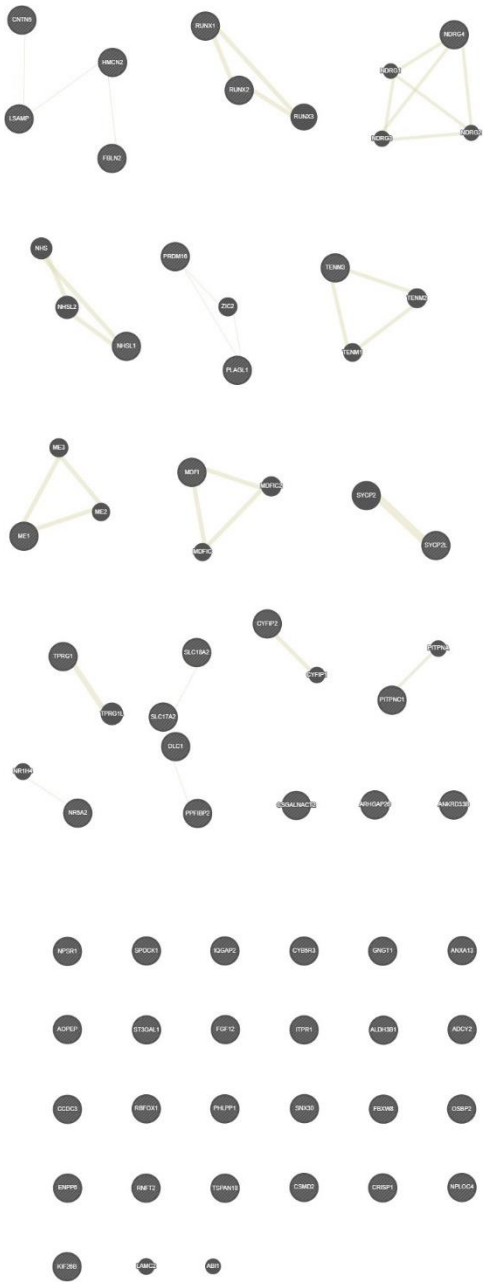

Figure 4.13. Topological relationship for Ensemble > 3.21 scored, only overlapped genes for shared protein domains. Nodes represent genes, genes that we have searched are indicated with stripes and strings represent relation between genes. 27 genes are not connected.

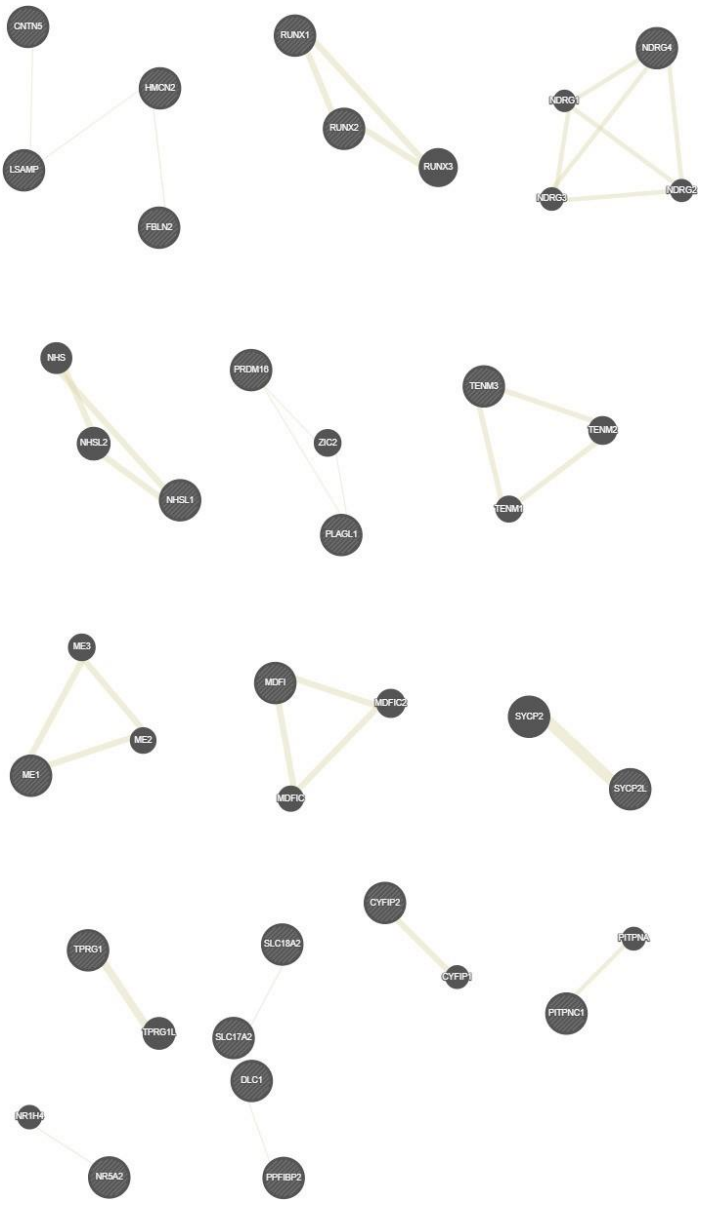

Figure 4.14. Topological relationship for Ensemble >3.21 scored, only overlapped genes for shared protein domains. Nodes represent genes, genes that we have searched are indicated with stripes, and strings represent the relation between genes. Only connected genes are shown.

