INTEGRATION OF MACHINE LEARNING AND ENTROPY METHODS FOR POST-GENOME-WIDE ASSOCIATION STUDIES ANALYSIS

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

BY

BURCU YALDIZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF MEDICAL INFORMATICS

AUGUST 2022

Approval of the thesis:

INTEGRATION OF MACHINE LEARNING AND ENTROPY METHODS FOR POST-GENOME-WIDE ASSOCIATION STUDIES ANALYSIS

Submitted by BURCU YALDIZ in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Health Informatics Department, 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 Dept., METU** _____

Examining Committee Members:

Prof. Dr. Cem İyigün
Industrial Engineering Dept., METU _____

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

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

Assoc. Prof. Ceren Sucularlı
Bioinformatics Dept., Hacettepe University _____

Assoc. Prof. Dr. Nurcan Tunçbağ
Chemical and Biological Engineering Dept., Koç University _____

Date: 31.August.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 : Burcu Yaldız

Signature : _____

ABSTRACT

INTEGRATION OF MACHINE LEARNING AND ENTROPY METHODS FOR POST-GENOME-WIDE ASSOCIATION STUDIES ANALYSIS

Yaldız, Burcu

Ph.D., Department of Health Informatics

Supervisor: Assoc. Prof. Dr. Yeşim Aydın Son

August 2022, 79 pages

Non-linear relationships between genotypes play an essential role in understanding the genetic interactions of complex disease traits. Genome-Wide Association Studies (GWAS) have revealed a statistical association between the SNPs in many complex diseases. As GWAS results could not thoroughly explain the genetic background of these disorders, Genome-Wide Interaction Studies started to gain importance. In recent years, various statistical approaches such as entropy-based methods have been suggested for revealing these non-additive interactions between variants. This study integrates an entropy-based 3-way interaction information method and machine learning (PLINK-Random Forest-Random Forest) workflow to capture the hidden patterns resulting from non-linear relationships between genotypes in Late-Onset Alzheimer's Disease (LOAD) to discover early and differential diagnosis markers. We have optimized an entropy-based approach that detects the third-order interactions in PLINK-RF-RF models from three different LOAD datasets. A reduced SNP set was selected for all three datasets by 3WII analysis of PLINK-RF-RF prioritized SNPs, promising a model minimization approach. Selected triplets of SNPs that show significant differences between case and control groups in terms of 3WII are proposed as candidate biomarkers for a genotyping-based LOAD diagnosis. Among SNPs prioritized by 3WII, four out of 19 SNPs from GenADA, one out of 27 from ADNI, and four out of 106 NCRAD are mapped to genes directly associated with Alzheimer's Disease. For the first time, we have integrated the RF-RF model with the entropy-based model for determining the three-way epistatic interactions for LOAD and discovered the common biological pathways for ADNI, GenADA, and NCRAD datasets.

Keywords: Biomarker, three-way interaction, entropy, GWAS, Alzheimer's Disease

ÖZ

GENOM BOYUNCA İLİŞKİLENDİRME ÇALIŞMALARI SONRASI ANALİZİ İÇİN MAKİNE ÖĞRENMESİ VE ENTROPİ YÖNTEMLERİNİN ENTEGRASYONU

Yaldız, Burcu

Doktora, Sağlık Bilişimi Bölümü
Tez Yöneticisi: Doç. Dr. Yeşim Aydın Son

Ağustos 2022, 79 sayfa

Genotipler arasındaki doğrusal olmayan ilişkiler, kompleks hastalık özelliklerinin genetik etkileşimlerini anlamada önemli bir rol oynar. Genom Boyutunda İlişkilendirme Çalışmaları (GWAS), birçok kompleks hastalıkta SNP'lerin istatistiksel ilişkisini ortaya çıkarmıştır. GWAS sonuçları bu bozuklukların genetik arka planını tam olarak açıklayamadığından, Genom Boyutunda Etkileşim Çalışmaları önem kazanmaya başlamıştır. Son yıllarda, varyantlar arasındaki bu toplamsal olmayan etkileşimleri ortaya çıkarmak için entropi tabanlı yöntemler gibi çeşitli istatistiksel yaklaşımlar önerilmiştir. Bu çalışmada, Geç Başlangıçlı Alzheimer Hastalığı'nda (LOAD) erken ve ayırıcı tanı belirteçlerini keşfetmek amacı ile genotipler arasındaki doğrusal olmayan ilişkilerden kaynaklanan gizli örüntüleri ortaya çıkarmak için entropi tabanlı üçlü etkileşim bilgi yöntemi (3WII) ile bir makine öğrenmesi(PLINK-Random Forest-Random Forest) iş akışı entegre edilmiştir. Üç farklı LOAD veri seti ıle geliştirilmiş PLINK-RF-RF modellerindeki üçüncü dereceden etkileşimlerin yakalanmasını sağlayan entropi tabanlı yaklaşımı optimize edilmiştir. PLINK-RF-RF ile önceliklendirilmiş SNPlerin 3WII analizi ile üç veri kümesinin tümü için indirgenmiş bir SNP seti seçilmiş olup bu yaklaşım, bir model minimizasyon yaklaşımı olarak umut vericidir. 3WII açısından vaka ve kontrol grupları arasında anlamlı farklılık gösteren seçilmiş SNP üçlüleri, genotiplendirmeye dayalı LOAD teşhisi için aday biyobelirteçler olarak önerilmektedir. 3WII tarafından önceliklendirilen SNP'lerden; GenADA'dan gelen 19 SNP'den 4'ü, ADNI'den gelen 27 SNP'den 1'i ve NCRADdan gelen 106 SNP'den 4'ü, Alzheimer Hastalığı ile doğrudan ilişkili genlere haritalandırılmıştır. Bu çalışmada, LOAD'da üç yönlü epistatik etkileşimleri belirlemek için ilk kez RF-RF modeli entropi tabanlı model ile entegre edilmiş ve ADNI, GenADA ve NCRAD veri setlerinin ortak biyolojik yolakları keşfedilmiştir.

Anahtar Sözcükler: Biyobelirteç, üç yönlü etkileşim, entropi, GWAS, Alzheimer Hastalığı

To My Family

ACKNOWLEDGMENTS

First, I would like to express my deepest gratitude to my supervisor, Assoc. Prof. Dr. Yeşim Aydın Son for her guidance, patience, and encouragement. She supported me throughout this process not only with her academic guidance but also with her personality, kindness and sincerity, and endless patience. I feel lucky to have walked this difficult path under her guidance.

Besides my supervisor, I would like to thank Prof. Dr. Cem İyigün for his recommendations on the statistical problems we encountered during this work.

I would also like to thank Assoc. Prof. Dr. Nurcan Tunçbağ, Assoc. Prof. Dr. Ceren Sucularlı and Assist. Prof. Aybar Can Acar for tracking and reviewing my work.

I thank my colleagues Onur Erdoğan and Sevda Rafatov for their help and contributions. They were always so kind to help and support me.

I am grateful to Kübra Narcı for her support, encouragement, and priceless friendship. We walked together on this path, and I was lucky to find her next to me whenever I stumbled.

I want to thank Esra Yazıcıoğlu, one of my sisters, who helped me in the preliminary studies of this thesis and did not spare me her endless love and support. I appreciate her and Özgür Gültekin for their unconditional friendship with me during one of my most difficult times and for motivating me to start this study during those times as a source of inspiration with their scientific identities.

I cannot thank enough my other sister Feruze Birer for her priceless friendship. Even if she was thousands of kilometers away but still next to me, she inspired me to make the right decisions with her calmness and wisdom.

I would like to thank to my brother Özgür and his wife Fatma, my aunts and women heroes Ümit Çağlar and Buket Çağlar, my grandmother Mürüvvet Çağlar, my uncle Alp, his wife Esra and cousins Ece and Arda Çağlar for their endless support and love.

Last but not least, special thanks to my mother, Çiğdem Yaldız, and my father, Kadir Yaldız. I would have never been able to achieve this without their endless support, love, trust, and patience in every step of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GWAS	Genome-Wide Association Studies
SNP	Single Nucleotide Polymorphism
MDR	Multifactor dimensionality reduction
RF	Random Forest
3WII	Three-way interaction information
TCI	Total correlation information
AD	Alzheimer's Disease
EOAD	Early onset Alzheimer's Disease
LOAD	Late onset Alzheimer's Disease
NGS	Next generation sequencing
WGS	Whole genome sequencing
WES	Whole exome sequencing
IG	Information gain
IIG	Interaction information gain
LD	Linkage Disequilibrium
CV	Cross Validation
GAD	Genetic Association of Complex Diseases and Disorders
OMIM	Online Mendelian Inheritance in Man
HGMD	The Human Gene Mutation Database

CHAPTER 1

INTRODUCTION

Genome-Wide Association Studies (GWAS) explore the statistical association of the SNPs in complex genetic disorders using high-dimensional datasets (Marees et al., 2018). Primarily, these associations are identified with single-locus approaches, whereby each SNP is tested individually for the association. However, the univariate approach could not explain a large proportion of genetic heritability in most complex diseases. Interactions at higher dimensions such as SNP-SNP, gene-gene, and gene-environment interactions can address the missing heritability problem (Cordell, 2009).

While some of these interactions are identified in small-scale studies, most are revealed in Genome-Wide Interaction Studies (GWIS). A variety of machine learning methods, such as multifactor dimensionality reduction (MDR) and random forest (RF), are used to disclose the complex interactions of the variants(Bureau et al., 2005; McKinney et al., 2006; Oki & Motsinger-Reif, 2011). Also, Entropy-based methods have been proposed to analyze non-linear relationships between genotypes in complex diseases(Ferrario & König, 2018).

Different study designs have suggested various entropy-based approaches for pairwise, third-order, and high-order interactions, such as family-based, case only, and case-control. Information gain, defined as the difference between the mutual information in case and control groups, is used to measure pairwise interactions between two markers. Three-way interaction information (3WII) and total correlation information are two quantities used for assessing third-order interactions. 3WII describes the amount of information common to all variables not present in any other subset alone. In contrast, total correlation information (TCI) reveals the total dependence among the attributes. Nevertheless, different methods have been introduced to assess the pairwise and third-order interactions based on these measurements(Fan, R, Zhong, M, Wang, 2011; T. Hu, Chen, Kiralis, Collins, et al., 2013a; Kwon et al., 2014; Su et al., 2015).

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that is the most common cause of late-onset dementia. More than 55 million people live with dementia worldwide currently, and it is estimated that AD may contribute to 60-70% of the cases[1]. AD is characterized by cognitive impairment; however, a significant

[1] *https://www.who.int/News-Room/Fact-Sheets/Detail/Dementia*

heterogeneity can be observed in clinical progression. It is depicted as early-onset (EOAD) and late-onset (LOAD) based on the age of onset, and LOAD constitutes approximately 95% of cases. While EOAD is more likely familial and inherited in a Mendelian pattern, LOAD presents complex genetic inheritance, where interactions of multiple genetic variations and environmental factors affect the phenotype of patients(Reitz & Mayeux, 2014). In early studies, APOE4 was established as a genetic risk factor for LOAD (Corder et al., 1993; Johanna Kuusisto, Keijo Koivisto, Kari Kervinen, Leena Mykkanen, Eeva-Liisa Helkala, Matti Vanhanen, Tuomo Hanninen, Kalevi Py6rala, Y Antero Kesaniemi, Paavo Riekkinen, 1994; Xu et al., 2021). Several risk variants have been revealed in recent years in several GWA studies(Harold et al., n.d.; Lambert et al., 2013; Lee et al., 2011; Reitz et al., 2011). Besides, various studies have identified epistatic interactions(Combarros, van Duijn et al., 2009; Granados et al., 2013; Grunin et al., 2020; Hohman, Bush, Jiang, Brown-Gentry, Torstenson, Dudek, Mukherjee, Naj, Kunkle, Ritchie, Martin, Schellenberg, Mayeux, Farrer, Pericak-Vance, Haines, & Thornton-Wells, 2016; Meda et al., 2013; Zieselman et al., 2014). However, only two-way interactions have been considered in previous studies.

This study aims to detect the third-order interactions in LOAD by calculating the total information common to all three attributes but not present in any subset (3WII) in a case-control study design. We integrated the machine learning algorithms and entropy-based 3-way interaction information method proposed by Fan et al. A large set of SNPs prioritized by PLINK-RF-RF analysis of the LOAD GWAS datasets are analyzed without mapping them to individual genes to reduce the bias. Then the significant SNP combinations are identified by using the entropy-based test statistics. These prioritized SNP combinations are proposed as potential early and differential diagnosis markers.

Chapter 2 explains the Genome Wide Association Studies, epistasis, and methods for detecting epistasis. Machine learning, data mining, and entropy-based methods are explained in detail as we used these approaches in this study. Molecular etiology, AD risk factors, and epistasis in AD are also explained.

Chapter 3 explains the way of obtaining data and methodology in detail. After we overview the methods, data acquisition is explained. Then, we explain data preprocessing, GWAS analysis, and SNP prioritization with Random Forest-Random Forest (RF-RF) method. Afterward, the entropy-based prioritization step is explained in detail. Subsequently, multiple testing comparisons, variants annotation, and functional enrichment analysis are explained.

In Chapter 4, the results are presented and explained. This chapter explains variant prioritization by the LOAD-RF-RF model, three-way interaction information, and two-way mutual information gain and functional enrichment analysis of the prioritized variants.

Chapter 5 discussed the study results and interpreted the importance of our integrated entropy-based approach for revealing the interactions among LOAD-associated variants.

CHAPTER 2

LITERATURE REVIEW AND BACKGROUND INFORMATION

2.1 Single Nucleotide Variants

A variation at a single position in DNA sequence among individuals is called a single nucleotide polymorphism (SNP). In order to classify a variation as an SNP, more than 1% of the population must contain the alternative nucleotide at a particular position in their genome sequence(*Single Nucleotide Polymorphism*, n.d.). SNPs are not distributed across the genome homogenously. They occur in non-coding regions more frequently than in coding regions. While negative selection has globally reduced population differentiation in amino acid-altering mutations, particularly in disease-related genes, positive selection has increased population differentiation in gene regions, particularly nonsynonymous and 5'-UTR variants, resulting in regional adaptation of human populations(Barreiro et al., 2008). Besides, recombination rate can also explain a substantial fraction of the variability in the intensity of nucleotide polymorphism across the human genome (Nachman, 2001).

There might be several different consequences depending on where SNPs are located. While SNPs in coding regions can cause monogenic disorders (Cordovado et al., 2012), SNPs in non-coding regions can be an indication of a higher risk of cancer (G. Li et al., 2014) or may play a role in gene expression level alterations as an expression quantitative trait locus (eQTL) (Nicolae et al., 2010). Besides, SNPs can lead to differences in traits such as susceptibility to various diseases and physical features between individuals. They can also affect how humans respond to pathogens, chemicals, drugs, and vaccines. Even though many SNPs associated with diseases are considered to increase the risk of disease, a growing number of studies have reported SNPs to be protective, decreasing the risk of certain diseases (Cohen et al., 2005; Steinthorsdottir et al., 2014).

2.2 Genome Wide Association Studies

Genome-Wide Association Studies (GWAS) explore the statistical association of the SNPs in complex genetic disorders using high-dimensional datasets. These studies compare the frequency of genetic variations between case and control groups to determine the significantly more frequent genetic variations in people with the disease. These variations lead to the development of new methods to diagnose, treat and prevent the disease by identifying the loci that influence the disease predisposition.

They can also give direction to other applications such as reconstructing population history, ancestry determination, and population substructure.

Sequencing the whole human genome as part of the Human Genome Project in the early 2000s and developing new analysis tools led to other projects that aimed to discover more about genomic variations, complex parts of the genome, and regulatory mechanisms. The International HapMap Project proceeded after the Human Genome Project intending to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared. It allowed the spreading of GWAS as it led to the optimization of association analysis by representing the haplotype blocks.

The first GWAS was published for age-related macular degeneration in 2005 (Tam et al., 2019). Afterward, a significant number of genomic regions have been identified for various common diseases such as type 2 diabetes mellitus, coronary artery disease, cancers, bipolar disorder, schizophrenia, and inflammatory bowel disease (De Lange et al., 2017; Z. Li et al., 2017; Mullins et al., n.d.; Nikpay et al., 2015; Sud et al., 2017; Zhao et al., 2017). The NHGRI-EBI GWAS Catalog, a publicly available resource of GWAS and their results, contains 71673 variant-trait associations from 3730 publications as of September 2018 (Buniello et al., 2019).

A typical GWAS is conducted by genotyping individuals using microarray or next-generation sequencing (NGS) methods such as whole-exome sequencing (WES) or whole-genome sequencing (WGS). It can be designed as a case-control study when the trait of interest is dichotomous or with a quantitative study design when the measurements on the whole study sample when the trait is quantitative (Uffelmann et al., 2021).

GWAS data is pre-processed before the association analysis through several quality control steps. Imputation of untyped variants using haplotype phasing and reference populations is performed, followed by the statistical test for association and optionally a meta-analysis. Later independent replications are determined, and the results are interpreted by multiple post-GWAS analysis steps (**Figure 1**). These standardized quality control and analysis protocols are required to rule out the possible biases and errors in each step.

While GWAS results determine the statistical associations of variants and their role in the biological context, they can also be used for disease risk prediction, heritability, and the proportion of variation in the trait that can be explained by genetic variation in the population. Besides, the role of gene-gene and gene-environment interactions in complex traits has been represented in various studies (Fenger et al., 2019; Hohman et al., 2013). However, this univariate approach in GWAS studies ignores complex interactions such as SNP-SNP, gene-gene, or gene-environment interactions (Cordell, 2009).