Figure 4.15. Topological relationship for Ensemble >3.21 scored, only overlapped genes for physical interaction attribute. Nodes represent genes, genes that we have searched are indicated with stripes, and strings represent the relation between genes. Sixty genes are not connected.

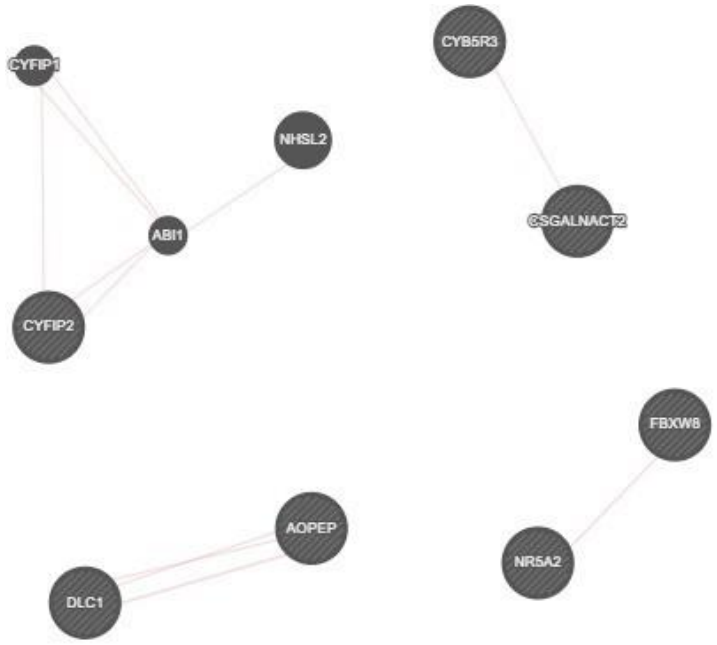

Figure 4.16. Topological relationship for Ensemble > 3.21 scored, only overlapped genes for physical interaction attribute. Nodes represent genes, genes that we have searched are indicated with stripes, and strings represent the relation between genes. Only connected genes are shown.

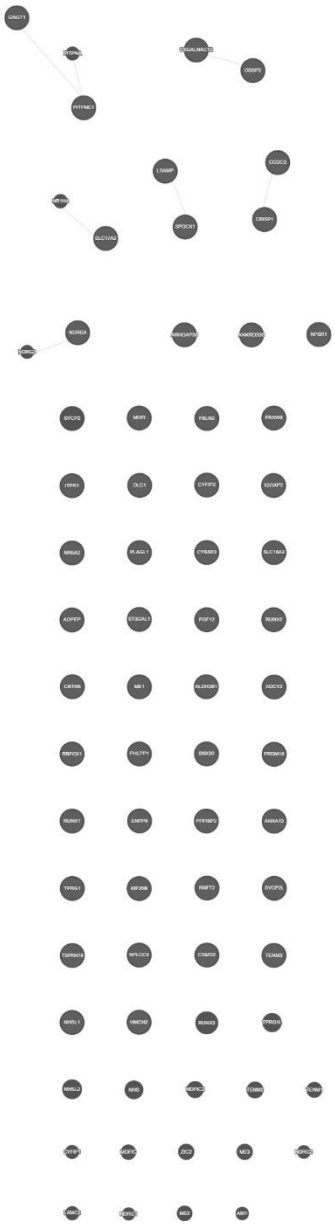

Figure 4.17. Topological relationship for Ensemble > 3.21 scored, only overlapped genes for co-localization attribute. Nodes represent genes, genes that we have searched are indicated with stripes, and strings represent the relation between genes. Fifty-seven genes are not connected.

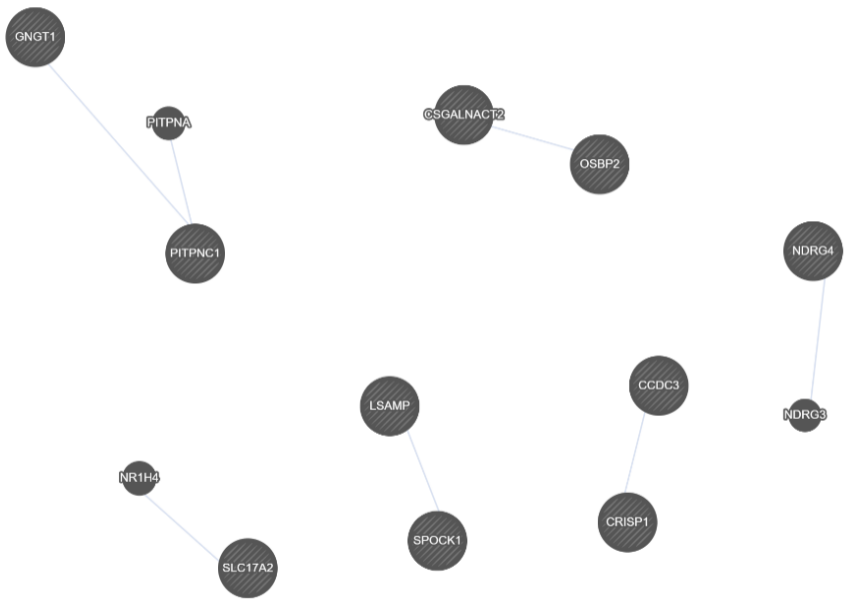

Figure 4.18. Topological relationship for Ensemble > 3.21 scored, only overlapped genes for co-localization attribute. Nodes represent genes, genes that we have searched are indicated with stripes, and strings represent the relation between genes. Only connected genes are shown.

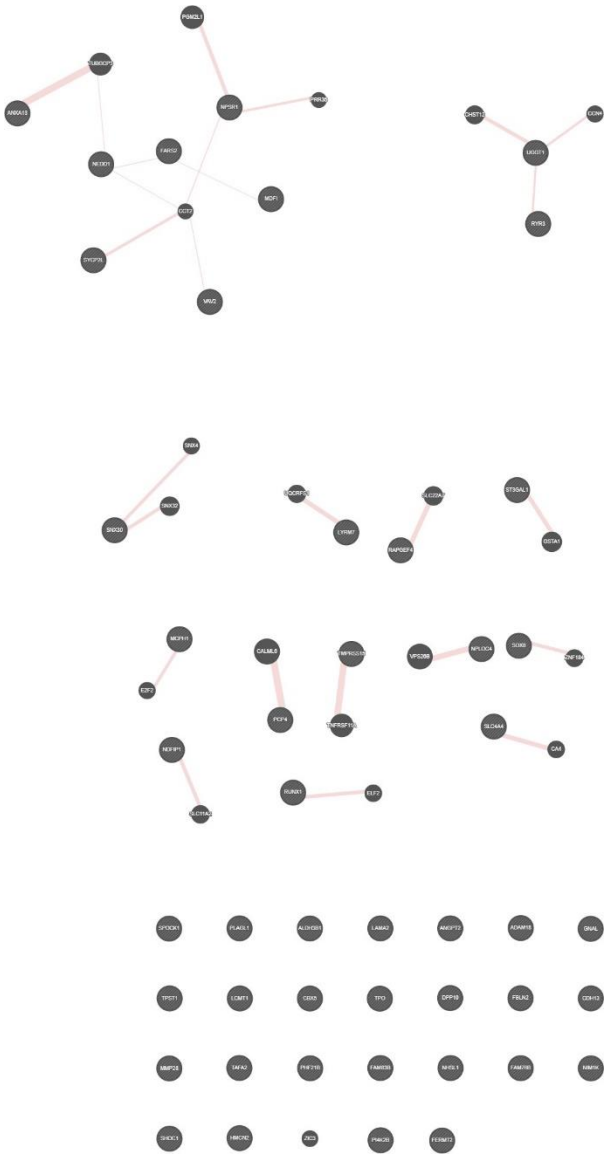

Figure 4.19. Topological relationship for Entropy only overlapped genes for physical interactions attribute. Nodes represent genes, genes that we have searched are indicated with stripes, and strings represent the relation between genes. Twenty-six genes are not connected.

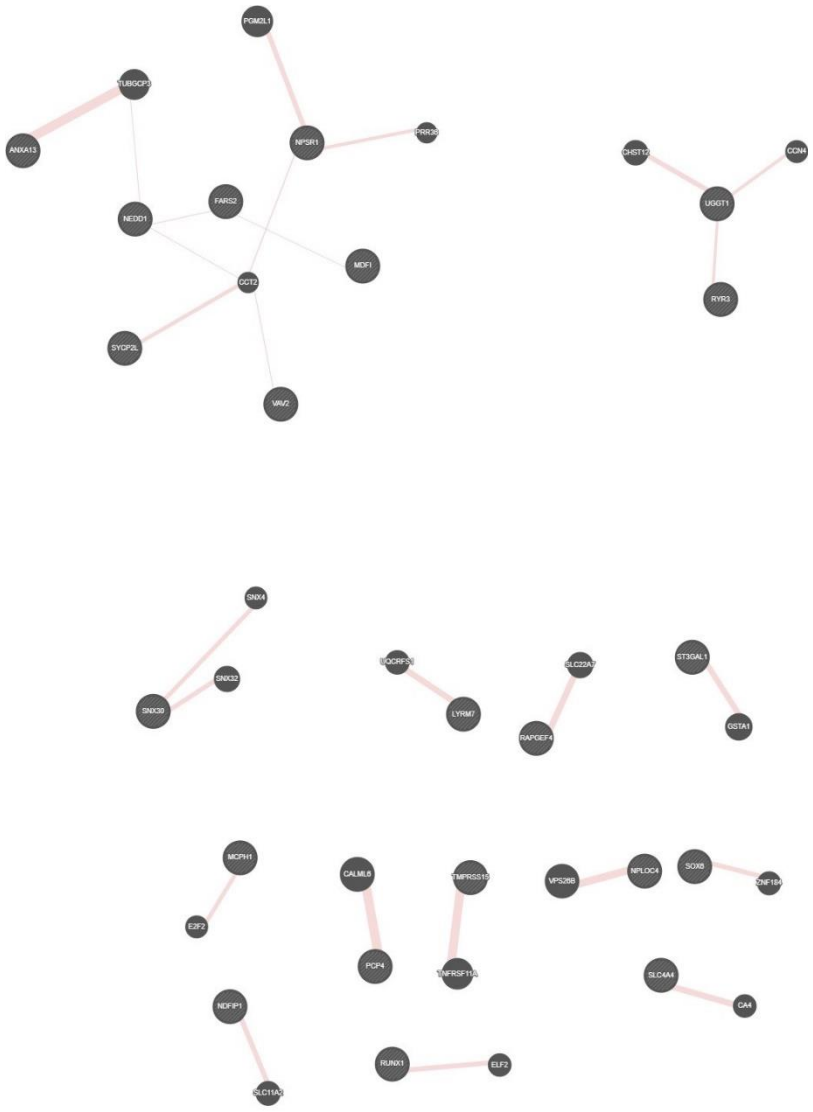

Figure 4.20. Topological relationship for Entropy only overlapped genes for physical interactions attribute. Nodes represent genes, genes that we have searched are indicated with stripes, and strings represent the relation between genes. Only connected genes are shown.