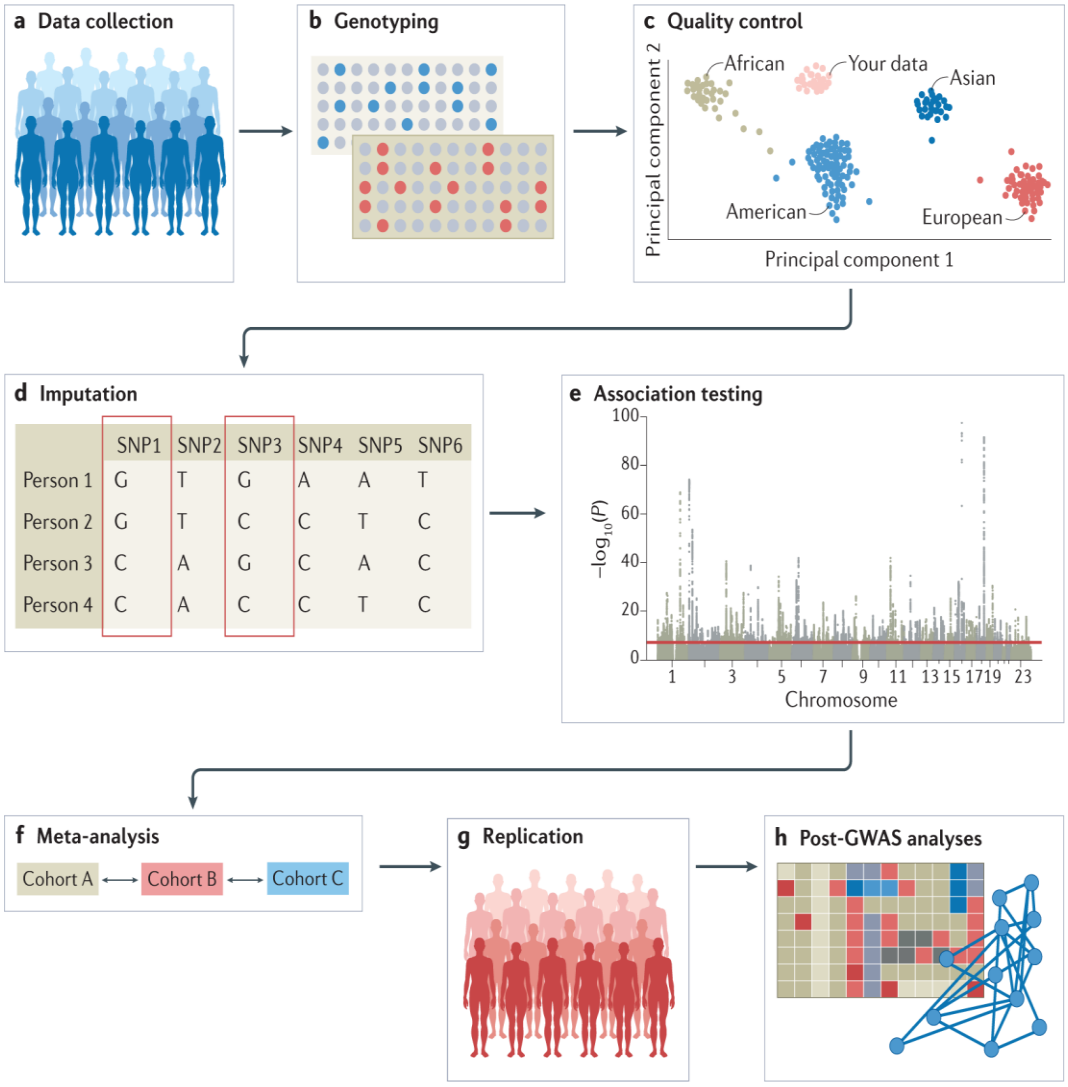


Figure 1 Overview of GWAS analysis steps **a)** Data can be collected from study cohorts, or available genetic and phenotypic information can be used from biobanks or repositories **b)** Microarrays or next generations sequencing methods (whole genome or whole exome sequencing) can be used in order to collect genotypic data **c)** Quality control can be performed at wet- laboratory or dry- laboratory stages (Clustering of individuals according to genetic substrata is demonstrated in the figure) **d)** Untyped genotypes can be imputed using information from matched reference populations from repositories such as 1000 Genomes Project or TopMed. **e)** Each variant can be tested for genetic association using a model such as additive, non-additive, linear, or logistic regression models. After correcting for confounders, including population strata, and controlling the multiple testing, the output is inspected for unusual patterns, and summary statistics are generated. **f)** Results from multiple smaller cohorts are combined using standardized statistical pipelines. **g)** Internal or external replication in an independent cohort can be used for replicating the results. **h)** Information from external resources can be used for in silico analysis of GWAS. Reprinted from Uffelmann et. al. (Uffelmann et al., 2021)

2.3 Epistasis

The additive effect of common variants associated with many complex diseases in GWA studies is relatively small, leading to a missing heritability phenomenon. Epistasis could partially explain the missing heritability in complex traits. Epistasis refers to non-additive interactions between a pair of loci or multiple loci contributing to a single phenotype (Carlborg & Haley, 2004) (**Figure 2).** Epistasis can be considered from two different perspectives, biological and statistical. While biological epistasis implies the interactions between biological components within gene regulatory networks and biochemical pathways, statistical epistasis implies the deviation from additive effects between factors in a model (Moore & Williams, 2005). It is difficult to make inferences about biological function and causation from any statistical result. Besides, correlation is not always an indicator of causality. Therefore, statistical epistasis may not indicate biological epistasis. Similarly, SNPs involved in epistatic interactions may have very low minor allele frequencies, which may lead to biological epistasis that does not imply statistical epistasis(Raghavan & Tosto, 2017). Both have their challenges and benefits.

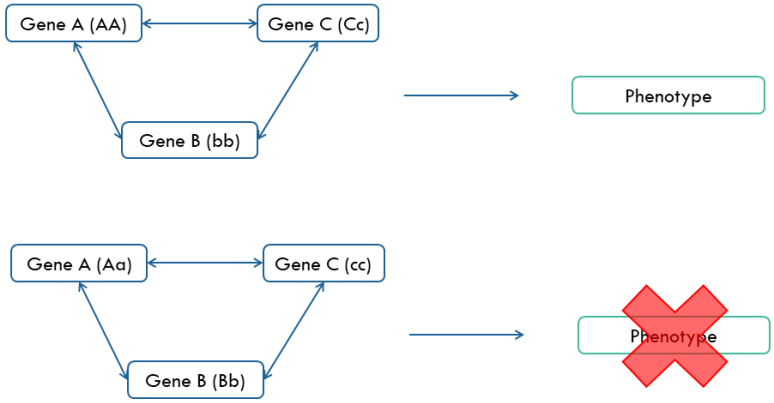


Figure 2 Representation of biological epistasis

Experiments to determine biological epistasis are costly, and interpretation of discovered interactions is challenging. However, discovering epistatic interactions at the biological level provides important functional and pathway information. Some studies demonstrated the proportion of heritability explained by biological epistasis in model organisms. However, due to the complex parameters such as common and rare variants, structural variations, and environmental effects that impact the susceptibility of complex traits, it is hard to estimate the contribution of biologic epistasis to heritability(Gusareva, Kristel, et al., 2014).

The major limitation of the statistical epistasis is the high rates of type 1 and type 2 errors due to the extensive multiple comparisons. An exhaustive search for detecting these interactions demands a high cost of computational resources and sufficiently large sample sizes. While these limitations avoid producing replicable results, statistical epistasis can still give insight into unknown associations by revealing the

non-additive effects of multiple genetic factors(M. T. W. Ebbert, Ridge, & Kauwe, 2015). Several methods have been developed for detecting statistical epistasis in different studies, as described in this section.

2.3.1 Methods for Detecting Epistasis

The methods developed to capture epistasis apply two major approaches: hypothesis-free and hypothesis-driven, depending on whether all SNPs or subset of SNPs are examined for interactions that contribute to the corresponding trait. Wei et al. grouped the methods used in human complex traits in the review they published in 2014(Wei et al., 2014) (**Figure 3**).

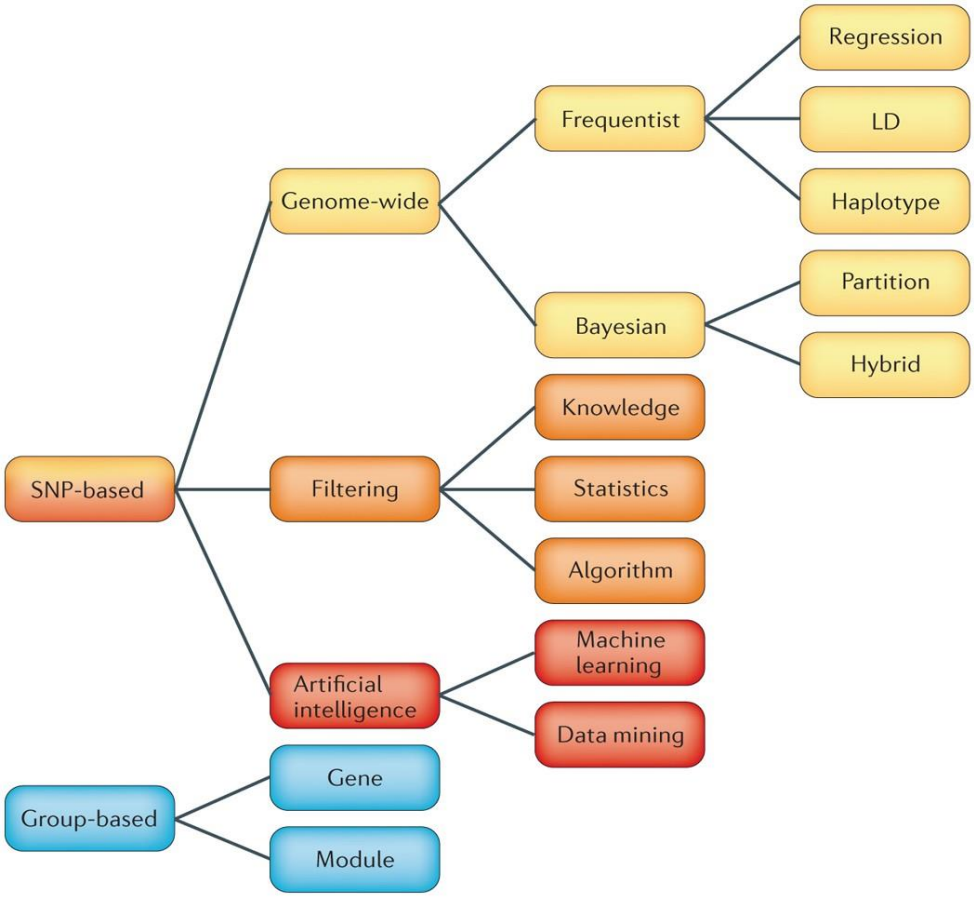


Figure 3 The different types of methods in two major groups based on single-nucleotide-based single-nucleotide polymorphisms (SNPs) and groups of SNPs are outlined. SNP-based methods can be further subdivided based on the type of data analysis (shown in yellow, orange, and red). LD, linkage disequilibrium. Reprinted from Wei et al. (Wei et al., 2014).

Regression-based methods are traditional methods used for assessing the SNP interactions in disease or quantitative traits. In this approach, Fischer's saturated and reduced SNP genotype models can be compared for testing the interactions (Cockerham & Zeng, 1996; Cordell, 2002, 2009; J. K. Hu et al., 2014; Ueki & Cordell,2012). In addition, regression models are also used to examine the subset of

SNP pairs, which are the output of the approximate interaction tests to avoid false positive and redundant signals. Although the advantage of modern computing infrastructure and technologies has been taken to reduce the computational cost of using these methods, genome-wide search requires large sample sizes due to low power. Focusing on SNPs whose marginal effects have been identified in GWAS studies will partially eliminate the problem of low power; however, variants that only affect the phenotype by interaction will be ignored.

Another approach used to define epistasis is LD-based methods. In these methods, the difference between the interactions between loci in the patient and control groups is tested by using joint genotype frequencies. Using chi-square statistics with one degree of freedom with these methods allows us to get faster results than regression-based methods using four degrees of freedom. Besides, SNP interactions between haplotypes can be detected by haplotype-based methods adapted from the LD-based statistics. However, these methods may ignore intra-locus interactions, leading to the effect of unobserved rare variants being overlooked and may also cause high false-positive levels in cases where SNPs may show high correlation or both SNPs have a high marginal effect.

On the other hand, Bayesian methods provide a flexible model and stochastic search for epistasis based on the idea that the difference between genotype frequency distributions between loci will indicate significant interactions. In particular, Bayesian model averaging and hybrid Bayesian (i.e., a combination of Bayesian framework and generalized linear model) approaches improve epistasis detection.

Data filtering methods are based on the idea of testing the interactions of a set of selected SNPs relying on existing biological information, statistical properties, or fast algorithms. These methods have higher power than exhaustive searches since they can be applied faster, require fewer multiple tests, and functional interpretation is less complicated. However, limitations in algorithms and existing knowledge should be considered when applying these methods to avoid potential bias.

It is also possible to test interactions between SNPs grouped on genes or functional modules. While this approach increases the capture power of interactions, it also reduces the burden of multiple testing.

2.3.2 Machine Learning and Data Mining Methods for Detecting Epistasis

Machine learning and data mining methods also capture hidden, novel, and significant patterns through GWAS datasets(Pirooznia M, Seifuddin F, Judy J, 2012). Besides, the high computational and statistical cost of exhaustive search could be reduced by these approaches' classifiers used for data reduction and feature selection.

Various algorithms have varying prediction accuracies, such as decision trees, support vector machines, and Bayesian and neural networks. Random forests (RF) are collections of decision trees that build an ensemble model for classification and regression. RF implementation in genomic analysis has become popular within data-mining techniques that can reveal many predictor variables and generate assembled

models with interactions (Botta et al., 2014; Sinoquet, 2018). Moreover, RF analysis produces variable importance measures that are useful to rank the predictor variables for the prioritization of SNPs.

Nevertheless, since the interactions are not tested saliently, detecting high-level interactions is challenging for data mining and machine learning methods. In addition, these methods also require new approaches for detecting genome-wide higher-order interactions due to limitations such as high computational and multiple testing requirements, along with insufficient sample sizes.

2.3.3 Entropy-Based Methods for Detecting Epistasis

Entropy-based methods analyze nonlinear gene-gene and gene-environment interactions in complex diseases (Dong et al., 2008). Shannon entropy is a nonlinear function that measures the uncertainty of random variables(Shannon, 1949). In other words, entropy is the average level of uncertainty associated with a random variable. The entropy of a discrete random variable X with probability mass function p(x) is defined as:

$$H(X)=-E[\log(P(X)]=-\sum_{X=x} \mathrm{p(x)}\log(\mathrm{p(x)}) \qquad \text{(Eq 1)}$$

Methods based on Shannon entropy measure the strength of prediction of one or 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 & Thomas, 1991).

In a case-control study design disease status of an individual is denoted by D (control=0; case=1), and three di-allelic SNPs are denoted by X, Y, and Z (reference=0, heterozygous=1, homozygous=2). The entropy of a marker X and the combined entropy of two markers X and Y in the general population are defined by the following equations respectively:

$$\mathrm{H(X)}=-\sum_{i=0}^{2} P(G_{X=i})\log\ \mathrm{P}(G_{X=i})$$

$$\mathrm{H(X,Y)}=-\sum_{i=0}^{2}\sum_{j=0}^{2} P\left(G_{X=i}, G_{Y=j}\right)\log\ \mathrm{P}(G_{X=i}, G_{Y=j}) \qquad \text{(Eq 2)}$$

Conditional entropies in the affected population are defined by:

$$\mathrm{H(X\backslash D)}=-\sum_{i=0}^{2} P(G_{X=i}\backslash D=1)\log\ \mathrm{P}(G_{X=i}\backslash D=1)$$

$$\mathrm{H(X,Y\backslash D)}=-\sum_{i=0}^{2}\sum_{j=0}^{2} P\left(G_{X=i}, G_{Y=j}\backslash D=1\right)\log\ \mathrm{P}(G_{X=i}, G_{Y=j}\backslash D=1) \quad \text{(Eq 3)}$$

The combined entropy of three markers in general and the affected population are defined similarly.

The interaction between two markers is measured by mutual information, which is defined in the general and affected population (Cover & Thomas, 1991; Shannon, 1949):

$$I(X,Y)=H(X)+H(Y)-H(X,Y)$$

$$I(X,Y\backslash D)=H(X\backslash D)+H(Y\backslash D)-H(X,Y\backslash D) \quad \text{(Eq 4)}$$

Different study designs have suggested various entropy-based approaches for pairwise, third-order, and high-order interactions, such as family-based, case only, and case-control. The interaction between two markers is measured by mutual information, and information gain, defined as the difference between the mutual information in case and control groups, is used to measure pairwise interactions between two markers. Three-way interaction information (3WII) and total correlation information are two quantities used for assessing third-order interactions. 3WII describes the amount of information common to all variables not present in any other subset alone (**Figure 4**). In contrast, total correlation information (TCI) reveals the total dependence among the attributes. Three-way interaction information is defined in the general population as:

$$I(X,Y,Z)=-H(X)-H(Y)-H(Z)+H(X,Y)+H(X,Z)+H(Y,Z)-H(X,Y,Z) \quad \text{(Eq 5)}$$

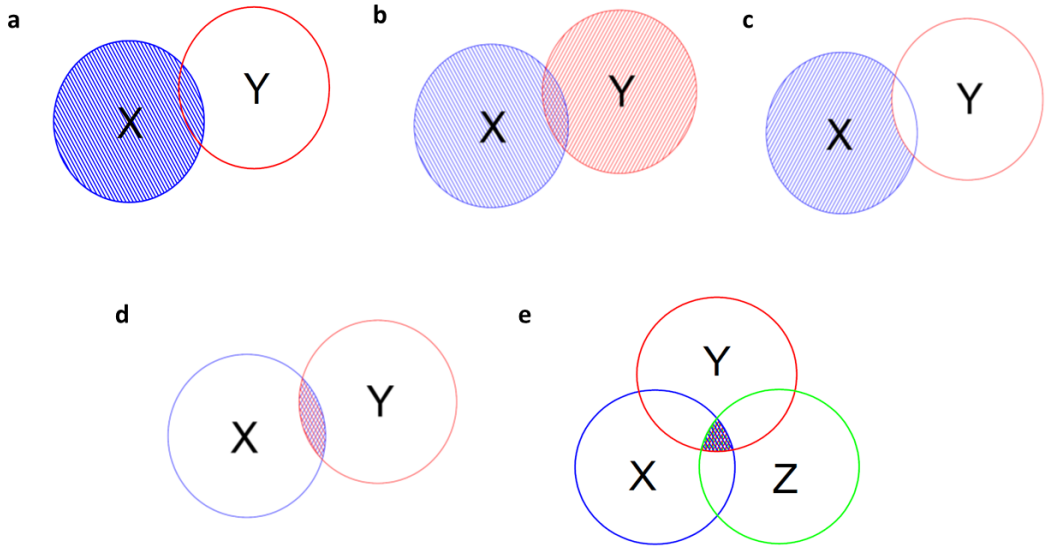


Figure 4 Schematic representation of entropy and entropy-based concepts a) Entropy of a random variable X (Eq 1). b) Joint entropy of two random variables X and Y (Eq 4). c) Conditional entropy of X given the variable Y (Eq 3). d) Mutual information of the variables X and Y. e) Three-way interaction information of three variables X, Y, and Z (Eq 5). (Adapted from Ferrario P.G and König I.R (Ferrario & König, 2018)

Nevertheless, different methods have been introduced based on these measurements to assess the pairwise and third-order interactions. For instance, Fan. et al. proposed an information gain approach based on mutual information, an entropy-based function that measures the interaction between markers (Fan, R, Zhong, M, Wang, 2011). They also developed one-dimensional test statistics to analyze sparse data for investigating 2-way, 3-way, and K-way interactions in case-control settings and applied them to the bladder cancer dataset.