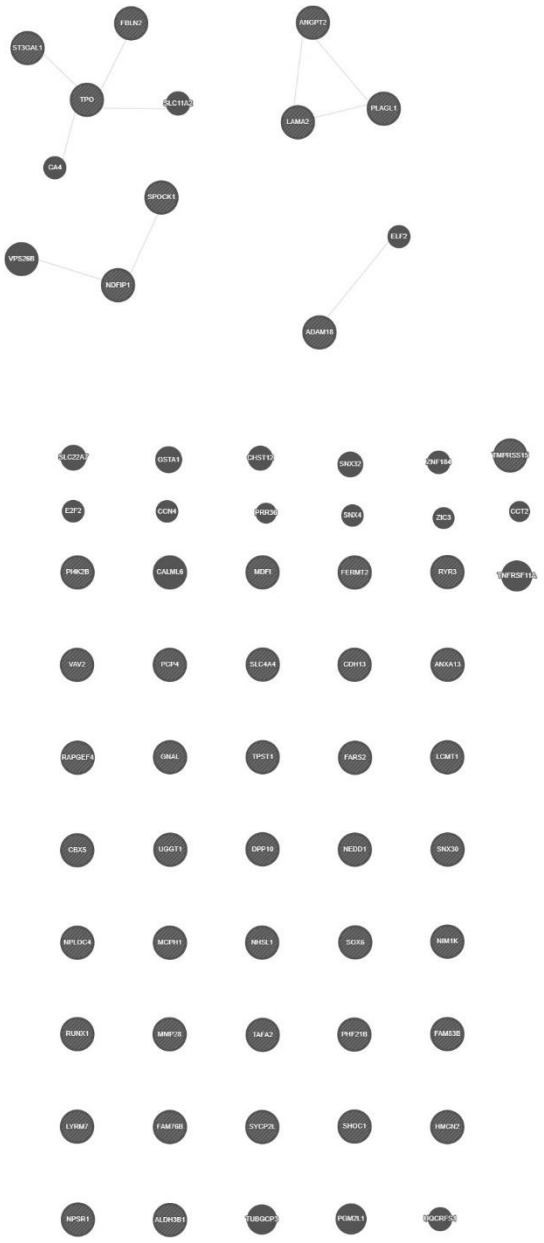

Figure 4.21. Topological relationship for Entropy only overlapped genes for co-localization interactions attribute. Nodes represent genes, genes that we have searched are indicated with stripes, and strings represent the relation between genes. Fifty-three genes are not connected.

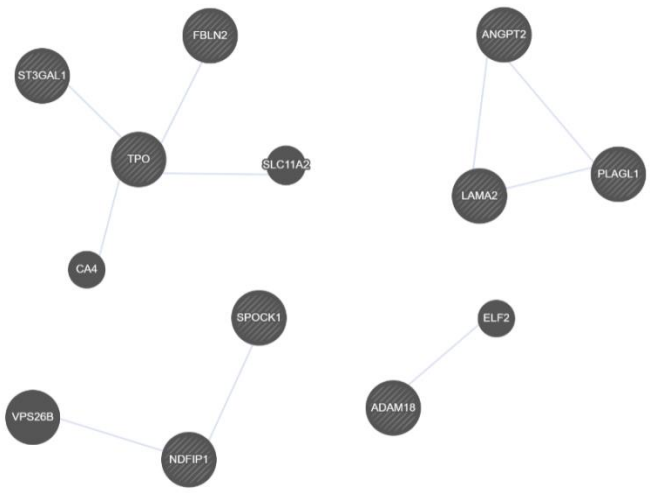

Figure 4.22. Topological relationship for Entropy only overlapped genes for co-localization interactions attribute. Nodes represent genes, genes that we have searched are indicated with stripes, and strings represent the relation between genes. Only connected genes are shown.

Table 4.2. Genes in the GeneMANIA networks for Ensemble > 2.31, Ensemble >3.21, and Entropy groups.

Ensemble > 2.31_physical_int	Ensemble >2.31_co_loc	Ensemble > 3.21_physical_int	Ensemble >3.21_co_loc	Entropy_physical_int	Entropy_co_loc
ABCB9,APP, C9orf3	ABCB9,ABCC8,ADAMTS12	DLC1	CSGALNACT2	RYR3,VAV2	LAMA2
ADCY2,ARMCX3, CNNM1	AOAH,APP,ARHGAP18	FBXW8	CRISP1	MDFI,NEDD1	TPO
ADCY5,BEX1, COL22A1	ARHGAP26, ARHGDIG,ARMCX3	ABI1	GNGT1	CALML6	PLAGL1
CSGALNACT2, CYP4F2, DMD	ASIC2,ATOH8,BEX1	CYB5R3	SPOCK1	VPS26B	SPOCK1
CYB5R3,DISC1, FAM168A	C9orf3,CACNA2D1,CALN1	CYFIP1	NR1H4	PGM2L1	VPS26B
CYP4F12,DLC1, FARS2	CCDC3,CDH13,CDH22	C9orf3	PITPNA	TNFRSF11A	SLC11A2
FBLN2,FGF12, GOPC	CHST11,CNGB3,CNNM1	NR5A2	NDRG3	TUBGCP3	CA4
FBXW8,FITM2, HUS1	CNR1,ZNF764	CYFIP2	OSBP2	GSTA1	ELF2
FERMT2,GLRA3, IGSF21	CNTN5,CSMD1,CYP4F12	CSGALNACT2	CCDC3	SLC22A7	FBLN2
KCNAB1,KCTD2, LMBRD1	CRISP1,CTDSPL,CYP4F2	NHSL2	PITPNC1	SNX32	ANGPT2
KCNIP4,LINGO1, LSAMP	CSGALNACT2,CYB5R3,DISC1		LSAMP	CHST12	ST3GAL1
KCNJ11,LINGO2, MDFI	DLGAP1, FGF12, GNAL		SLC17A2	SLC11A2	NDFIP1
MLYCD,NFATC1, NTM	DMD, FHIT, GNGT1		NDRG4	UQCRFS1	ADAM18
MSH3,NPLOC4, OPCML	EMCN, GLRA3, HUS1			ZNF184	
NEGR1,NR5A2,OSBP2	IGSF21,KCNIP4,LAMA2			CA4	
PCDHA1,PCDHA12, PCDHA3	ITPR1, KCNK7, LRFN3			ELF2	
PCDHA10, PCDHA13, PCDHA4	KCNAB1,KIF26B,LSAMP			E2F2	
PCDHA11,PCDHA2, PCDHA5	MAN1A2, ME1, NAV2			CCN4	
PCDHA6,PCDHA9, PLEKHG2	MCF2, MLYCD, NDFIP1			SNX4	
PCDHA7, PCDHAC1, PPFIBP2	MDFI, MOCOS, NDRG4			CCT2	
PCDHA8,PCDHAC2, PPP2R5C	NR5A2,OSBP2,PCDHA11			PRR36	
PRKCE,RAPGEF4,RYR2	NTM,PCDHA1,PCDHA12			FARS2	
PTPRD,RBFOX1,RYR3	OPCML,PCDHA10,PCDHA13			UGGT1	
PVRL2,RUNX1,SLC17A2	PCDHA2,PCDHA5,PCDHA8			PCP4	
SPRY1, TPST2, VAV2	PCDHA3,PCDHA6,PCDHA9			NPLOC4	
SSBP3,TSPAN18,ZNF747	PCDHA4,PCDHA7,PCDHAC1			NPSR1	

SYCP2L,TTN,ZNF764	PCDHAC2,PIK3C3,PRKCE			TMPRSS15	
ZNF786	PCSK6,PITPNC1,PTPRD			ANXA13	
	PEX14, PLAGL1, PTPRG			ST3GAL1	
	PTPRT, RBFOX1, RORA			RAPGEF4	
	RAPGEF4,RCN1,RPS6KL1			SNX30	
	RASGRF1,RNFT2,RUNX1			NDFIP1	
	RUNX2,SEMA3C,SPOCK1			LYRM7	
	RYR2, SEMA6A, SSBP3			SOX6	
	RYR3,SLCO1A2,ST3GAL1			SLC4A4	
	STAB2, TPO, UST			RUNX1	
	THSD4, TRIM9, VAV2			MCPH1	
	TNFRSF13C,TTN,ZNF747			SYCP2L	

APPENDIX C

Results

Table 4.7. Genes and diseases which are found in **Figure 4.29**.

SHARED NAME	ENTREZ ID	Type	Role	SHARED NAME	ENTREZ ID	Type	Role
PCP4	5121	Gene/Protein	Candidate-Gene/Protein	PLAGL1	5325	Gene/Protein	Candidate-Gene/Protein
ANXA6	309	Gene/Protein	Candidate-Gene/Protein	KCNAB1	7881	Gene/Protein	Candidate-Gene/Protein
PRDM16	63976	Gene/Protein	Candidate-Gene/Protein	LAMA2	3908	Gene/Protein	Candidate-Gene/Protein
FERMT2	10979	Gene/Protein	Candidate-Gene/Protein	DGKB	1607	Gene/Protein	Candidate-Gene/Protein
FARS2	10667	Gene/Protein	Candidate-Gene/Protein	CTNNA3	29119	Gene/Protein	Training-Gene/Protein
ARHGAP18	93663	Gene/Protein	Training-Gene/Protein	CYB5R3	1727	Gene/Protein	Candidate-Gene/Protein
SEMA3C	10512	Gene/Protein	Candidate-Gene/Protein	NFATC1	4772	Gene/Protein	Candidate-Gene/Protein
FRMPD4	9758	Gene/Protein	Candidate-Gene/Protein	ROBO2	6092	Gene/Protein	Candidate-Gene/Protein
ERC2	26059	Gene/Protein	Candidate-Gene/Protein	TPST1	8460	Gene/Protein	Candidate-Gene/Protein
NPLOC4	55666	Gene/Protein	Candidate-Gene/Protein	TPST2	8459	Gene/Protein	Candidate-Gene/Protein
OSBP2	23762	Gene/Protein	Candidate-Gene/Protein	CDH13	1012	Gene/Protein	Candidate-Gene/Protein
SSBP3	23648	Gene/Protein	Candidate-Gene/Protein	CTDSPL	10217	Gene/Protein	Candidate-Gene/Protein
STAB2	55576	Gene/Protein	Candidate-Gene/Protein	PPP2R5C	5527	Gene/Protein	Candidate-Gene/Protein