The difference between the mutual information in the affected population and the general population is defined as information gain:

$$IG(X,Y\backslash D)=I(X,Y\backslash D)-I(X,Y) \quad \text{(Eq 6)}$$

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

$$IIG(X,Y,Z\backslash D)=I(X,Y,Z\backslash D)-I(X,Y,Z) \quad \text{(Eq 7)}$$

A test statistic is constructed to detect two-way or three-way interactions. Information gain-based test statistic (TIG) 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.

IGENT, introduced by Kwon et al., uses information gain to detect gene-gene interactions associated with the phenotype (Kwon et al., 2014). They used the difference between the entropy of the phenotype and the conditional entropy of the phenotype, given two genetic variants ($H(P)-H(P|(G_i, G_j))$) as information gain definition. This equality provides information about the association of the genotype interaction with the phenotype.

Su et al. proposed the Interaction Gain method similar to Fan et al.'s approach, defined as the difference between the conditional mutual information of pair of variants given the phenotype and the mutual information of the same pair ($I(G_{i,} G_j|P)- I(G_{i,} G_j)$). They also proposed a parallelization method that speeds up the interaction testing procedure.

On the other hand, unlike Fan et al., Chanda et al. suggested an approach for 3WII in which they used the phenotype as a third variable alongside two genetic variants(Chanda et al., 2008, 2009). While with Fan et al.'s approach, we can assess the correlation between variant interactions and phenotype, Chanda et al.'s method assess the interaction between variants and phenotype.

Then, Hu et al. suggested a novel approach for measuring pure three-way interactions after removing the one-way and two-way effects (T. Hu, Chen, Kiralis, Collins, et al., 2013b). However, they also used Chanda et al.'s method as a definition of information gain in which they also used the phenotype as a parameter ($IG=I(G_i, G_{j,} P)-I(G_i, P)-I(G_j-P)$).

Additionally, entropy-based methods have been performed to demonstrate the statistical interactions among the output SNPs of machine learning models. In 2014, Zieselman et al. proposed an approach that combines machine learning analysis of gene-gene interactions with large-scale functional genomics data for assessing biological relationships (Zieselman et al., 2014). This approach uses SNPs mapped to genes for machine learning analysis. The study was designed in two phases. In the first phase, QMDR is used as a feature selection method. Then, selected genes were used as an input for functional genomics analysis. Three-way and four-way interactions between genes in significantly enriched pathways were identified by applying QMDR analysis again. Finally, gene-gene interactions were assessed by performing entropy-based analyses using the visualization of the statistical epistasis networks (ViSEN) software package(T. Hu, Chen, Kiralis, & Moore, 2013). A statistically significant synergistic interaction among two SNPs located in the intergenic region of an olfactory gene cluster was found in this study. Similarly, in 2018 Dorani et al. implemented random forests (RF) and gradient boosting machine (GBM) algorithms to reveal the SNPs contributing to the colorectal cancer risk. Afterward, significant two-way and three-way interactions were determined for the SNPs identified with the RF and GBM algorithms (Dorani et al., 2018).

2.4 Alzheimer's Disease

More than 55 million people live with dementia worldwide, with over 60% living in low- and middle-income countries. This number is projected to triple by 2050 as the incidence and prevalence of dementia increase dramatically with age, and the elder people population is increasing worldwide.

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that is the most common cause of late-onset dementia. It is estimated that AD may contribute to 60-70% of the cases. AD is characterized by cognitive impairment; however, a significant heterogeneity can be observed in clinical progression. Besides, the primary pathological signs of AD are amyloid-beta (Aβ) plaques, neurofibrillary tangles, and neuronal loss. It is depicted as early-onset (EOAD) and late-onset (LOAD) based on the age of onset, and LOAD constitutes approximately 95% of cases. While EOAD is more likely familial and inherited in a Mendelian pattern and is diagnosed by the formation of AD symptoms before age 65 (Wu et al., 2022), LOAD emerges with cognitive decline symptoms such as memory impairment and loss of intellectual abilities after the age of 65. Besides, LOAD presents complex genetic inheritance, where interactions of multiple genetic variations and environmental factors affect the phenotype of patients (Reitz & Mayeux, 2014).

2.4.1 Environmental Risk Factors

The association between modifiable risk factors, such as medical comorbidities, lifestyle choices, and the reduced risk for cognitive decline is strongly evident from the population-based perspective (Baumgart et al., n.d.; Norton et al., 2014). There is increasing evidence that exposure to aluminum and air pollution may play a role in the development of AD (K.-H. Chang et al., n.d.; Flaten, n.d.; Oudin et al., n.d.; Rondeau et al., 2009). Association between cardiovascular risk factors, diabetes mellitus,

midlife hypertension, midlife obesity, and AD have been demonstrated in various studies(Joas et al., 2012; Norton et al., 2014; Ritchie et al., n.d.). Other AD-related factors include sleep disturbances, traumatic brain injury and other neurodegenerative disorders, late-life depression, education attainment, physical activity, and social engagement. However, links between many of these factors and the specific pathobiology of AD are not well established (Rabinovici, 2019).

2.4.2 Genetic Risk Factors

The heritability of LOAD is estimated to be 60 to 80% (Gatz et al., 2006) and studies on its molecular mechanisms are still ongoing. In early studies, APOE4 was established as a genetic risk factor for LOAD (Corder et al., 1993; Johanna Kuusisto, Keijo Koivisto, Kari Kervinen, Leena Mykkanen, Eeva-Liisa Helkala, Matti Vanhanen, Tuomo Hanninen, Kalevi Py6rala, Y Antero Kesaniemi, Paavo Riekkinen, 1994; Xu et al., 2021). While ε4, one of the three common alleles of APOE4, is associated with an increased risk of developing AD, the genotypes $\varepsilon_2/\varepsilon_2$ and $\varepsilon_2/\varepsilon_3$ of the other two common alleles are found to be protective. However, since the ε_4 allele represents a risk factor rather than a deterministic gene, APOE4 genotyping is not currently recommended in the clinical evaluation of patients with suspected AD (Knopman et al., 2001).

Besides, there are more than 20 LOAD candidate loci on genes such as ABCA7, BIN1, CD33, CLU, CR1, CD2AP, EPHA1, MS4A6A–MS4A4E, and PICALM identified in recent years in several GWA studies. In 2013, 11 new GWAS loci, 19 in total, were identified through a meta-analysis of 74,046 individuals within the scope of the International Genomics of Alzheimer's Project (I-GAP) (Lambert et al., 2013). In 2019, in another large-scale meta-analysis study, 29 associated loci, 16 of which overlapped with the loci identified in the I-GAP study, were identified, and nine novel significantly associated loci in this study were mapped to ADAMTS4, HESX1, CLNK, CNTNAP2, ADAM10, APH1B, KAT8, ALPK2, and AC074212.3 genes (Jansen et al., 2019). Another 13 new AD candidate loci were identified by Prokopenko et.al. in 2021(Prokopenko et al., 2021). Similarly, in another recent study, four new variants near APP, CHRNE, PRKD3/NDUFAF7, PLCG2, and two exonic variants in the SHARPIN gene were identified as associated with AD risk(Rojas et al., n.d.). Nonetheless, the functional consequences of all LOAD risk genes have not been fully discovered. Amyloid, inflammation/immune system, lipid transport and metabolism, synaptic cell functioning/endocytosis, and tau protein binding are the primary pathways clustered by many AD-associated genes.

2.4.3 Epistasis in Alzheimer’s Disease

GWA studies identify common variants that increase the disease risk. However, genes significantly associated with AD by GWA studies can explain only 25 percent of the phenotypic variance (Ridge et al., 2013). The extensive usage of next-generation sequencing (NGS) technologies give rise to studies identifying rare variants associated with AD (Prokopenko et al., 2021). Besides, polygenic risk scores are computed to predict risk based on the risk variant profile per individual. Although predictive power is not extremely high with current knowledge of heritability, it is a promising approach

for identifying subjects at high risk for developing AD and diagnosing timely those affected by the disease.

Epistasis is the other component that can elucidate the missing heritability. Various studies have identified several epistatic interactions in the last two decades. The first two epistatic interactions were reported between IL6 and IL10 and TF and HFE genes in the early 2000s (Infante et al., 2010; Warden & Smith, 2004), and both these interactions were replicated in independent samples (Combarros, Van Duijn, et al., 2009; Kauwe et al., 2010). In 2014, Ebbert et al. identified two significant interactions between CLU-MS4A4E and CD33-MS4A4E genes, and only CLU-MS4A4E was replicated by an independent sample group (M. T. Ebbert et al., 2016; M. T. W. Ebbert, Ridge, Wilson, et al., 2015). Some other studies focused on epistatic interactions associated with the disease's endophenotypes, such as amyloid deposition, brain atrophy, and grey matter density (Hohman et al., 2013; Koran et al., 2014; Zieselman et al., 2014). The replication of Gusareva et al. identified an interaction between KHDRBS2 and CRYL1 using a thorough, genome-wide screening approach, and replicable epistasis associated with AD was identified using a hypothesis-free screening approach for the first time with this study(Gusareva, Carrasquillo, et al., 2014). These studies have revealed interactions between genes previously known to be associated with AD and between genes not associated with the disease before and do not affect their own.

CHAPTER 3

MATERIALS AND METHODS

We obtained the SNPs used in this study from GWAS datasets provided by ADNI, GenADA, and NCRAD initiatives. Firstly, quality control was conducted in genotyping the datasets, and a filtering procedure was applied. After performing PLINK analysis to find out the LOAD-associated SNPs from each dataset, associated SNPs were prioritized by the RF-RF approach (performed by Onur Erdoğan). Then triplets with significant 3-way interactions were obtained by performing an entropy-based prioritization step. Subsequently, multiple testing comparisons were performed, and then obtained triplets were classified as 'risk' or 'protective' in terms of LOAD by checking the value of information gained between the affected and general population. Lastly, variants of the obtained triplets were annotated, and functional enrichment analyses were performed for those variants (Figure 5).

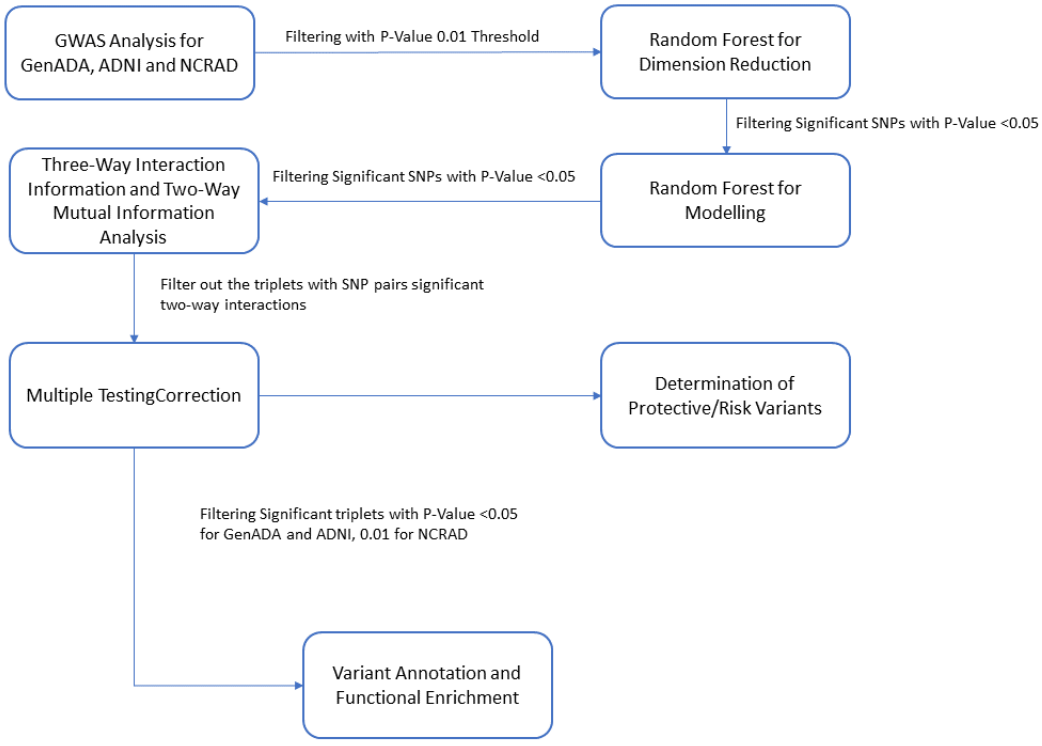


Figure 5 Overview of the Methodology

3.1 Data

High throughput sequencing technologies quickly scan entire genomes of large numbers of subjects to find genetic variants associated with a trait or disease. This advantage of HTS enables the design of large-scale research studies and generates more comprehensive information about the genomic signatures of various complex diseases.

Alzheimer's Disease Neuroimaging Initiative (ADNI), GenADA, and the National Centralized Repository for Alzheimer's Disease and Related Dementias (NCRAD) are three different initiatives that collect data for studies focusing on the etiology, early detection, disease progression, and therapeutic development of AD. This study obtains three high-dimensional datasets from these initiatives via dbGaP control access (Filippini et al., 2009; H. 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. 210 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 (**Table 1**).

Table 1 Overview of data used in this study

	Number of Cases	**Number of Controls**
GenADA (phs000168.v1.p1)	798	777
ADNI (adni-info.org)	344	210
NCRAD (phs000219.v1.p1)	1289	1310

3.1.1 Data Preprocessing

Irrelevant and redundant information, or noisy and unreliable data, can produce misleading results. In order to ensure or enhance performance, the data preprocessing step, which comprises manipulation and dropping of data operations, is performed before modeling. Accordingly, quality control was conducted in terms of genotyping in the datasets used within the scope of this study. Autosomal SNPs with Hardy Weinberg Equilibrium (HWE) <5e-07 are filtered out with MAF<0.05 and call rate<0.99. The remaining SNPs with a call rate <0.95 and no valid map location were also excluded from the genotyping data.

3.2 PLINK Analysis

GWAS analysis is done for the initial dimension reduction to discover statistically significant SNPs for building a LOAD model from each different dataset. PLINK is an open-source C/C++ tool developed for GWAS. PLINK enables rapid manipulation and analysis of datasets consisting of up to millions of markers genotyped for thousands of individuals (C. C. Chang et al., 2015; Purcell et al., 2007). In order to perform a fundamental case/control association analysis for a disease trait, allele frequencies between cases and controls are compared using a chi-square test.

In this study, PLINK analysis is performed using the "--assoc" function for identifying the independent statistical significance of variants related to the LOAD, and SNPs were filtered using a 0.05 basic allelic test chi-square p-value threshold. Multiple testing correction was not considered since a further SNP prioritization analysis would be performed using machine learning algorithms. PLINK results were used for filtering and reducing redundant SNPs that were not directly related to the disease.

3.3 SNP Prioritization with RF-RF Approach

RFs are collections of decision trees that build an ensemble model for classification and regression(Breiman, 2001). Random forests are constituted by generating random vectors with the same distribution that govern each tree's growth in the ensemble. After generating many trees, the model's prediction is determined by voting for the most popular class among those trees. In this process, the importance of a feature could be calculated according to its ability to improve the purity of the node by each tree and then averaged for all trees. In other words, the features that improve the purity lead to more significant information gain.

Accordingly, after filtering with PLINK, SNPs found to be significantly associated with LOAD were used as input for multistep RF modeling. We performed the initial RF for dimension reduction using feature importance and the second RF step as the modeling algorithm. RF was implemented using the RANGER package in R (Breiman, 2001).

Overfitting occurs when a model fits precisely against its training data and cannot perform accurately for unseen data. A 5-fold cross-validation (CV) resampling method was applied to avoid overfitting by splitting data into 5 subsets. Then the model was trained on four subsets and tested on the remaining one. This process was performed iteratively until each group was held as the test group, then the average of the models was used as the resulting model.

In the modeling phase, tuning was performed on "mtry", which refers to the number of variables to split at in each node possibly, and "ntree", which indicates the number of trees. Ntree parameter values up to 1000 were used for all datasets and "mtry" values were selected to create a manual grid. For GenADA, the model was tuned with mtry values 2, 5, 10, 20, 41, 83, 166. The mtry parameter values 2, 4, 9, 19, 39, 78, 157 were used for the ADNI dataset, and 5, 10, 20, 41, 83, 166, and 333 were used for the NCRAD dataset for RF model tuning. This grid was created with the SNP number's

square root and the square root's fold. RF tuning split rule was selected to be the "gini index." Each decision tree in the forest was created as a tree of maximum size. Importance and importance p-values were also calculated, and after the first RF, features with an importance value smaller than 0.05 were used for the modeling step. This multimodal approach was implemented for each dataset for prioritization.

3.4 Entropy-Based Prioritization

We aim to minimize the model by reducing the level of complex interactions revealed by LOAD-RF-RF. We focused on non-linear three-way interactions in this study. As mentioned above, several entropy-based approaches were suggested for pairwise, third-order, and high-order interactions for different study designs, such as family-based, case-only, and case-control. Total correlation information (TCI) and 3WII are the two quantities used for detecting three-way interactions. While TCI refers to the amount of information common to all three attributes, 3WII implies the total amount of information common to all three attributes but not present in any subset, which is in line with the purpose of this study. These definitions have been modified differently by different groups. Fan et al. express the effects of interaction on a phenotype by the difference of those interactions of variants between case and control groups (Eq 8).

$$IIG=3WII(X,Y,Z)_{cases}-3WII(X,Y,Z) \quad \text{(Eq 8)}$$

Chanda et al. used the phenotype as the third variable to explain the TCI as 'the information that cannot be obtained without observing all variables and the phenotype at the same time'. In order to avoid false positives, Hu et al. proposed another information gain model based on Chanda et al.'s, which subtracts all lower order effects from the total IG, including the main effects of the three attributes and all pairwise synergies between them (T. Hu, Chen, Kiralis, Collins, et al., 2013b) (Eq 9).

$$IG_{strict}(X,Y,Z,P)=I(X,Y,Z,P)-\max\{I(X,Y,P),0\}-\max\{I(X,Z,P),0\}-\max\{I(Y,Z,P),0\}-I(X,P)-I(Y,P)-I(Z,P) \quad \text{(Eq 9)}$$

Nevertheless, Chanda et al.'s IG definition and the other methods developed based on this definition assume that variants interact with the phenotype and interact with variants among themselves. In this study, our goal is to demonstrate the effect of the interaction of genetic variants on LOAD. Therefore we used the IG definition of Fan et al. to assess the effect of three-way interactions. Individual steps of the workflow followed are listed below:

1- The SNPs prioritized by the LOAD-RF-RF model are filtered from BED files and divided into case-control groups using the following P-LINK commands. Firstly bed files were converted to ped files:

.\plink –bfile bedFileName –recode12 –tab –out outPutPedFileName

2- The SNPs prioritized by the LOAD-RF-RF model are filtered from generated ped files:

.\plink --file outPutPedFileName --out filteredSNPs --recode12 --extract listOfPrioritizedSNPs

3- Filtered SNP ped files divided into case and control groups:

.\plink --file filteredSNPs --filter-cases --recode12 --out caseSNPs

.\plink --file filteredSNPs --filter-controls --recode12 --out controlSNPs

4- In the last step of PLINK analysis, members belonging to each genotype are listed for case and control groups separately:

.\plink --file caseSNPs --list --out caseSNPsList

.\plink --file controlSNPs --list –out controlSNPsList

5- The filtered SNPs' genotype frequencies are calculated using a custom Python script (**Figure6; Appendix A.1.**). As reviewed in the previous section, these frequencies are used as parameters for 3-way interaction information gain and two-way mutual information gain functions for identifying SNPs that explain the susceptibility of LOAD. These two functions are implemented using custom R scripts adapted from Fan R. (Ruzong Fan, n.d.) (**Appendix A.2. , A.3 and A.4**).

IIG (Eq 7) values, the difference of 3WII of variants between case and control groups, for each triplet were calculated by the implementation of 3-way interaction information gain function. Then, we calculated the test statistic by dividing the IIG of each dataset's prioritized SNPs 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 as mentioned in Literature Review section. Significantly different interactions are identified by using p-values assigned in these test statistics.

Then, two-way mutual information gain test statistics are calculated for the triplets' variants, which are found to have significant interactions in the previous step. 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.

3.5 Multiple Test Correction

In cases where we perform multiple simultaneous statistical tests, we encounter multiple comparison problems. As the number of tests applied increases, the probability of obtaining false-positive results also increases. Validation tests based on multiple testing corrections and resampling techniques such as permutation-based tests have been performed to counteract the multiple comparisons problem. Multiple testing correction techniques such as Bonferroni (Haynes, 2013) and The Benjamini and Hochberg (B&H) (Hochberg, 1995) corrections make the statistical tests more stringent by adjusting the p-values. The total number of observations in a population

sample is resampled to build an empirical estimate of the null distribution from which the test statistic has been drawn by performing the permutation test (Belmonte & Yurgelun-Todd, 2001).

```
#This function remove the family IDs and recode the genotypes in case control genotyping files

function convertGtypeCoding(SNP list file path):
   open SNP file
   read SNP file line by line and generate a SNP list
   for the element in SNP list
      Remove the familiy IDs from the element
   Then convert the genotype values
   for the element in SNP list
      if genotype=='11'
         genotype='0'
      if genotype='12' or '21'
         genotype='1'
      if genotype'22'
         genotype='2'

   return SNP List

#This function generates a csv file with frequencies of triplets or pairs.
function convertgtypeII(SNP List,interactionType (3 way or 2 way),outputPath (path for the
output)):

   if interactionType is " 3WII"
      for i is 0 to (length of SNP List)-11 increase by 4
         for j is i+4 to (length of SNP List)-7 increase by 4
            for k is j+4 (length of SNP List)-3 increase by 4
              for l in i to i+4
                  for m is j to j+4
                     for n is k to k+4
                        if the remainder is not equal to 3 for l/4, m/4 and n/4
                           Generate sets of individuals for each three SNPs at three genotype levels
                           Intersect those sets to calculate the frequency of the relevant triplet
                           Assign this frequency to the relevant column
             if the remainder is equal to 3 for l/4
               Assing the name of the SNPs to the relevant columns
    else
      for j is 0 to (length of SNP List)-7 increase by 4
         for k is j+4 to (length of SNP List)-3 increase by 4
            for m is j to j+4:
               for n is k to k+4:
                  if the remainder is not equal to 3 for m/4 and n/4
                     Generate sets of individuals for each two SNPs at two genotype levels
                     Intersect those sets to calculate the frequency of the relevant pairs
                     Assign this frequency to the relevant column
            if the remainder is 3 for m/4:
               Assing the name of the SNPs to the relevant columns
   Save the resulting dataframe to a CSV file
   Return
```

Figure 6 Pseudocode of the algorithm developed for genotype frequency calculation

This study validates significant interactions for multiple comparisons using permutation testing (Camargo et al., 2008) by performing the following steps:

1. Disease status labels are randomly shuffled

2. Information gain-based test statistic is calculated in each iteration for performing 1000-fold permutation testing for GenADA and ADNI datasets. 10000-fold permutation testing is performed for the NCRAD dataset to accommodate prioritized SNPs and triplets, which had more significant triplets than the other datasets.

3. Then the ratio of test statistics greater than the observed test statistic is calculated to assign a p-value to the permutation testing (**Appendix A.5**).

Triplets with a p-value greater than 0.05 and 0.01 are filtered out for GenADA, ADNI, and NCRAD. As the variants are previously prioritized by the multistep random forest model, revealing the interactions, we defined the p-value thresholds by considering the Type II errors more than Type I errors. Besides, we adjusted different thresholds based on the number of triplets in each dataset separately.

3.6 Determining Risk and Protective Variants for LOAD

As described above, information gain is the difference between the affected and the general populations' mutual information. If the disease is not associated with the markers, the information gain equals 0. If the mutual information in the affected population is greater than the general population, information gain is greater than zero; otherwise, it is smaller than zero. Accordingly, we checked the triplets' mutual information values, which show significantly different three-way interactions. Then we identified the triplets with positive information gain value as ‘Risk Variants’ and the triplets with negative information gain value as ‘Protective Variants’.

3.7 Variant Annotation

The process of assigning functional information to variants is called variant annotation. This information could involve sequence conservation measures, predictions about a variant's effect on protein structure and function, and genomic mapping. SNPNexus (Chelala et al., 2009; Dayem Ullah et al., 2012, 2013, 2018) and SNiPA (Arnold et al., 2015) tools have been used to annotate the variants in the filtered triplets.

Genomic mapping, variant annotation, gene/protein consequences, and phenotype/disease association information have been obtained from these tools. We selected the GRCh37 human reference genome assembly for both tools for our query. Gene/protein consequences were obtained from NCBI RefSeq (Pruitt et al., 2014) and Ensembl (Hubbard et al., 2007) datasets in the SNPNexus tool. Ensembl was selected in the SNiPA tool for the same purpose. SNiPA linkage disequilibrium data and allele frequencies were computed for 1000 Genomes Phase 3 v5 dataset in the European

population, and we obtained the Non-Finnish European gnomAD frequencies from SNPNexus.

While phenotype/disease associations were obtained from the Genetic Association of Complex Diseases and Disorders (GAD) and ClinVar databases in SNPNexus, this information was obtained from OMIM, Orphanet, NHGRI GWAS Catalog, HGMD, dbGap, and ClinVar datasets.

3.8 Functional Enrichment

Functional enrichment analysis was performed for the triplets' variants with significant three-way interactions. Firstly, we collected the triplets with all variants mapped to a gene. Then we queried those triple genes separately in the Reactome database, open-source, open access, manually curated, and peer-reviewed pathway database (Fabregat et al., 2018) to determine if all three genes are found in pathways associated with AD.

Then, GO Molecular Function, GO Cellular Component, GO Biological Process, and Reactome pathways are used using the g: GOSt component of the g: Profiler tool (Raudvere et al., 2019) for the variants reported in each triplet for all datasets. All analyses have been done with default attributions with a significance threshold of 0.05. The p-value of the enrichment of pathways has been computed using Fisher's exact test, and the Bonferroni correction method has been used for multiple testing corrections.

Then, EnrichmentMap (Merico et al., n.d.), a plugin for the Cytoscape tool (P. Shannon et al., n.d.), has been used to create networks from Gene Ontology annotations and Reactome pathways in order to derive overrepresented functional groups from functional annotation. All analyses have been done with a p-value of 0.05, FDR q-value cutoff of 0.01, and edge similarity cutoff of 0.3(Jaccard metric).

CHAPTER 4

RESULTS

This study used an entropy-based approach to detect the third-order interactions in PLINK-RF-RF models from three different LOAD datasets. Firstly, LOAD-associated variants were determined by performing a PLINK analysis in three different datasets. Then a two-step RF-RF model is conducted for variant prioritization. Variants obtained in this analysis (performed by Onur Erdoğan) were used in further analysis.

32 SNPs found to be associated with LOAD for the ADNI dataset, 36 SNPs and 218 SNPs for GenADA and NCRAD datasets, respectively, were used to examine the three-way interactions related to the LOAD. Firstly, a three-way interaction information analysis was performed for these variants to determine significant three-way interactions. Afterward, a two-way mutual information analysis was done. Then, triplets involving variant pairs with significant two-way interactions were filtered out since we focused on the information common to all three attributes that the two-way mutual information cannot explain. Multiple testing correction was applied to validate the significant three-way interactions. Subsequently, the selected triplets are examined based on the value of information gained for determining the risk and protective triplets.

The variants in triplets with significant three-way interactions and all of which were mapped to a gene, were queried in the Reactome database to determine if all three genes are found in pathways associated with AD. In the last step, Functional enrichment analysis was performed with overlapped genes, nearest downstream genes, and nearest upstream genes of all variants in all datasets.

Among SNPs prioritized by 3WII, four out of 19 SNPs from GenADA, one out of 27 from ADNI, and four out of 106 NCRAD are mapped to genes directly associated with Alzheimer's Disease.

4.1 Variants Associated with LOAD

After filtering out the SNPs that failed the quality control, PLINK association analysis was performed for each controlled accessed GWAS dataset. Significance values were calculated based on comparing allele frequencies between cases and controls. 7639 SNPs for ADNI, 3767 SNPs for GenADA, and 16404 SNPs for NCRAD with p-values smaller than 0.01 selected. We found only one common SNP for these datasets, 18 common SNPs for only NCRAD and GenADA, 206 common SNPs for only NCRAD and ADNI, and nine common SNPs for only ADNI and GenADA (**Figure 7**).

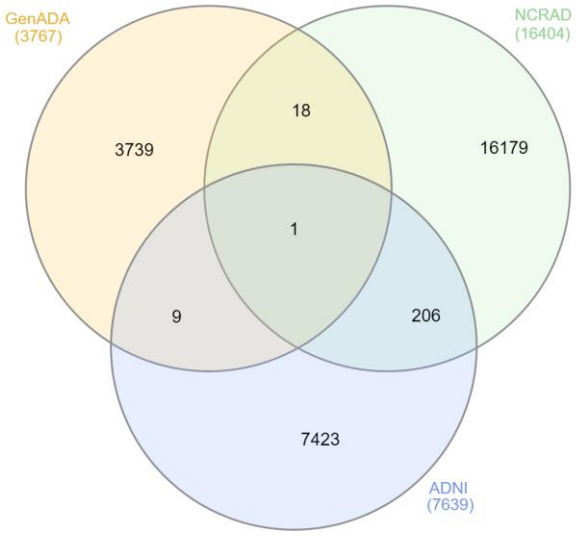


Figure 7 Venn Diagram for Number of Filtered SNPs by GWAS Analysis

4.2 SNP Prioritization by LOAD-RF-RF Model

After the PLINK association analysis, the RF model was implemented for the feature selection from each dataset. Depending on the evaluation of the model tuning with 5-fold fold cross-validation results, the best mtry and ntree parameters were determined as in Table 2.

Table 2 The best mtry and ntree parameters determined by parameter tuning

	ADNI	**GenADA**	**NCRAD**
mtry	39	2	83
ntree	50	900	1000

The permutation hypothesis test calculates the contribution of random change in the value of the variation to the accuracy rate. After 100 permutations, 390 variants from ADNI, 1740 from NCRAD, and 434 from GenADA datasets related to the disease were selected as the input set for the modeling step with second RF at a 95% confidence level (Type I error = 0.05).

In the second step with the multistep LOAD-RF-RF model, 32 SNPs are identified and selected for the disease at a 95% confidence level for the ADNI dataset. Besides, 36 SNPs and 218 SNPs for GenADA and NCRAD datasets were associated with the disease at a 99% confidence level (**Table 10 in Appendix B**) (**Figure 8)**. These

prioritized SNPs are used to examine the 3-way interactions related to the LOAD interactions.

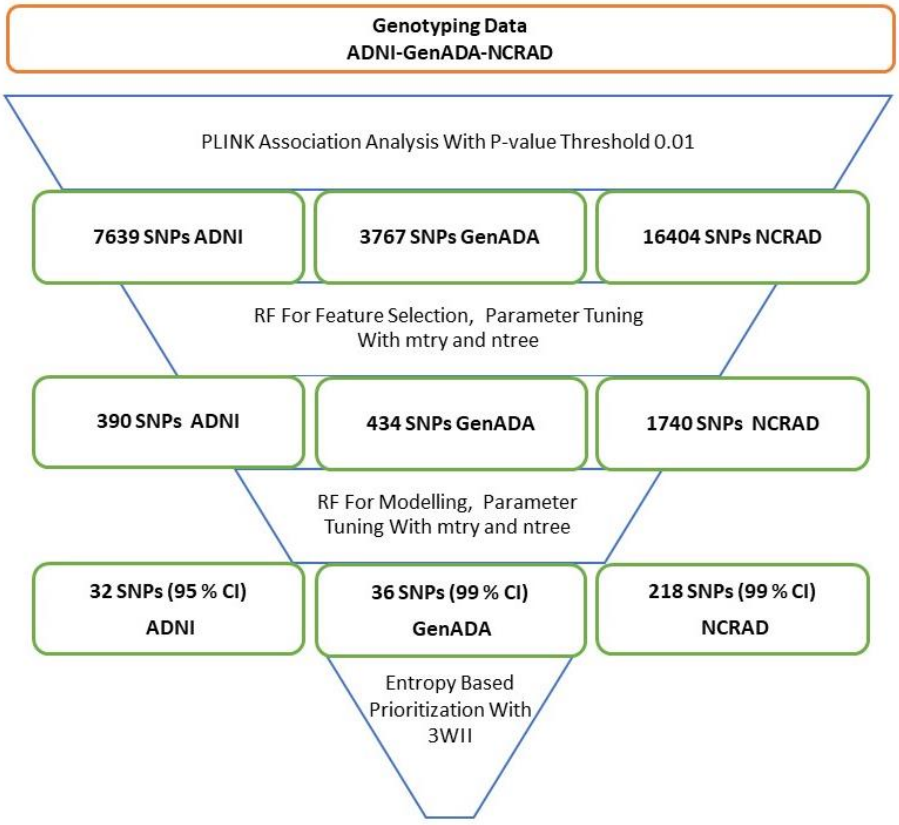


Figure 8 Workflow and Data Summary In the first step, PLINK association analysis is performed for genotyping data of three datasets. Then RF-RF is conducted for feature selection and modeling, respectively. Output variants of the PLINK-RF-RF model are prioritized by 3WII analysis

4.3 Three-Way Interaction Information (3-WII) Analysis

Prioritized SNPs from the RF-RF model of each dataset have been used to determine the significant three-way interactions. ADNI dataset has been tested for 4960 three-way interactions. GenADA and NCRAD datasets have been tested for 7140 and 1,703,016 interactions, respectively.

Firstly, the rate of prioritized genotype triplets is calculated separately in case and control groups for each dataset. Then, the difference of each variant triplet's three-way interaction (3WI) information is estimated between the case and control groups. For the GenADA dataset, nine out of 7140 triplets had significant three-way interactions. However, one of the triplets consisting of rs17067596, rs4895529, and rs16993582 is filtered out since it includes an SNP pair in strong linkage disequilibrium. Likewise,

ADNI and NCRAD datasets have 17 out of 4960 and 86 out of 1,703,016 significantly interacted triplets.