TRIM9	114088	Gene/Protein	Candidate-Gene/Protein	UGGT1	56886	Gene/Protein	Candidate-Gene/Protein
EMCN	51705	Gene/Protein	Candidate-Gene/Protein	CHAT	1103	Gene/Protein	Training-Gene/Protein
FBXW8	26259	Gene/Protein	Candidate-Gene/Protein	PI4K2B	55300	Gene/Protein	Candidate-Gene/Protein
NAV2	89797	Gene/Protein	Candidate-Gene/Protein	SCHIZOPHRENIA; SCZD	MIM181500	Disease	Candidate-Disease
SLC6A5	9152	Gene/Protein	Candidate-Gene/Protein	PIK3C3	5289	Gene/Protein	Training-Gene/Protein
PTPRT	11122	Gene/Protein	Candidate-Gene/Protein	COL5A1	1289	Gene/Protein	Candidate-Gene/Protein
RCN1	5954	Gene/Protein	Candidate-Gene/Protein	RASGRF1	5923	Gene/Protein	Candidate-Gene/Protein
TMPRSS15	5651	Gene/Protein	Candidate-Gene/Protein	IQGAP2	10788	Gene/Protein	Candidate-Gene/Protein
PTPRD	5789	Gene/Protein	Candidate-Gene/Protein	ITPR1	3708	Gene/Protein	Candidate-Gene/Protein
LCMT1	51451	Gene/Protein	Training-Gene/Protein	VAV2	7410	Gene/Protein	Candidate-Gene/Protein
PCDHA6	56142	Gene/Protein	Training-Gene/Protein	ARHGAP26	23092	Gene/Protein	Candidate-Gene/Protein
EXD3	54932	Gene/Protein	Candidate-Gene/Protein	PLEKHG2	64857	Gene/Protein	Candidate-Gene/Protein
NXN	64359	Gene/Protein	Candidate-Gene/Protein	CHRM3	1131	Gene/Protein	Candidate-Gene/Protein
OPCML	4978	Gene/Protein	Candidate-Gene/Protein	MCF2	4168	Gene/Protein	Candidate-Gene/Protein
SLCO1A2	6579	Gene/Protein	Candidate-Gene/Protein	ZMIZ1	57178	Gene/Protein	Candidate-Gene/Protein
NDFIP1	80762	Gene/Protein	Candidate-Gene/Protein	PCDHA4	56144	Gene/Protein	Candidate-Gene/Protein
LINGO1	84894	Gene/Protein	Candidate-Gene/Protein	MDFI	4188	Gene/Protein	Candidate-Gene/Protein
PVRL2	5819	Gene/Protein	Training-Gene/Protein	ANXA13	312	Gene/Protein	Candidate-Gene/Protein
SPOCK1	6695	Gene/Protein	Candidate-Gene/Protein	DLGAP1	9229	Gene/Protein	Candidate-Gene/Protein
MAN1A2	10905	Gene/Protein	Candidate-Gene/Protein	RYR2	6262	Gene/Protein	Training-Gene/Protein

RAPGEF4	11069	Gene/Protein	Candidate-Gene/Protein	IGSF21	84966	Gene/Protein	Candidate-Gene/Protein
KCNIP4	80333	Gene/Protein	Candidate-Gene/Protein	DLC1	10395	Gene/Protein	Candidate-Gene/Protein
LSAMP	4045	Gene/Protein	Candidate-Gene/Protein	NR5A2	2494	Gene/Protein	Candidate-Gene/Protein
PEX14	5195	Gene/Protein	Candidate-Gene/Protein	PHLPP1	23239	Gene/Protein	Candidate-Gene/Protein
SHROOM3	57619	Gene/Protein	Candidate-Gene/Protein	AHNAK	79026	Gene/Protein	Candidate-Gene/Protein
SPRY1	10252	Gene/Protein	Candidate-Gene/Protein	FBLN2	2199	Gene/Protein	Training-Gene/Protein
GNAL	2774	Gene/Protein	Candidate-Gene/Protein	ADCY2	108	Gene/Protein	Candidate-Gene/Protein
GNGT1	2792	Gene/Protein	Candidate-Gene/Protein	CIITA	4261	Gene/Protein	Training-Gene/Protein
CYFIP2	26999	Gene/Protein	Candidate-Gene/Protein	DISC1	27185	Gene/Protein	Training-Gene/Protein
RYR3	6263	Gene/Protein	Candidate-Gene/Protein	PRKCE	5581	Gene/Protein	Candidate-Gene/Protein
PCSK6	5046	Gene/Protein	Candidate-Gene/Protein	TTN	7273	Gene/Protein	Candidate-Gene/Protein
HUS1	3364	Gene/Protein	Candidate-Gene/Protein	ETV6	2120	Gene/Protein	Candidate-Gene/Protein
FGF12	2257	Gene/Protein	Candidate-Gene/Protein	GOPC	57120	Gene/Protein	Candidate-Gene/Protein
PPFIBP2	8495	Gene/Protein	Candidate-Gene/Protein	ROS1	6098	Gene/Protein	Candidate-Gene/Protein
PTPRG	5793	Gene/Protein	Candidate-Gene/Protein	DIABETES MELLITUS	MIM125853	Disease	Candidate-Disease
RBFOX1	54715	Gene/Protein	Training-Gene/Protein	FAM76B	143684	Gene/Protein	Candidate-Gene/Protein
ELL2	22936	Gene/Protein	Candidate-Gene/Protein	PITPNC1	26207	Gene/Protein	Candidate-Gene/Protein

APPENDIX D

Whole gene lists used in this work.