In the next step, two-way mutual information gain is calculated for the variants found in the significant triplets. 171 two-way mutual information gain is calculated for 19 unique SNPs in GenADA triplets, 325 for 26 unique SNPs in ADNI triplets, and 7750 for 125 unique NCRAD triplets.

The SNP triplets with SNP combinations with significant two-way mutual information gain are excluded. For GenADA and ADNI, no SNP pairs with significant two-way mutual information gain are found. For NCRAD, 22 triplets with SNP pairs with significant two-way mutual information gain are filtered out (**Table 11 and Table 12 in Appendix B**). After this filtering, eight significant triplets with 19 unique SNPs from GenADA, 17 with 26 unique SNPs from ADNI, and 64 significant triplets with 116 unique SNPs from the NCRAD dataset were selected. There were no common SNPs between these groups.

Lastly, for the validation of p-values assigned by test statistics, a permutation test was performed. Significant triplets have been selected using 0.05 for the p-value threshold for GenADA and ADNI datasets. Since there were more triplets for the NCRAD dataset, 0.01 was used as the p-value threshold. Accordingly, 1000 permutations for GenADA and ADNI and 10000 permutations for the NCRAD dataset were performed in this step (**Table 3**).

Table 3 Overview of the SNPs after each filtering step

	ADNI	**GenADA**	**NCRAD**
Number of prioritized SNPs by RF-RF	32	36	218
Number of triplets	4960	7140	1,703,016
Number of significantly interacted triplets	17 (26 SNPs)	8 (19 SNPs)	86 (125 SNPs)
Number of SNP pairs tested for 2WI	325	171	7750
Number of triplets with significantly interacted SNP pairs	0	0	22
Number of significantly interacted triplets after permutation testing	17 (26 SNPs)	8 (19 SNPs)	64 (116 SNPs)

4.3.1 GenADA Dataset Results

Following the workflow in **Figure 8**, eight triplets with 19 unique SNPs for GenADA (**Table 4**) had significant 3-way interaction (3WI) information. All GenADA 3WII SNPs are categorized as modifiers based on their impact with the SnpEff tool (Cingolani et al., 2012). Four SNPs are mapped to FBLN2, ADAM 10, NHSL1, and ST3GAL1 genes, previously associated with Alzheimer's Disease (Nobrega et al., 2017; Patel et al., 2019). The SNP mapped to ADAM10 is also associated with reticulate acropigmentation of Kitamura. Lastly, one variant is mapped to the RUNX1 gene associated with chronic myeloid leukemia **(Table 5).**

Table 4 Test Statistics and Permutation Testing Results for GenADA Dataset

SNP1	SNP2	SNP3	T_{IG}	P-value	Permutation p-value	Gene1	Gene2	Gene3
rs17793957	rs605928	rs9911460	8.50	0.003	0.001	FBLN2	ADAM10	NPLOC4
rs7045548	rs1795977	rs11652714	8.25	0.004	0.001	-	-	-
rs1879019	rs17081694	rs605928	6.65	0.009	0.002	-	-	ADAM10
rs1608169	rs11862388	rs16993582	7.14	0.007	0.003	-	-	RUNX1
rs4895529	rs9314604	rs17081694	8.47	0.003	0.006	NHSL1	ANGPT2 / MCPH1	-
rs17067596	rs9314604	rs17081694	8.46	0.003	0.008	NHSL1	-	
rs10050568	rs2978012	rs6098412	7.76	0.005	0.015	SPOCK1	ST3GAL1	-
rs1879019	rs1519959	rs136687	8.48	0.003	0.036	-	-	PHF21B

Table 5 The SNPs in GenADA Dataset Mapped to a Gene Associated with a Disease

rsID	Chr.	Pos	Gene	Phenotype
rs17067596	6	138,767,012	NHSL1	Alzheimer's Disease
rs2978012	8	134,539,196	ST3GAL1	Alzheimer's Disease
rs4895529	6	138,770,275	NHSL1	Alzheimer's Disease
r605928	15	59,046,163	ADAM10	Alzheimer's Disease/ Reticulate acropigmentation of Kitamura
rs17793957	3	13,649,920	FBLN2	Alzheimer Diesase
rs9314604	8	6,411,499	MCPH1/ ANGPT2	Microcephaly
rs16993582	21	37,110,209	RUNX1	Platelet disorder, familial, with associated myeloid malignancy

4.3.2 ADNI Dataset Results

Seventeen triplets with 27 unique SNPs showed significant 3-way interaction (3WI) information after permutation testing are identified for the ADNI Dataset **(Table 6).** One variant was mapped to FERMT2, a risk factor for Alzheimer's disease. Four SNPs mapped to SYCP2L, VAV2, SEPSECS, and TMPRSS15 are associated with age-related hearing impairment, multiple sclerosis, pontocerebellar hypoplasia type 2, enterokinase deficiency, respectively. PLAGL1 is associated with transient neonatal diabetes mellitus, paternal uniparental disomy of chromosome 6, and NPSR1 is associated with asthma-related traits **(Table 7).** Although some triplets have common SNP pairs, none of their elements are in the same LD. All these SNPs are also categorized as modifiers.

4.3.3 NCRAD Dataset Results

Fifty-two triplets with 106 unique SNPs were significant 3WI for NCRAD (**Table 13 in Appendix B**). Only two triplets shared an SNP pair. The third SNPs of the triplets, rs12663008, and rs17830067, were found to be in the same LD region. So, one of these triplets could be used as a representative. There were also three other SNPs in the same LD region, which do not interact with other common SNPs.

Variants mapped to PVRL2, TOMM40, LCMT1, and RAB3GAP1 genes were previously associated with Alzheimer's disease. Besides, another variant mapped to HNRNPA1 is associated with amyotrophic lateral sclerosis and inclusion body myopathy with Paget disease of bone and frontotemporal dementia. Other associated diseases can be seen in Table 8.

4.3.4 Protective-Risk Triplets Analysis

As noted previously, the interaction information gain IIG value represents the difference between 3-way interaction information of disease and control groups. A positive IIG represents the gain of 3WII in the presence of a disease. In contrast, a negative IIG represents the gain of 3WII in the general population versus the affected population.

In GenADA dataset for only two triplets, rs1608169; rs11862388; rs16993582 and rs10050568; rs2978012; rs6098412, the IIG was positive, and it was negative for the rest of the triplets. Additionally, in ADNI dataset only two triplets (rs6705017;rs10017010;rs557098 and rs685677;rs7157639;rs2824808) had positive IIG **(Table 9)**. The SNPs are categorized as modifiers as in the other dataset groups. However, unlike the other datasets, IIG values are mostly positive in NCRAD (**Table 13 in Appendix B**).

Table 6 Test Statistics and Permutation Testing Results for ADNI Dataset

SNP1	SNP2	SNP3	T_{IG}	P-value	Permutation p-value	Gene1	Gene2	Gene3
rs9366664	rs3780792	rs1150360	8.53	0.003	0.001	SYCP2L	VAV2	FAM76B
rs6705017	rs10017010	rs557098	7.51	0.006	0.001	UGGT1	PI4K2B	ALDH3B1
rs11749731	rs3780792	rs7157639	6.74	0.009	0.001	NDFIP1	VAV2	FERMT2
rs1023276	rs324389	rs2824808	8.81	0.002	0.002	-	NPSR1-AS1	TMPRSS15
rs6751810	rs4561856	rs1023276	7.92	0.004	0.002	-	-	-
rs4561856	rs4409091	rs2633466	7.61	0.005	0.002	-	-	-
rs4561856	rs10807701	rs2824808	7.55	0.005	0.002	-	TPST1	TMPRSS15
rs7091014	rs11006011	rs2633466	7.47	0.006	0.002	-	-	-
rs6856771	rs7157639	rs2824808	8.18	0.004	0.003	-	FERMT2	TMPRSS15
rs9366664	rs10960174	rs1150360	7.69	0.005	0.003	SYCP2L	-	FAM76B
rs11749731	rs10807701	rs7157639	7.42	0.006	0.003	NDFIP1	TPST1	FERMT2
rs6705017	rs11006011	rs2633466	7.77	0.005	0.005	UGGT1	-	-
rs9313264	rs12056012	rs2633466	7.69	0.005	0.005	-	-	-
rs6705017	rs2633466	rs462074	7.38	0.006	0.006	UGGT1	-	-
rs10017010	rs9313264	rs2207851	8.65	0.003	0.008	PI4K2B	-	PLAGL1
rs4561856	rs9896368	rs2824808	8.88	0.002	0.009	-	MMP28	TMPRSS15
rs4561856	rs7157639	rs717840	7.784	0.005	0.022	-	FERMT2	CDH13

Table 7 The SNPs in ADNI Dataset Mapped to a Gene Associated with a Disease

rsID	Chr.	Pos	Gene	Phenotype
rs7157639	14	53,388,161	FERMT2	Alzheimer's Disease/Hereditary Spastic Paraplegia
rs9366664	6	10,892,499	SYCP2L	Age-related hearing impairment
rs10017010	4	25,188,718	SEPSECS	Pontocerebellar hypoplasia type 2es
rs2207851	6	144,337,886	PLAGL1	Transient neonatal diabetes mellitus / Paternal uniparental disomy of chromosome 6
rs2824808	21	19,775,220	TMPRSS15	Enterokinase deficiency
rs324389	7	34,777,714	NPSR1	Asthma-Related Traits
rs3780792	9	136,835,343	VAV2	Multiple Sclerosis

Table 8 The SNPs in ADNI Dataset Mapped to a Gene Associated with a Disease

rsID	Chr.	Pos	Gene	Phenotype
rs2075650	19	45,395,619	TOMM40	Alzheimer's Disease
rs6859	19	45,382,034	PVRL2	Alzheimer's Disease
rs10445686	2	135,893,372	RAB3GAP1	Alzheimer's Disease/Leiomyoma, Uterine
rs11645986	16	25,127,645	LCMT1	Alzheimer's Disease
rs1920045	12	54,670,398	HNRNPA1	Amyotrophic lateral sclerosis/ inclusion body myopathy with Paget disease of bone and frontotemporal dementia
rs7181139	15	77,977,667	LINGO1	Mental retardation / Essential tremor & Parkinson's
rs3775162	4	72,397,710	SLC4A4	Proximal renal tubular acidosis with ocular abnormalities
rs4076290	2	1,378,969	TPO	Thyroid dyshormonogenesis
rs1560964	15	33,766,809	RYR3	Epileptic encephalopathy
rs1530498	5	13,902,220	DNAH5	Primary ciliary dyskinesia
rs17576289	3	45,458,733	LARS2	Perrault syndrome
rs17742907	22	18,890,615	DGCR6	Velocardiofacial syndrome
rs2108392	5	130,533,828	LYRM7	Mitochondrial Complex iii Deficiency
rs2432762	6	5,435,756	FARS2	Combined oxidative phosphorylation defect type 14
rs3785113	16	68,369,213	PRMT7	Pseudohypoparathyroidism-like disorder
rs3888795	18	11,863,899	GNAL	Dystonia
rs991974	6	70,481,267	LMBRD1	Methylmalonic acidemia with homocystinuria

Table 9 Risk/Protective Triplets Distribution Among Datasets

	Number of Risk Triplets	Number of Protective Triplets
GenADA	2	6
ADNI	2	15
NCRAD	38	14

4.4 Functional Enrichment

We queried the triplets variants with significant three-way interactions, all of which were mapped to a gene in the Reactome database to determine if all three genes are found in pathways associated with AD. Variants of one triplet for GenADA, four triplets for ADNI, and 15 triplets for NCRAD were examined.

No 3WII triplets had variants enriched in the same pathway. ADAM10 and FBLN2 genes were significantly enriched in the extracellular matrix organization, to which two GenADA variants are mapped. The third variant in the identical triplet was mapped to the NPLOC4 gene, enriched in the post-translational protein modification pathway in ADAM10. VAV2 and FERMT2 genes with two ADNI variants mapped were significantly enriched in five different pathways. PI4K2B and ALDH3B1, another gene pair that two of the ADNI variants are mapped to, were significantly enriched in the metabolism of the lipids pathway (**Table 14 in Appendix B**). No triplets in NCRAD dataset involve triplets with variants pairs enriched in the same pathway.

Functional enrichment analysis was conducted for the gene set, which combined the genes mapped in all three datasets. These variants are annotated with SNPNexus. The functional enrichment analysis involves overlapped nearest upstream and downstream genes (**Table 15 in Appendix B**).

Firstly, GO Molecular Function, GO Cellular Component, GO Biological Process, and Reactome pathways are obtained for the resulting dataset. Calcium ion binding, extracellular matrix, external encapsulating structure, and RUNX1 regulates estrogen receptor-mediated transcription pathways are significantly enriched.

Then, functional enrichment networks are created by Enrichment Map. The common functions of extracellular matrix and external encapsulating structure pathways are observed on the same network (**Figure 9**).

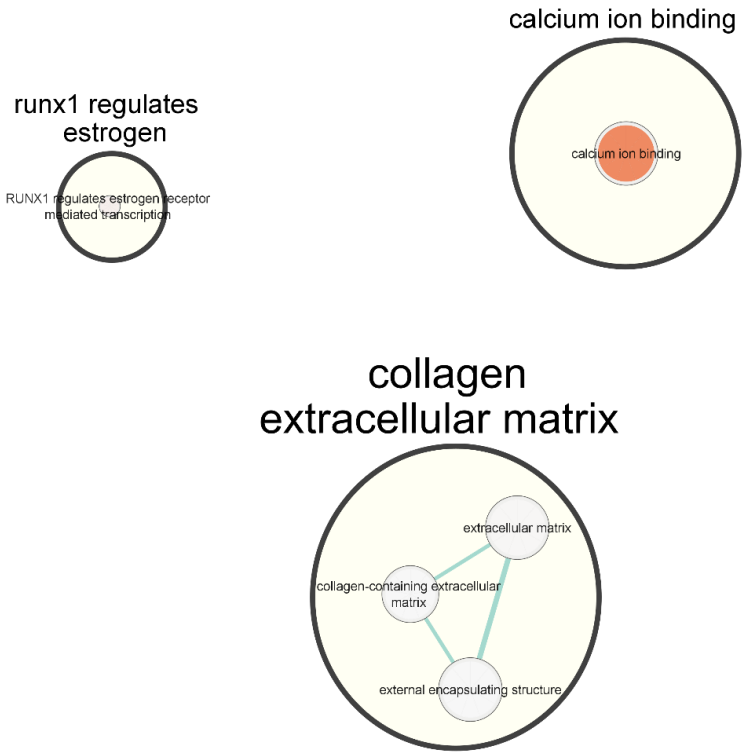


Figure 9 Functional enrichment network created by Enrichment Map

CHAPTER 5

DISCUSSION

Similar to most complex diseases, unveiling the missing heritability is a challenge in AD. Various studies have revealed epistatic relationships and offer a potential solution to this problem. In earlier studies, interactions between genes in pathways associated with AD are examined (Infante et al., 2010; Robson et al., 2004), while recent studies focus on the interactions between GWAS-identified LOAD genes (M. T. W. Ebbert, Ridge, Wilson, et al., 2015). Also, few studies reported epistatic effects associated with AD's endophenotype or intermediate traits such as amyloid deposition and brain atrophy (Hohman et al., 2013; Meda et al., 2013; Zieselman et al., 2014).

The limitations of epistatic interaction studies are the extensive multiple comparisons and low power, leading to high error rates of type 1 and type 2. Besides, the high cost of computational resources required for an exhaustive search is another constraint that challenges the interactions' detection. These limitations also avoid producing replicated results.

We have proposed an integrative approach where multistep LOAD-RF-RF is followed by 3WII entropy analysis to overcome these limitations. The LOAD-RF-RF model is developed using PLINK and RF-RF workflow for three LOAD datasets from different datasets. Then, triplets with significant 3-way interactions are identified among the prioritized SNPs. RF is a powerful tool to address the missing heritability problem, revealing high-dimensional interactions between variants. Applying the 3 WII methods after the RF let us focus on the non-linear high-dimensional interactions found with RF. Prioritization of SNPs through multistep RF before 3-WII analysis reduces the need for extensive computational resources for exhaustive analysis. Since we examined the interactions between disease-associated variants, we ignored the variants that only affect the phenotype by interaction as a part of the limitations of this study.

For all three datasets, the number of unique SNPs is reduced after the 3WII analysis, promising for applying 3-way interaction information analysis for model minimization. Zieselman et al. and Dorani et al. performed similar integrated entropy-based approaches to demonstrate the statistical interactions among the output SNPs of machine learning models (Dorani et al., 2018; Zieselman et al., 2014). In Zieselman's model, machine

learning analysis of gene-gene interactions was combined with large-scale functional genomics data for assessing biological relationships. While they used SNPs mapped to genes for machine learning analysis, we did not perform a biological-knowledge-driven approach. In addition, they applied this method to grey matter density as an endophenotype for late-onset Alzheimer's disease rather than the disease itself. In Dorani et al.'s study, after performing data preprocessing and feature selection steps, random forests (RF) and gradient boosting machine (GBM) algorithms were implemented separately to reveal the SNPs contributing to the colorectal cancer risk. Common SNPs prioritized by both algorithms were used for further examination. In both studies, significant two-way and three-way interactions of prioritized SNPs were determined using the information gain approach proposed by Hu et al. (T. Hu, Chen, Kiralis, Collins, et al., 2013b). In those studies this step was applied only to demonstrate the existing interactions instead of being used for model minimization as in our study. Besides, Hu et al.'s method assessed the interaction between variants and phenotype rather than the correlation between variant interactions and phenotype as we did use Fan et al.'s approach.

Implementing the 3WII step allowed us to recognize the informative variant triplets among the LOAD-RF-RF prioritized SNPs, which could be informative candidate biomarkers. On the other side, Hohman et al. and Gusareva et al. also performed an exhaustive genome-wide interaction information analysis(Gusareva, Carrasquillo, et al., 2014; Hohman, Bush, Jiang, Brown-Gentry, Torstenson, Dudek, Mukherjee, Naj, Kunkle, Ritchie, Martin, Schellenberg, Mayeux, Farrer, Pericak-Vance, Haines, Thornton-Wells, et al., 2016). Hohman et al. used a biological knowledge-driven approach to assess the interactions, while we did not use prior knowledge to reveal the significant interactions. Apart from that, both studies focus on 2-way rather than 3-way interactions.

IIGs are calculated for triplets dependent on the disease based on test statistics (TIG). Only a few IIGs are positive for GenADA and ADNI datasets. IIGs for 38 triplets out of 52 significant triplets have positive IIG. In various studies, while positive gain of information is referred to as synergy between variables, loss of information indicates the redundancy between them (Anastassiou, 2007; T. Hu, Chen, Kiralis, Collins, et al., 2013b; Moore & Hu, 2015). In those studies, the phenotype was treated as the third variable in the definition of IG. This definition of IG enables us to examine the interaction between variants and phenotype. However, in this study, we assess the correlation between variant interactions and phenotype using Fan et al.'s approach. Therefore, in our study, positive IIG demonstrates that the disease group's 3-way interaction information is greater than the control group. Accordingly, the triplets with positive IIG are considered a risk factor for LOAD, while the IIG negative triplets should be further investigated as protective markers.

3WII analysis is performed to prioritize SNPs by the LOAD-RF-RF model for all datasets. The number of prioritized SNPs differs between different datasets, so the number of

significant triplets differed. Also, we observed a higher number of significant interactions with the NCRAD dataset, which was the most extensive dataset. Although the NCRAD SNPs are reduced with the LOAD-RF-RF model, similar to the other two datasets before the 3WII analysis, the high number of SNPs prioritized leads to a higher number of significant triplets identified.

Complex diseases are also polygenic since multiple genes and environmental interactions contribute to the phenotype. Here we have observed that most common disease-risk variants map to noncoding sequences, known as modifiers. Also, in literature, complex disease genes overlap with genes related to Mendelian disorders. Our observations parallel the literature as some significant interactions revealed for LOAD are SNPs mapped to Mendelian disorder genes. Up to ten variants are mapped to Alzheimer's disease genes within the prioritized SNPs in all datasets. Five SNPs are associated with neurological disorders like multiple sclerosis, epileptic encephalopathy, and Parkinson's disease. 3-way interaction information describes the amount of information gained for all variables but does not present any subset alone. Therefore, although prioritized SNPs are not previously annotated as LOAD-associated SNPs, their interaction could still inform the LOAD risk. New clinical studies can validate the prioritized SNPs and triplets' association with the LOAD.

Functional enrichment analysis of 3WI variants from three different LOAD datasets at the gene level showed enrichment of collagen-containing extracellular matrix (ECM) and external encapsulating structure, and estrogen receptor (ER) mediated transcription pathways. The ECM supports the basement membranes and microcircular environment of the tissues. Several recent studies have reported the link between ECM changes and aging and neurodegenerative diseases (Damodarasamy et al., 2020; Ma et al., 2020). Even though the exact molecular impact of changes in the ECM proteins during AD development is still under investigation, its effects on synaptic transmission, amyloid-β-plaque generation and degradation, Tau-protein production, oxidative-stress response, and inflammatory response have been reviewed (Wilhelm Steinbusch et al., 2021).

Additionally, the role of ERs in cognition and memory has been investigated, and it has been demonstrated that ERs act as a neuroprotectant by modulating several neuroprotective pathways, including immune response, neurogenesis, glial cell functions, and response to excitotoxicity. As the female predominance in developing AD suggests the involvement of gender-specific factor(s), the potential role of ER alpha in AD pathogenesis has been explored in many studies (Maioli et al., 2021; Wang et al., 2016).

These functional-level observations support the proposed entropy-based post-GWAS analysis, LOAD-RF-RF followed by 3WII, as the prioritized variants and genes show association with LOAD and provide insights into early LOAD pathogenesis. Nevertheless, although the variants of the prioritized triplets are not enriched in AD-associated functional pathways, their interactions can still imply the LOAD risk.

CHAPTER 6

CONCLUSION

Random forest and entropy-based methods reveal non-linear genetic and environmental factors contributing to complex traits. The proposed workflow in this study demonstrates an efficient framework for revealing the complex interactions that contribute significantly as genetic factors for LOAD. 3WII is used as a model minimization method and determines the significant 3-way interactions between the prioritized SNPs by PLINK-RF-RF.

In this study:

- The proposed integrated method was applied to three different LOAD GWAS datasets.
- We used the three-way interaction information method proposed by Fan et al. and integrated this method with a two-step RF-based model for the first time to examine the correlation between the interaction of variants and the phenotype.
- The correlation between the three-way interactions rather than two-way of the potential LOAD-associated variants and their protective or risk status for LOAD was examined for the first time.

In the future, the SNPs detected by this optimized in-silico model could be examined in a clinical context to decide if the resulting triplets have predictive power for early or differential LOAD diagnosis.

This framework is a promising approach for post-GWAS analysis of other complex genetic disorders. The method can be improved by applying it to the GWAS data obtained from large-scale data repositories such as UK BioBank (Sudlow et al., 2015) and Estonian Biobank (Leitsalu et al., 2015), and the FinnGen[2] study. Besides, the proposed method in

[2] https://finngen.gitbook.io/documentation/

this study could be improved by using larger datasets of other complex diseases and LOAD, and it could also be modified by integrating other machine learning and entropy-based interaction method.

REFERENCES

Anastassiou, D. (2007). *Computational analysis of the synergy among multiple interacting genes*. *83*. https://doi.org/10.1038/msb4100124

Arnold, M., Raffler, J., Pfeufer, A., Suhre, K., & Kastenmüller, G. (2015). SNiPA: An interactive, genetic variant-centered annotation browser. *Bioinformatics*, *31*(8), 1334–1336. https://doi.org/10.1093/bioinformatics/btu779

Barreiro, L. B., Laval, G., Ne Quach, H., Patin, E., & Quintana-Murci, L. (2008). *Natural selection has driven population differentiation in modern humans*. https://doi.org/10.1038/ng.78

Baumgart, M., Snyder, H. M., Carrillo, M. C., Fazio, S., Kim, H., & Johns, H. (n.d.). *Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective*. https://doi.org/10.1016/j.jalz.2015.05.016

Belmonte, M., & Yurgelun-Todd, D. (2001). Permutation testing made practical for functional magnetic resonance image analysis. *IEEE Transactions on Medical Imaging*, *20*(3), 243–248. https://doi.org/10.1109/42.918475

Botta, V., Louppe, G., Geurts, P., & Wehenkel, L. (2014). Exploiting SNP correlations within random forest for genome-wide association studies. *PloS One*, *9*(4), e93379–e93379. https://doi.org/10.1371/journal.pone.0093379

Breiman, L. (2001). *Random Forests* (Vol. 45).

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., & Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, *28*(2), 171–182. https://doi.org/10.1002/gepi.20041

Camargo, A., Azuaje, F., Wang, H., & Zheng, H. (2008). Permutation - Based statistical tests for multiple hypotheses. *Source Code for Biology and Medicine*, *3*(Dcm), 1–8. https://doi.org/10.1186/1751-0473-3-15

Carlborg, Ö., & Haley, C. S. (2004). *Epistasis: too often neglected in complex trait studies?* www.nature.com/reviews/genetics

Chanda, P., Sucheston, L., Zhang, A., Brazeau, D., Freudenheim, J. L., Ambrosone, C., & Ramanathan, M. (2008). *AMBIENCE: A Novel Approach and Efficient Algorithm for Identifying Informative Genetic and Environmental Associations With Complex Phenotypes*. https://doi.org/10.1534/genetics.108.088542

Chanda, P., Sucheston, L., Zhang, A., & Ramanathan, M. (2009). The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors. *European Journal of Human Genetics*, *17*, 1274–1286. https://doi.org/10.1038/ejhg.2009.38

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1). https://doi.org/10.1186/s13742-015-0047-8

Chang, K.-H., Chang, M.-Y., Muo, C.-H., Wu, T.-N., Chen, C.-Y., & Kao, C.-H. (n.d.). *Increased Risk of Dementia in Patients Exposed to Nitrogen Dioxide and Carbon Monoxide: A Population-Based Retrospective Cohort Study*. https://doi.org/10.1371/journal.pone.0103078

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. https://doi.org/10.1093/bioinformatics/btn653

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Ruden, D. M., & Lu, X. (2012). *© 2012 Landes Bioscience . Do not distribute . © 2012 Landes Bioscience . Do not distribute . June*, 1–13.

Cockerham, C. C., & Zeng, Z.-B. (1996). Design I11 With Marker Loci. In *Genetics* (Vol. 143).

Cohen, J., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K., & Hobbs, H. H. (2005). Low LDL cholesterol in individuals of African descent resulting from

frequent nonsense mutations in PCSK9. *Nature Genetics*, *37*(2), 161–165. https://doi.org/10.1038/ng1509

Combarros, O., van Duijn, C. M., Hammond, N., Belbin, O., Arias-Vásquez, A., Cortina-Borja, M., Lehmann, M. G., Aulchenko, Y. S., Schuur, M., Kölsch, H., Heun, R., Wilcock, G. K., Brown, K., Kehoe, P. G., Harrison, R., Coto, E., Alvarez, V., Deloukas, P., Mateo, I., … Lehmann, D. J. (2009). Replication by the Epistasis Project of the interaction between the genes for IL-6 and IL-10 in the risk of Alzheimer's disease. *Journal of Neuroinflammation*, *6*, 22. https://doi.org/10.1186/1742-2094-6-22

Combarros, O., Van Duijn, C. M., Hammond, N., Belbin, O., Arias-Vásquez, A., Cortina-Borja, M., Lehmann, M. G., Aulchenko, Y. S., Schuur, M., Kölsch, H., Heun, R., Wilcock, G. K., Brown, K., Kehoe, P. G., Harrison, R., Coto, E., Alvarez, V., Deloukas, P., Mateo, I., … Lehmann, D. J. (2009). *Journal of Neuroinflammation Replication by the Epistasis Project of the interaction between the genes for IL-6 and IL-10 in the risk of Alzheimer's disease*. https://doi.org/10.1186/1742-2094-6-22

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, *11*(20), 2463–2468. https://doi.org/10.1093/hmg/11.20.2463

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, *10*(6), 392–404. https://doi.org/10.1038/nrg2579

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L., & Pericak-Vance, M. A. (1993). Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families. *Science*, *261*(5123), 921–923. http://www.jstor.org/stable/2882127

Cordovado, S. K., Hendrix, M., Greene, C. N., Mochal, S., Earley, M. C., Farrell, P. M., Kharrazi, M., Hannon, W. H., & Mueller, P. W. (2012). CFTR mutation analysis and haplotype associations in CF patients. *Molecular Genetics and Metabolism*, *105*(2), 249–254. https://doi.org/10.1016/j.ymgme.2011.10.013

Cover, T. M., & Thomas, J. A. (1991). Elements of Information Theory [Book Review]. In *Elements of Information Theory* (Issue 1). https://doi.org/10.1109/tit.1993.1603955

Damodarasamy, M., Vernon, R. B., Pathan, J. L., Keene, C. D., Day, A. J., Banks, W. A., & Reed, M. J. (2020). The microvascular extracellular matrix in brains with Alzheimer's disease neuropathologic change (ADNC) and cerebral amyloid angiopathy (CAA). *Fluids Barriers CNS*, *17*, 60. https://doi.org/10.1186/s12987-

020-00219-y

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). https://doi.org/10.1093/nar/gks364

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. https://doi.org/10.1093/bib/bbt004

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. https://doi.org/10.1093/nar/gky399

De Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., Heap, G., Nimmo, E. R., Ahmad, T., Mansfield, J. C., Anderson, C. A., & Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Jeremy Sanderson*, *49*(2). https://doi.org/10.1038/ng.3760

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T., Huang, W., & Li, Y. (2008). Exploration of gene-gene interaction effects using entropy-based methods. *European Journal of Human Genetics*, *16*(2), 229–235. https://doi.org/10.1038/sj.ejhg.5201921

Dorani, F., Hu, T., Woods, M. O., & Zhai, G. (2018). Ensemble learning for detecting gene-gene interactions in colorectal cancer. *PeerJ*, *6*, e5854. https://doi.org/10.7717/peerj.5854

Ebbert, M. T., Boehme, K. L., Wadsworth, M. E., Staley, L. A., Mukherjee, S., Crane, P. K., Ridge, P. G., Kauwe, J. S., & Dement Author manuscript, A. (2016). Interaction between variants in CLU and MS4A4E modulates Alzheimer's disease risk HHS Public Access Author manuscript. *Alzheimers Dement*, *12*(2), 121–129. https://doi.org/10.1016/j.jalz.2015.08.163

Ebbert, M. T. W., Ridge, P. G., & Kauwe, J. S. K. (2015). *Bridging the Gap between Statistical and Biological Epistasis in Alzheimer's Disease*. https://doi.org/10.1155/2015/870123

Ebbert, M. T. W., Ridge, P. G., Wilson, A. R., Sharp, A. R., Norton, M. C., Tschanz, J. T., Munger, R. G., & Christopher, D. (2015). *Population-based analysis of*

Alzheimer's disease risk alleles implicates genetic interactions. *75*(9), 732–737. https://doi.org/10.1016/j.biopsych.2013.07.008.Population-based

Fan, R, Zhong, M, Wang, S. (2011). Entropy-Based Information Gain Approaches to Detect and to Characterize Gene-Gene and Gene-Environment Interactions/ Correlations of Complex Diseases. *Genetic Epidemiology*, *35*(7), 706–721. https://doi.org/10.1002/gepi.20621.Entropy-Based

Fenger, M., Wang, T., Yu, Z., Lin, W.-Y., W-y, L., C-c, H., Y-l, L., S-j, T., P-h, K., Huang, C.-C., Liu, Y.-L., Tsai, S.-J., & Kuo, P.-H. (2019). Genome-Wide Gene-Environment Interaction Analysis Using Set-Based Association Tests. *Front. Genet*, *9*, 715. https://doi.org/10.3389/fgene.2018.00715

Ferrario, P. G., & König, I. R. (2018). Transferring entropy to the realm of GxG interactions. *Briefings in Bioinformatics*, *19*(1), 136–147. https://doi.org/10.1093/bib/bbw086

Filippini, N., Rao, A., Wetten, S., Gibson, R. A., Borrie, M., Guzman, D., Kertesz, A., Loy-English, I., Williams, J., Nichols, T., Whitcher, B., & 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. https://doi.org/10.1016/J.NEUROIMAGE.2008.10.003

Flaten, P. (n.d.). *Geographical associations between aluminium in drinking water and death rates with dementia (including Alzheimer's disease), Parkinson's disease and amyotrophic lateral sclerosis in Norway*.

Gatz, M., Reynolds, C. A., Fratiglioni, L., Mortimer, J. A., Berg, S., Fiske, A., & Pedersen, N. L. (2006). Role of Genes and Environments for Explaining Alzheimer Disease. In *Arch Gen Psychiatry* (Vol. 63).

Granados, E. A. O., Vásquez, L. F. N., & Granados, H. A. (2013). Characterizing genetic interactions using a machine learning approach in Colombian patients with Alzheimer's disease. *Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013*, 1–2. https://doi.org/10.1109/BIBM.2013.6732588

Grunin, M., Wheeler, N. R., Bush, W. S., & Haines, J. L. (2020). Using linkage analysis to identify novel gene-gene interactions in Alzheimer's disease. *Alzheimer's & Dementia*, *16*(S2). https://doi.org/10.1002/alz.043435

Gusareva, E. S., Carrasquillo, M. M., Bellenguez, C., Cuyvers, E., Colon, S., Graff-radford, N. R., Petersen, R. C., Dickson, D. W., Mahachie, J. M., Bessonov, K.,

Broeckhoven, C. Van, Consortium, G., Harold, D., Williams, J., Amouyel, P., Sleegers, K., Ertekin-taner, N., Lambert, J., & Steen, K. Van. (2014). *Genome-wide association interaction analysis for Alzheimer's disease*. *35*, 2436–2443. https://doi.org/10.1016/j.neurobiolaging.2014.05.014

Gusareva, E. S., Kristel, ·, & Steen, V. (2014). Practical aspects of genome-wide association interaction analysis. *Hum Genet*, *133*, 1343–1358. https://doi.org/10.1007/s00439-014-1480-y

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M., Singh Pahwa, J., Moskvina, V., Dowzell, K., Williams, A., Jones, N., Thomas, C., Stretton, A., Morgan, A., Lovestone, S., Powell, J., Proitsi, P., Lupton, M. K., Brayne, C., … Williams, J. (n.d.). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease, and shows evidence for additional susceptibility genes. *Sebastiaan Engelborghs*, *18*, 31. https://doi.org/10.1038/ng.440

Haynes, W. (2013). *Bonferroni Correction BT - Encyclopedia of Systems Biology* (W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.); p. 154). Springer New York. https://doi.org/10.1007/978-1-4419-9863-7_1213

Hochberg, Y. (1995). *Controlling the False Discovery Rate : a Practical and Powerful Approach to Multiple Testing*. *I*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Hohman, T. J., Bush, W. S., Jiang, L., Brown-Gentry, K. D., Torstenson, E. S., Dudek, S. M., Mukherjee, S., Naj, A., Kunkle, B. W., Ritchie, M. D., Martin, E. R., Schellenberg, G. D., Mayeux, R., Farrer, L. A., Pericak-Vance, M. A., Haines, J. L., & Thornton-Wells, T. A. (2016). Discovery of gene-gene interactions across multiple independent data sets of late onset Alzheimer disease from the Alzheimer Disease Genetics Consortium. *Neurobiology of Aging*, *38*, 141–150. https://doi.org/10.1016/j.neurobiolaging.2015.10.031