Ensembl>2.31	Ensembl3.21>	Entropy-all	Entropy-ADNI	Entropy-GenADA	Entropy-NCRAD
PRDM16	PRDM16	AADACL2-AS1	ALDH3B1	AP003398.2	HIF1AN
PEX14	CSMD2	AC009117.1	FAM76B	AC079362.1	PAX2
IGSF21	RP5-1007G16.1	AC009509.1	LINC00507	AC024909.1	RP11-309P22.1
MYOM3	NR5A2	AC009869.1	FERMT2	None	GPAM
CSMD2	KIF26B	AC010255.3	CDH13	AC034268.2	RP11-396O20.2
RP5-1007G16.1	CCDC3	AC022784.1	MMP28	NPLOC4	RP11-222N13.1
SSBP3	CSGALNACT2	AC024909.1	AC092594.1	RUNX1	SOX6
RP11-14O19.2	SLC18A2	AC025252.1	UGGT1	PHF21B	RP11-1060J15.4
SLC6A17	PPFIBP2	AC026396.1	TMPRSS15	FBLN2	CBX5
FAM19A3	TSPAN18	AC026884.1	PI4K2B	SPOCK1	RP11-968A15.2
MAN1A2	ALDH3B1	AC034268.2	NDFIP1	NHSL1	FAM19A2
PBX1	CNTN5	AC068787.1	SYCP2L	MCPH1	Y_RNA
NR5A2	RP11-13A1.3	AC073592.2	RP11-637O19.3	ANGPT2	NCOR2
DISC1	RNFT2	AC073592.8	PLAGL1	ST3GAL1	SCARB1
MLK4	FBXW8	AC079362.1	NPSR1-AS1	HMCN2	RP11-955H22.2
RYR2	RBFOX1	AC079380.1	NPSR1		RP11-351O2.1
CHRM3	NDRG4	AC087379.1	TPST1		LINC00430
KIF26B	PITPNC1	AC090515.3	RP11-3P22.2		UBBP5
CCDC3	NPLOC4	AC091895.1	VAV2		RP11-33N16.3
RP11-462L8.1	PHLPP1	AC092100.1			RYR3
CSGALNACT2	AC107057.2	AC092364.1			LCMT1
CHAT	AC108025.2	AC093277.1			GNAL

SLC18A3	RP4-568F9.6	AC093462.1			MBD2
CTNNA3	RUNX1	AC093840.1			POLI
ZMIZ1	OSBP2	AC096992.1			RP11-420K14.8
HPSE2	CYB5R3	AC104662.2			PSG5
CNNM1	AC026188.1	AC105252.1			PSG4
SLC18A2	ITPR1	AC105450.1			TPO
PPFIBP2	FBLN2	AC106744.1			DPP10
ABCC8	LSAMP	AC106800.1			RAPGEF4
NAV2	RP11-454C18.2	AC117382.2			DNER
SLC6A5	TPRG1	AC123767.1			TRIP12
LUZP2	FGF12	ADAM10			PCP4
RP11-587D21.4	TENM3	ADAM18			AC008132.13
RCN1	ENPP6	AF127936.2			DGCR6
TSPAN18	ADCY2	AL024498.2			EEFSEC
AHNAK	ANKRD33B	AL137230.2			DNAJB8
ALDH3B1	CTC-340D7.1	AL138885.1			RP11-102M11.2
FAM168A	IQGAP2	AL139317.5			RP11-731C17.1
KCTD21	SPOCK1	AL161716.1			RP11-454C18.2
FAM76B	ARHGAP26	AL353595.1			SLC4A4
CNTN5	CYFIP2	AL589740.1			RP11-285A15.1
SLC35F2	SYCP2L	AL589826.1			RP11-427M20.1
NTM	RP11-637O19.3	AL713851.1			RNU1-150P
OPCML	SLC17A2	ALDH3B1			AC108105.1
ETV6	MDFI	ANGPT2			NIM1K
SLCO1A2	RUNX2	ANXA13			RP11-447H19.1
RP11-13A1.3	CRISP1	AP003398.2			CTC-441N14.4
UHRF1BP1L	ME1	APOOP2			CTC-210G5.1
STAB2	RP11-378G13.2	BX119904.1			RP11-166A12.1
CHST11	NHSL1	C9orf84			LYRM7

RP11-438N16.1	PLAGL1	CBX5			FARS2
RNFT2	NPSR1-AS1	CDH13			FAM83B
FBXW8	NPSR1	CTC-210G5.1			RP3-523K23.2
LINC00507	GNGT1	CTC-441N14.4			HCRTR2
LINC00347	AC009784.3	DGCR6			RP11-406O16.1
MIR4500HG	DLC1	DNAJB8			RP3-453D15.1
SLC25A21	RP11-145O15.3	DNER			RP11-67P15.1
TRIM9	ANXA13	DPP10			LAMA2
FERMT2	ST3GAL1	EEFSEC			LINC-PINT
RP11-33N16.3	C9orf3	FAM19A2			RP11-63E5.6
PPP2R5C	SNX30	FAM230F			FAM87A
RP11-580I1.2	HMCN2	FAM76B			RP11-10A14.4
RYR3		FAM83B			CTD-2024D23.1
RP11-108K3.1		FAM87A			ADAM18
RORA		FARS2			ANXA13
THSD4		FBLN2			MELK
LINGO1		FERMT2			MIR4475
RASGRF1		GNAL			C9orf84
AEN		GPAM			SNX30
PCSK6		HCRTR2			RP11-232D9.1
RBFOX1		HIF1AN			RNU6-1323P
CIITA		HMCN2			NEDD1
LCMT1		HSPD1P15			MDFI
NDRG4		LAMA2			
CDYL2		LCMT1			
CDH13		LINC00430			
ATP2C2		LINC00507			
NXN		LINC01376			
DNAH9		LINC01440			