Hohman, T. J., Bush, W. S., Jiang, L., Brown-Gentry, K. D., Torstenson, E. S., Dudek, S. M., Mukherjee, S., Naj, A., Kunkle, B. W., Ritchie, M. D., Martin, E. R., Schellenberg, G. D., Mayeux, R., Farrer, L. A., Pericak-Vance, M. A., Haines, J. L., Thornton-Wells, T. A., & Macdonald, J. T. (2016). Discovery of gene-gene interactions across multiple independent datasets of Late Onset Alzheimer Disease from the Alzheimer Disease Genetics Consortium HHS Public Access. *Neurobiol Aging*, *38*, 141–150. https://doi.org/10.1016/j.neurobiolaging.2015.10.031

Hohman, T. J., Koran, M. E., & Thornton, T. (2013). *Epistatic Genetic Effects among Alzheimer's Candidate Genes*. https://doi.org/10.1371/journal.pone.0080839

Hu, J. K., Wang, X., & Wang, P. (2014). Testing Gene-Gene Interactions in Genome Wide Association Studies. *Genet Epidemiol*, *38*, 123–134. https://doi.org/10.1002/gepi.21786

Hu, T., Chen, Y., Kiralis, J. W., Collins, R. L., Wejse, C., Sirugo, G., Williams, S. M., & Moore, J. H. (2013a). An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *Journal of the American Medical Informatics Association*, *20*(4), 630–636. https://doi.org/10.1136/amiajnl-2012-001525

Hu, T., Chen, Y., Kiralis, J. W., Collins, R. L., Wejse, C., Sirugo, G., Williams, S. M., & Moore, J. H. (2013b). An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *Journal of the American Medical Informatics Association*, *20*(4), 630–636. https://doi.org/10.1136/amiajnl-2012-001525

Hu, T., Chen, Y., Kiralis, J. W., & Moore, J. H. (2013). ViSEN: Methodology and software for visualization of statistical epistasis networks. *Genetic Epidemiology*, *37*(3), 283–285. https://doi.org/10.1002/gepi.21718

Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., … Birney, E. (2007). Ensembl 2007. *Nucleic Acids Research*, *35*(suppl_1), D610–D617. https://doi.org/10.1093/nar/gkl996

Infante, J., Rodríguez-Rodríguez, E., Mateo, I., Llorca, J., Vázquez-Higuera, J. L., Berciano, J., & Combarros, O. (2010). Gene–gene interaction between heme oxygenase-1 and liver X receptor-β and Alzheimer's disease risk. *Neurobiology of Aging*, *31*(4), 710–714. https://doi.org/10.1016/J.NEUROBIOLAGING.2008.05.025

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., Voyle, N., Proitsi, P., Witoelar, A., Stringer, S., Aarsland, D., Almdahl, I. S., Andersen, F., Bergh, S., Bettella, F., … Posthuma, D. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics*, *51*(3), 404–413. https://doi.org/10.1038/s41588-018-0311-9

Joas, E., Bäckman, K., Gustafson, D., Östling, S., Waern, M., Guo, X., & Skoog, I. (2012). Blood pressure trajectories from midlife to late life in relation to dementia in women followed for 37 years. *Hypertension*, *59*(4), 796–801. https://doi.org/10.1161/HYPERTENSIONAHA.111.182204

Johanna Kuusisto, Keijo Koivisto, Kari Kervinen, Leena Mykkanen, Eeva-Liisa Helkala, Matti Vanhanen, Tuomo Hanninen, Kalevi Py6rala, Y Antero Kesaniemi, Paavo Riekkinen, M. L. (1994). Association ofapolipoprotein E phenotypes with late onset Alzheimer's disease: population based study. *BMJ*, *309*(8), 309–636.

Kauwe, J., Bertelsen, S., Mayo, K., Cruchaga, C., Abraham, R., Hollingworth, P., Harold, D., Owen, M., Williams, J., Lovestone, S., Morris, J., Goate, A., Neuroimaging Initiative, D., Goate, A. M., & Ludwig Professor, M. S. (2010). Suggestive synergy between genetic variants in TF and HFE as risk factors for Alzheimer's disease. *Am J Med Genet B Neuropsychiatr Genet*, *153*(4), 955–959. https://doi.org/10.1002/ajmg.b.31053

Knopman, D. S., DeKosky, S. T., Cummings, J. L., Chui, H., Corey–Bloom, J., Relkin, N., Small, G. W., Miller, B., & Stevens, J. C. (2001). Practice parameter: Diagnosis of dementia (an evidence-based review). *Neurology*, *56*(9), 1143–1153. https://doi.org/10.1212/wnl.56.9.1143

Koran, M. E. I., Hohman, T. J., Meda, S. A., Thornton-Wells, T. A., & Initiative, for the A. D. N. (2014). Genetic Interactions within Inositol-Related Pathways are Associated with Longitudinal Changes in Ventricle Size. *Journal of Alzheimer's Disease*, *38*, 145–154. https://doi.org/10.3233/JAD-130989

Kwon, M. S., Park, M., & Park, T. (2014). IGENT: Efficient entropy based algorithm for genome-wide gene-gene interaction analysis. *BMC Medical Genomics*, *7*(SUPPL.1), 1–11. https://doi.org/10.1186/1755-8794-7-S1-S6

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., Grenier-Boley, B., Russo, G., Thornton-Wells, T. A., Jones, N., Smith, A. V., Chouraki, V., Thomas, C., Ikram, M. A., Zelenika, D., … Seshadri, S. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, *45*(12), 1452–1458. https://doi.org/10.1038/ng.2802

Lee, J. H., Cheng, R., Barral, S., Reitz, C., Medrano, M., Lantigua, R., Jiménez-Velazquez, I. Z., Rogaeva, E., St George-Hyslop, P. H., & Mayeux, R. (2011). ONLINE FIRST Identification of Novel Loci for Alzheimer Disease and Replication of CLU, PICALM, and BIN1 in Caribbean Hispanic Individuals. *Arch Neurol*, *68*(3), 320–328. https://doi.org/10.1001/archneurol.2010.292

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P. C., Mägi, R., Milani, L., Fischer, K., & Metspalu, A. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International Journal of Epidemiology*, *44*(4), 1137–1147. https://doi.org/10.1093/ije/dyt268

Li, G., Pan, T., Guo, D., & Li, L.-C. (2014). *Regulatory Variants and Disease: The E-Cadherin −160C/A SNP as an Example*. https://doi.org/10.1155/2014/967565

Li, H., Wetten, S., Li, L., St. Jean, P. L., Upmanyu, R., Surh, L., Hosford, D., Barnes, M. R., Briley, J. D., Borrie, M., Coletta, N., Delisle, R., Dhalla, D., Ehm, M. G., Feldman, H. H., Fornazzari, L., Gauthier, S., Goodgame, N., Guzman, D., … Roses, A. D. (2008). Candidate Single-Nucleotide Polymorphisms From a Genomewide Association Study of Alzheimer Disease. *Archives of Neurology*, *65*(1), 45–53. https://doi.org/10.1001/archneurol.2007.3

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J., Song, Z., Ji, W., Wang, M., Zhou, J., Chen, B., Liu, Y., Wang, J., Wang, P., Yang, P., … Ji, J. (2017). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature GeNetics Ripke*, *49*(11), 12. https://doi.org/10.1038/ng.3973

Ma, J., Ma, C., Li, J., Sun, Y., Ye, F., Liu, K., & Zhang, H. (2020). Extracellular Matrix Proteins Involved in Alzheimer's Disease. *Chemistry – A European Journal*, *26*(53), 12101–12110. https://doi.org/https://doi.org/10.1002/chem.202000782

Maioli, S., Leander, K., Nilsson, P., & Nalvarte, I. (2021). Estrogen receptors and the aging brain. *Essays in Biochemistry*, *65*, 913–925. https://doi.org/10.1042/EBC20200162

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2), 1–10. https://doi.org/10.1002/mpr.1608

McKinney, B. A., Reif, D. M., Ritchie, M. D., & Moore, J. H. (2006). Machine learning for detecting gene-gene interactions: A review. *Applied Bioinformatics*, *5*(2), 77–88. https://doi.org/10.2165/00822942-200605020-00002

Meda, S. A., Koran, M. E. I., Pryweller, J. R., Vega, J. N., & Thornton-Wells, T. A. (2013). Genetic interactions associated with 12-month atrophy in hippocampus and entorhinal cortex in Alzheimer's Disease Neuroimaging Initiative. *Neurobiology of Aging*, *34*(5), 1518.e9-1518.e18. https://doi.org/10.1016/j.neurobiolaging.2012.09.020

Moore, J. H., & Hu, T. (2015). *Epistasis Analysis Using Information Theory BT - Epistasis: Methods and Protocols* (J. H. Moore & S. M. Williams (Eds.); pp. 257–268). Springer New York. https://doi.org/10.1007/978-1-4939-2155-3_13

Moore, J. H., & Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. *BioEssays*, *27*(6), 637–646. https://doi.org/10.1002/bies.20236

Mullins, N., Forstner, A. J., OConnell, K. S., Coombes, B., I Coleman, J. R., Qiao, Z., Als, T. D., Bigdeli, T. B., Børte, S., Bryois, J., Charney, A. W., Kristian Drange, O., Gandal, M. J., Hagenaars, S. P., Ikeda, M., Kamitaki, N., Kim, M., Krebs, K., Panagiotaropoulou, G., … Andreassen, O. A. (n.d.). *Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology*. https://doi.org/10.1038/s41588-021-00857-4

Nachman, M. W. (2001). Single nucleotide polymorphisms and recombination rate in humans. *Trends in Genetics*, *17*(9), 481–485. https://doi.org/10.1016/S0168-9525(01)02409-X

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). *Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS*. https://doi.org/10.1371/journal.pgen.1000888

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., Webb, T. R., Zeng, L., Dehghan, A., Alver, M., Armasu, S. M., Auro, K., Bjonnes, A., Chasman, D. I., Chen, S., … Consortium, the Cardi. (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, *47*(10), 1121–1130. https://doi.org/10.1038/ng.3396

Nobrega, C., Yanagawa, Y., University, G., Zheng Wu, J., Ayciriex, S., Djelti, F., Alves, S., Regazzetti, A., Gaudin, M., Varin, J., Langui, D., Bièche, I., Hudry, E., Dargère, D., Aubourg, P., Auzeil, N., Laprévote, O., & Cartier, N. (2017). Neuronal Cholesterol Accumulation Induced by Cyp46a1 Down-Regulation in Mouse Hippocampus Disrupts Brain Lipid Homeostasis. *Frontiers in Molecular Neuroscience | Www.Frontiersin.Org*, *1*, 211. https://doi.org/10.3389/fnmol.2017.00211

Norton, S., Matthews, F. E., Barnes, D. E., Yaffe, K., & Brayne, C. (2014). Potential for primary prevention of Alzheimer's disease: An analysis of population-based data. *The Lancet Neurology*, *13*(8), 788–794. https://doi.org/10.1016/S1474-4422(14)70136-X

Oki, N. O., & Motsinger-Reif, A. A. (2011). Multifactor dimensionality reduction as a filter-based approach for genome wide association studies. *Frontiers in Genetics*, *2*(NOV), 1–17. https://doi.org/10.3389/fgene.2011.00080

Oudin, A., Forsberg, B., Nordin Adolfsson, A., Lind, N., Modig, L., Nordin, M., Nordin, S., Adolfsson, R., & Nilsson, L.-G. (n.d.). *Traffic-Related Air Pollution and Dementia Incidence in Northern Sweden: A Longitudinal Study*. https://doi.org/10.1289/ehp.1408322

Patel, H., Hodges, A. K., Curtis, C., Lee, S. H., Troakes, C., Dobson, R. J. B., & Newhouse, S. J. (2019). Transcriptomic analysis of probable asymptomatic and symptomatic alzheimer brains. *Brain, Behavior, and Immunity*, *80*(May), 644–656. https://doi.org/10.1016/j.bbi.2019.05.009

Pirooznia M, Seifuddin F, Judy J, et al. (2012). Data Mining Approaches for Genome-Wide Association of Mood Disorders. *Psychiatric Genetics*, *22*(2), 55–61. https://doi.org/10.1097/YPG.0b013e32834dc40d.Data

Prokopenko, D., Morgan, S. L., Mullin, K., Hofmann, O., Chapman, B., Kirchner, R., Amberkar, S., Wohlers, I., Lange, C., Hide, W., Bertram, L., & 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 and Dementia*, *17*(9), 1509–1527. https://doi.org/10.1002/alz.12319

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., Murphy, M. R., O'Leary, N. A., Pujar, S., Rajput, B., Rangwala, S. H., Riddick, L. D., Shkeda, A., Sun, H., Tamez, P., … Ostell, J. M. (2014). RefSeq: An update on mammalian reference sequences. *Nucleic Acids Research*, *42*(D1), 756–763. https://doi.org/10.1093/nar/gkt1114

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rabinovici, G. D. (2019). *Late-onset Alzheimer Disease*.

Raghavan, N., & Tosto, G. (2017). Genetics of Alzheimer's Disease: the Importance of Polygenic and Epistatic Components. *Current Neurology and Neuroscience Reports*, *17*(78). https://doi.org/10.1007/s11910-017-0787-1

Reitz, C., Cheng, R., Rogaeva, E., Lee, J. H., Tokuhiro, S., Zou, F., Bettens, K., Sleegers, K., Tan, E. K., Kimura, R., Shibata, N., Arai, H., Kamboh, M. I., Prince, J. A., Maier, W., Riemenschneider, M., Owen, M., Harold, D., Hollingworth, P., … Mayeux, R. (2011). Meta-analysis of the association between variants in SORL1 and Alzheimer

disease. *Archives of Neurology*, *68*(1), 99–106. https://doi.org/10.1001/archneurol.2010.346

Reitz, C., & Mayeux, R. (2014). Alzheimer disease: Epidemiology, diagnostic criteria, risk factors and biomarkers. In *Biochemical Pharmacology* (Vol. 88, Issue 4, pp. 640–651). Elsevier Inc. https://doi.org/10.1016/j.bcp.2013.12.024

Ridge, P. G., Mukherjee, S., Crane, P. K., & Kauwe, J. S. K. (2013). Alzheimer's Disease: Analyzing the Missing Heritability. *PLoS ONE*, *8*(11). https://doi.org/10.1371/journal.pone.0079771

Ritchie, K., Ritchie, C. W., Yaffe, K., Skoog, I., & Scarmeas, N. (n.d.). *Is late-onset Alzheimer's disease really a disease of midlife?* https://doi.org/10.1016/j.trci.2015.06.004

Robson, K. J. H., Lehmann, D. J., Wimhurst, V. L. C., Livesey, K. J., Combrinck, M., Merryweather-Clarke, A. T., Warden, D. R., & Robson, K. (2004). Synergy between the C2 allele of transferrin and the C282Y allele of the haemochromatosis gene (HFE) as risk factors for developing Alzheimer's disease. *J Med Genet*, *41*, 261–265. https://doi.org/10.1136/jmg.2003.015552

Rojas, I. de, Moreno-Grau, S., Tesi, N., Grenier-Boley, B., Andrade, V., Jansen, I. E., Pedersen, N. L., Stringa, N., Zettergren, A., Hernández, I., Montrr, L., & Tankard, R. M. (n.d.). *Common variants in Alzheimer's disease and risk stratification by polygenic risk scores*. https://doi.org/10.1038/s41467-021-22491-8

Rondeau, V., Lè Ne Jacqmin-Gadda, H., Commenges, D., Helmer, C., & Dartigues, J.-F. X. (2009). Original Contribution Aluminum and Silica in Drinking Water and the Risk of Alzheimer's Disease or Cognitive Decline: Findings From 15-Year Follow-up of the PAQUID Cohort. *Am J Epidemiol*, *169*(4), 489–496. https://doi.org/10.1093/aje/kwn348

Ruzong Fan, P. D. (n.d.). *R codes for Entropy-based Information Gain Approaches to Detect Gene-Gene and Gene-Environment Interactions/Correlations of Complex Diseases, R. Fan, 2015*. https://georgetown.app.box.com/s/ptf0niqqquc5m3zxstehdoormpgh7hq5

Shannon, C. E. (1949). *A Mathematical Theory of Communication*. *April 1924*, 1–54.