ASIC2		LINC02199			
MMP28		LINC-PINT			
PITPNC1		LYRM7			
KCTD2		MBD2			
SLC38A10		MCPH1			
NPLOC4		MELK			
DLGAP1		MGAT5			
GNAL		MIR4475			
AQP4-AS1		MMP28			
MOCOS		MRPS35P3			
PIK3C3		NDFIP1			
KIAA1468		NHSL1			
PHLPP1		NIM1K			
NFATC1		NPLOC4			
CTC-548K16.2		NPSR1			
CYP4F2		NPSR1-AS1			
CTC-513N18.7		PCP4			
RP11-420K14.8		PHF21B			
PLEKHG2		PLAGL1			
PVRL2		PSG4			
CTC-326K19.6		PSG5			
TPO		PSMC1P6			
AC019118.2		RAPGEF4			
AC107057.2		RF00017			
AC108025.2		RF00019			
LINC00298		RN7SKP190			
AC092594.1		RN7SKP93			
PRKCE		RNU1-150P			
RP11-444A22.1		RNU6-1323P			

AC007131.2		RNU6ATAC21P			
ATOH8		RP11-166A12.1			
CNTNAP5		RP11-285A15.1			
UGGT1		RP11-309P22.1			
RAPGEF4		RP11-406O16.1			
TTN-AS1		RP11-427M20.1			
TTN		RP11-63E5.6			
RP4-568F9.6		RP11-968A15.2			
PTPRT		RPL12P4			
CDH22		RUNX1			
TMPRSS15		RYR3			
APP		SAMSN1			
RUNX1		SEPSECS-AS1			
AF064860.5		SLC4A4			
PCP4		SNX30			
TPST2		SOX6			
OSBP2		SPARC			
RP1-272J12.1		SPOCK1			
CYB5R3		ST3GAL1			
PHF21B		SYCP2L			
AC026188.1		TAF15			
ITPR1		TEMN3-AS1			
FBLN2		TMPRSS15			
CTDSPL		TPO			
LARS2		TPST1			
ERC2		TRIP12			
FHIT		UBBP5			
PTPRG		UGGT1			
ROBO2		VAV2			

LSAMP					
SLC9A9					
RP11-454C18.2					
KCNAB1					
TPRG1					
FGF12					
KCNIP4					
PI4K2B					
RP11-725D20.1					
SHROOM3					
FRAS1					
EMCN					
SPRY1					
RP11-93I21.3					
TENM3					
ENPP6					
RP11-217E13.1					
ADCY2					
ANKRD33B					
ADAMTS12					
AMACR					
RP11-1084J3.4					
GDNF					
CTC-340D7.1					
IQGAP2					
PDE8B					
MSH3					
ELL2					
SEMA6A					

SPOCK1					
PCDHA1					
PCDHA2					
PCDHA3					
PCDHA4					
PCDHA5					
PCDHA6					
PCDHA7					
PCDHA8					
PCDHA9					
PCDHA10					
PCDHA11					
PCDHA12					
PCDHA13					
PCDHAC1					
NDFIP1					
ARHGAP26					
MIR143HG					
ANXA6					
CYFIP2					
RP11-541P9.3					
FARS2					
SYCP2L					
RP11-637O19.3					
MBOAT1					
CASC15					
LRRC16A					
SLC17A2					
ZSCAN12P1					

MDFI					
RUNX2					
CRISP1					
RP11-406O16.1					
LMBRD1					
COL19A1					
ME1					
RP11-378G13.2					
RP3-399L15.3					
ROS1					
GOPC					
LAMA2					
ARHGAP18					
NHSL1					
PLAGL1					
UST					
DGKB					
NPSR1-AS1					
NPSR1					
AOAH					
HUS1					
RP11-1217F2.15					
TPST1					
RP11-3P22.2					
CALN1					
SEMA3C					
CACNA2D1					
GNGT1					
AC009784.3					

ERICH1					
CSMD1					
DLC1					
RP11-145O15.3					
RP11-383H13.1					
CNGB3					
RP11-398G24.2					
ANXA13					
ST3GAL1					
COL22A1					
PTPRD					
LINGO2					
C9orf3					
C9orf84					
SNX30					
DENND1A					
HMCN2					
VAV2					
COL5A1					
EXD3					
FRMPD4					
DMD					
MCF2					

Acknowledgements

** The investigators within the ADNI, GenADA and NCRAD contributed to the design and implementation of and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Data collection and sharing for ADNI data for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding support for the dbGaP datasets: National Institute on Aging - Late Onset Alzheimer's Disease Family Study: Genome-Wide Association Study for Susceptibility Loci. dbGaP Study Accession: phs000168.v1.p1. and Multi-Site Collaborative Study for Genotype-Phenotype Associations in Alzheimer's disease and Longitudinal follow-up of Genotype-Phenotype Associations in Alzheimer's disease and Neuroimaging component of Genotype-Phenotype Associations in Alzheimer's disease. dbGaP Study Accession: phs000219.v1.p1