Single Nucleotide Polymorphism. (n.d.). https://www.nature.com/scitable/definition/snp-295/

Sinoquet, C. (2018). A method combining a random forest-based technique with the

modeling of linkage disequilibrium through latent variables, to run multilocus genome-wide association studies. *BMC Bioinformatics*, *19*(1), 106. https://doi.org/10.1186/s12859-018-2054-0

Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., Helgadottir, H. T., Johannsdottir, H., Magnusson, O. T., Gudjonsson, S. A., Justesen, J. M., Harder, M. N., Jørgensen, M. E., Christensen, C., Brandslund, I., Sandbæk, A., Lauritzen, T., Vestergaard, H., Linneberg, A., … Stefansson, K. (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nature Genetics*, *46*(3), 294–298. https://doi.org/10.1038/ng.2882

Su, L., Liu, G., Wang, H., Tian, Y., Zhou, Z., Han, L., & Yan, L. (2015). Research on single nucleotide polymorphisms interaction detection from network perspective. *PLoS ONE*, *10*(3), 1–19. https://doi.org/10.1371/journal.pone.0119146

Sud, A., Kinnersley, B., & Houlston, R. S. (2017). *Genome-wide association studies of cancer: current insights and future perspectives*. https://doi.org/10.1038/nrc.2017.82

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, *12*(3), 1–10. https://doi.org/10.1371/journal.pmed.1001779

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. In *Nature Reviews Genetics* (Vol. 20, Issue 8, pp. 467–484). Nature Publishing Group. https://doi.org/10.1038/s41576-019-0127-1

Ueki, M., & Cordell, H. J. (2012). Improved statistics for genome-wide interaction analysis. *PLoS Genetics*, *8*(4). https://doi.org/10.1371/journal.pgen.1002625

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 59. https://doi.org/10.1038/s43586-021-00056-9

Wang, C., Zhang, F., Jiang, S., Siedlak, S. L., Shen, L., Perry, G., Wang, X., Tang, B., & Zhu, X. (2016). *Estrogen receptor-α is localized to neurofibrillary tangles in Alzheimer's disease OPEN*. https://doi.org/10.1038/srep20352

Warden, D. R., & Smith, A. D. (2004). *Synergy between the C2 allele of transferrin and the C282Y allele of the haemochromatosis gene (HFE) as risk factors for developing Alzheimer's disease*. 261–265. https://doi.org/10.1136/jmg.2003.015552

Wei, W., Hemani, G., & Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Publishing Group*, *15*(11), 722–733. https://doi.org/10.1038/nrg3747

Wilhelm Steinbusch, H., Schmitt, K., Carel Wildering, W., Yang, Q., Bai, Z., Sun, Y., Xu, S., Jiang, M., Liu, X., & Yang, L. (2021). *Role of the Extracellular Matrix in Alzheimer's Disease*. https://doi.org/10.3389/fnagi.2021.707466

Wu, L., Rosa-Neto, P., Hsiung, G.-Y. R., Sadovnick, A. D., Masellis, M., Black, S. E., Jia, J., & Gauthier, S. (2022). Early-Onset Familial Alzheimer's Disease (EOFAD). *Can J Neurol Sci*, *39*, 436–445. https://doi.org/10.1017/S0317167100013949

Xu, X., Zhang, B., Wang, X., Zhang, Q., Wu, X., Zhang, J., Bai, Y., & Gu, X. (2021). A meta-analysis of Alzheimer's disease's relationship with human ApoE gene variants. In *Am J Transl Res* (Vol. 13, Issue 9). www.ajtr.org

Zhao, W., Rasheed, A., Tikkanen, E., Lee, J.-J., Butterworth, A. S., M Howson, J. M., Assimes, T. L., Chowdhury, R., Orho-Melander, M., Damrauer, S., Small, A., Asma, S., Imamura, M., Yamauch, T., Chambers, J. C., Chen, P., Sapkota, B. R., Shah, N., Jabeen, S., … Saleheen, D. (2017). *Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease*. https://doi.org/10.1038/ng.3943

Zieselman, A. L., Fisher, J. M., Hu, T., Andrews, P. C., Greene, C. S., Shen, L., Saykin, A. J., Moore, J. H., & Neuroimaging, D. (2014). *Computational genetics analysis of grey matter density in Alzheimer ' s disease*. 1–7.

APPENDICES

Appendix A Scripts Used In This Study

A.1. Python Script for the Calculations of Frequencies of Genotypes in Case and Control Groups

```
import json
import configparser
import argparse
import pandas as pd

parser=argparse.ArgumentParser()
parser.add_argument('--snpListFile', type=str,help='Prioritized SNPs genotype list file
path for case group')
parser.add_argument('--interactionType',type=str,help='Type of interaction
information can be 3WII or 2WI')
parser.add_argument('--outputFilePath',type=str,help='Json file path for the
frequencies of triplets or pairs of SNPs')
args=parser.parse_args()

############################################################
#This function remove the family IDs and recode the genotypes in case control
genotyping files
############################################################

def convertGtypeCoding(snpListFilePath):
    f = open(snpListFilePath, "r")
    #g = open(controlFilePath, "r")
    snplist = f.readlines()
    f.close()
    for n, i in enumerate(snplist):
        snplist[n]=snplist[n].split()
        temp_list=snplist[n][3:len(snplist[n])]
        #Remove the familiy IDs with this step
        del temp_list[0::2]
        temp_list=list(dict.fromkeys(temp_list))
```

```
        snplist[n][3:len(snplist[n])]=temp_list

    for i in range (0,len(snplist)-1):
        if snplist[i][2]=='11':
            snplist[i][2]='0'
        elif snplist[i][2]=='12' or snplist[i][2]=='21' :
            snplist[i][2]='1'
        elif snplist[i][2] == '22' :
            snplist[i][2]='2'
    return(snplist)


#################################################################
#This function generates a csv file with frequencies of triplets or pairs
#################################################################
def convertgtypeII(snplist,interactionType,outputPath):
    freq="Freq_"
    nofcom={}
    temp_dic={}
    x=0
    if interactionType=="3WII":
        for i in range (0,len(snplist)-11,4) :
            for j in range (i+4,len(snplist)-7,4):
                for k in range (j+4,len(snplist)-3,4):
                    for l in range (i,i+4):
                        for m in range (j,j+4):
                            for n in range (k,k+4):
                                if l%4 != 3 and m%4 != 3 and n%4 != 3:
                                    name_col=freq+snplist[l][2]+snplist[m][2]+snplist[n][2]
                                    set_1=set(snplist[l][3:len(snplist[l])])
                                    set_2=set(snplist[m][3:len(snplist[m])])
                                    set_3=set(snplist[n][3:len(snplist[n])])
                                    temp_dic.update({name_col:len(set_1 & set_2 & set_3)})
                    if l%4==3:

temp_dic.update({'SNP_1':snplist[l][1],'SNP_2':snplist[m][1],'SNP_3':snplist[n][1]})
                        nofcom.update({x:temp_dic})
                        x=x+1
                        temp_dic={}
```

```
    else:
        for j in range (0,len(snplist)-7,4):
            for k in range (j+4,len(snplist)-3,4):
                for m in range (j,j+4):
                    for n in range (k,k+4):
                        if m%4 != 3 and n%4 != 3:
                            name_col=freq+snplist[m][2]+snplist[n][2]
                            set_2=set(snplist[m][3:len(snplist[m])])
                            set_3=set(snplist[n][3:len(snplist[n])])
                            temp_dic.update({name_col:len(set_2 & set_3)})
                if m%4==3:
                    temp_dic.update({'SNP_1':snplist[m][1],'SNP_2':snplist[n][1]})
                    nofcom.update({x:temp_dic})
                    x=x+1
                    temp_dic={}

    nofcom=pd.DataFrame.from_dict(nofcom,orient='index')
    nofcom.to_csv(outputPath,index=False)
    return


interaction=args.interactionType
snpListFilePath=args.snpListFile
outputPath=args.outputFilePath

snpListEncoded=convertGtypeCoding(snpListFilePath)
convertgtypeII(snpListEncoded,interaction,outputPath)
```

A.2. R Program that can calculate p-values of example of 3 way with test statistics T_IIG adapted from Fan R. (Ruzong Fan, n.d.)

```
p_value_func_T_TIIG <- function(X1, Y1)
   {
   M  <- sum(X1)
   P1 <- X1/M
   P  <- P1[c(-27)]

   N  <- sum(Y1)
   Q1 <- Y1/N
   Q  <- Q1[c(-27)]

        Sigma <- diag(P) - P %*% t(P)

       PGABE1 <- array(P1, dim=c(3,3,3))

       h   <- array (0, dim=c(3,3,3) )
       dh  <- matrix (0, nrow = 1, ncol = 26)
       tmp<- 0

       if (PGABE1[3,3,3] > 0)
       tmp                  <-                  tmp                  -                  log2(
PGABE1[3,3,3]*sum(PGABE1[3,,])*sum(PGABE1[,3,])*sum(PGABE1[,,3])/(
sum(PGABE1[3,3,]) * sum(PGABE1[3,,3]) * sum(PGABE1[,3,3]) ) )
       for (i in 1:3)
         for (j in 1:3)
            for (e in 1:3)

              if (PGABE1[i,j,e] > 0)
                {
                  h[i,j,e]            <-            PGABE1[i,j,e]            *            log2(
PGABE1[i,j,e]*sum(PGABE1[i,,])*sum(PGABE1[,j,])*sum(PGABE1[,,e])/(
sum(PGABE1[i,j,]) * sum(PGABE1[i,,e]) * sum(PGABE1[,j,e]) ) )

                  if ( i+j+e < 9 )
                    {
                     tmp1                                 <-                                 log2(
PGABE1[i,j,e]*sum(PGABE1[i,,])*sum(PGABE1[,j,])*sum(PGABE1[,,e])/(
sum(PGABE1[i,j,]) * sum(PGABE1[i,,e]) * sum(PGABE1[,j,e]) ) )
                     dh[9*(e-1)+3*(j-1)+i] <- tmp + tmp1
                     }
```

```
                }

        SigmaD <- diag(Q) - Q %*% t(Q)

        QGABE1 <- array(Q1, dim=c(3,3,3))

        l <- array (0, dim=c(3,3,3) )
        dl <- matrix (0, nrow = 1, ncol = 26)
        tmp2<- 0

        if ( QGABE1[3,3,3] > 0)
        tmp2 <- tmp2 - log2(
QGABE1[3,3,3]*sum(QGABE1[3,,])*sum(QGABE1[,3,])*sum(QGABE1[,,3])/(
sum(QGABE1[3,3,]) * sum(QGABE1[3,,3]) * sum(QGABE1[,3,3]) ) )
         for (i in 1:3)
           for (j in 1:3)
             for (e in 1:3)

                if (QGABE1[i,j,e] > 0)
                   {
                     l[i,j,e] <- QGABE1[i,j,e] * log2(
QGABE1[i,j,e]*sum(QGABE1[i,,])*sum(QGABE1[,j,])*sum(QGABE1[,,e])/(
sum(QGABE1[i,j,]) * sum(QGABE1[i,,e]) * sum(QGABE1[,j,e]) ) )

                    if ( i+j+e < 9 )
                       {
                        tmp3 <- log2(
QGABE1[i,j,e]*sum(QGABE1[i,,])*sum(QGABE1[,j,])*sum(QGABE1[,,e])/(
sum(QGABE1[i,j,]) * sum(QGABE1[i,,e]) * sum(QGABE1[,j,e]) ) )
                        dl[9*(e-1)+3*(j-1)+i] <- tmp2 + tmp3
                        }
                  }

        Lambda <- dh %*% Sigma %*% t(dh)/M +   dl %*% SigmaD %*% t(dl)/N
        IIG <- sum(h)-sum(l)
        T_IIG <-(sum(h)-sum(l))^2/Lambda
        p_value<- 1-pchisq(T_IIG,1)
        T_IIG_pvalue<-c(IIG,T_IIG,p_value)
    }
```

A.3. R Program than can calculate p_value of example of 2 way with test statistic T_IG adapted from Fan R. (Ruzong Fan, n.d.)

```
p_value_func_T_IG <- function(X1, Y1)
  {
  M  <- sum(X1)
  P1 <- X1/M
  P  <- P1[c(-9)]

  N  <- sum(Y1)
  Q1 <- Y1/N
  Q  <- Q1[c(-9)]

  Sigma <- diag(P) - P %*% t(P)

  PGAB1 <- t( matrix(P1, nrow =3, ncol =3) )

  f   <- matrix (0, nrow = 3, ncol =3)
  df  <- matrix(0, nrow = 1, ncol = 8)
  tmp <- 0

  if (PGAB1[3, 3] > 0)
  tmp <- tmp - log2( PGAB1[3, 3]/( sum(PGAB1[3,]) * sum(PGAB1[,3]) ) )
  for (i in 1:3)
     for (j in 1:3)
          if (PGAB1[i, j] > 0)
            {
            f[i, j] <- PGAB1[i, j] * log2( PGAB1[i, j]/( sum(PGAB1[i,]) *
sum(PGAB1[,j]) ) )
            if ( 3*(i-1) + j < 9 )
              {
              tmp1 <- log2( PGAB1[i, j]/( sum(PGAB1[i,]) * sum(PGAB1[,j]) ) )
              df[3*(i-1)+j] <- tmp + tmp1
              }
            }


  SigmaD <- diag(Q) - Q %*% t(Q)

  QGAB1 <- t( matrix(Q1, nrow =3, ncol =3) )

  g <- matrix(0, nrow=3, ncol=3)
```

```
  dg <- matrix(0, nrow=1, ncol=8)

  tmp2 <- 0
  if ( QGAB1[3, 3] > 0)
    tmp2 <- tmp2 - log2( QGAB1[3, 3]/( sum(QGAB1[3,]) * sum(QGAB1[,3]) ) )
  for (i in 1:3)
     for (j in 1:3)
           if (QGAB1[i, j] > 0)
             {
             g[i, j] <- QGAB1[i, j] * log2( QGAB1[i, j]/( sum(QGAB1[i,]) *
sum(QGAB1[,j]) ) )
             if ( 3*(i-1) + j < 9 )
               {
               tmp3 <- log2( QGAB1[i, j]/( sum(QGAB1[i,]) * sum(QGAB1[,j]) ) )
               dg[3*(i-1)+j] <- tmp2 + tmp3
               }
             }

  Lambda <- df %*% Sigma %*% t(df)/M +   dg %*% SigmaD %*% t(dg)/N

  TIG <-(sum(g)-sum(f))^2/Lambda
  p_value<- 1-pchisq(TIG,1)

  TIG_pvalue<-c(TIG,p_value)
  }
```

A.4. R Program that we implemented interaction information analysis

```
library(dplyr)
library("optparse")
source("3-way_function_2015.R")
source("2way_TIG_function_2015.R")

option_list = list(

make_option(c("--caseFile"), type="character", default=NULL,
        help="Frequencies of triplets or pairs of SNPs for case group ", metavar="character"),

make_option(c("--controlFile"), type="character", default=NULL,
        help="Frequencies of triplets or pairs of SNPs for control group ", metavar="character"),

make_option(c("--intType"), type="character", default=NULL,
        help="Type of interaction:3-WII or 2WI", metavar="character"),

make_option(c("--outputFile"), type="character", default=NULL,
        help="Interaction information results output path", metavar="character")  );

opt_parser = OptionParser(option_list=option_list);
opt = parse_args(opt_parser);
casePath=opt$caseFile
controlPath=opt$controlFile
interactionType=opt$intType
outputPath=opt$outputFile

implementII <-function (caseFile, controlFile,interactionType ,outputFile){
        result_case <- read.csv(casePath)
        result_control <- read.csv(controlFile)

        if(interactionType=="3WII")
        {
                TIIG_Results<-data.frame(SNP_1<-
as.character(),SNP_2<as.character(),  SNP_3<-as.character(),  IIG<-as.numeric(), TIIG<-as.numeric(),TIIG_p_value<-as.numeric())

                for (i in 1:nrow(result_case))
                {
```

```
                X1<-c(result_control[i,1],result_control[i,2],result_control[i,3],
                      result_control[i,4],result_control[i,5],result_control[i,6],
                      result_control[i,7],result_control[i,8],result_control[i,9],
                    result_control[i,10],result_control[i,11],result_control[i,12],
                    result_control[i,13],result_control[i,14],result_control[i,15],
                    result_control[i,16],result_control[i,17],result_control[i,18],
                    result_control[i,19],result_control[i,20],result_control[i,21],
                    result_control[i,22],result_control[i,23],result_control[i,24],
                    result_control[i,25],result_control[i,26],result_control[i,27])

                Y1<-c(result_case[i,1],result_case[i,2],result_case[i,3],
                      result_case[i,4],result_case[i,5],result_case[i,6],
                      result_case[i,7],result_case[i,8],result_case[i,9],
                      result_case[i,10],result_case[i,11],result_case[i,12],
                      result_case[i,13],result_case[i,14],result_case[i,15],
                      result_case[i,16],result_case[i,17],result_case[i,18],
                      result_case[i,19],result_case[i,20],result_case[i,21],
                      result_case[i,22],result_case[i,23],result_case[i,24],
                      result_case[i,25],result_case[i,26],result_case[i,27])



                T_IIG_pvalue = p_value_func_T_TIIG(X1, Y1)
                 temp_results<-data.frame(SNP_1<-result_case[i,28],
                 SNP_2<-result_case[i,29], SNP_3<-result_case[i,30],
                 IIG<-T_IIG_pvalue[1],          TIIG<-T_IIG_pvalue[2],
                 TIIG_p_value<-T_IIG_pvalue[3])
TIIG_Results<-rbind(TIIG_Results,temp_results,stringsAsFactors = FALSE)
temp_results<-c()
        }
        colnames(TIIG_Results)=c("SNP1", "SNP2", "SNP3", "IIG_Value",
        "T_IIG_Value", "T_IIG_p_value")

}
else
{
        TIIG_Results<-data.frame(SNP_1<-as.character(),          SNP_2<-
        as.character(), TIG<-as.numeric(),TIG_p_value<-as.numeric())
        for (i in 1:nrow(result_case))
        {
```

```
                        X1<-c(result_control[i,1],result_control[i,2],result_control[i,3],
                                result_control[i,4],result_control[i,5],result_control[i,6],
                                result_control[i,7],result_control[i,8],result_control[i,9])

                        Y1<-c(result_case[i,1],result_case[i,2],result_case[i,3],
                                result_case[i,4],result_case[i,5],result_case[i,6],
                                result_case[i,7],result_case[i,8],result_case[i,9])

                        T_IG_pvalue = p_value_func_T_IG(X1, Y1)
                         temp_results<-data.frame(SNP_1<-result_case[i,10],  SNP_2<-
                         result_case[i,11],
        TIG<-T_IG_pvalue[1],TIG_p_value<-T_IG_pvalue[2])
        TIIG_Results<-rbind(TIIG_Results,temp_results,stringsAsFactors = FALSE)
                        temp_results<-c()

                }

        colnames(TIIG_Results)=c("SNP1","SNP2","TIG_Value","TIG_p_value")
        }
        write.csv(TIIG_Results,outputPath,row.names=FALSE)
}

implementII(casePath,controlPath,interactionType,outputPath)
```

A.5. R Program we used for performing permutation testing

```
require(tidyverse)

setwd(pathForWD)
file_list <- list.files(path=pathForFiles, full.names =TRUE)
for (i in 1:length(file_list)){
  tmpData<-read.csv(file_list[i])
    tmpData$Status<-as.factor(tmpData$Status)
  case_cont<-summary(tmpData$Status)
  resultPath<- pathForResults
  fileName<-basename(file_list[i])
  outpath<-paste(resultPath,fileName,sep="")
  genotypeData<-tmpData
  TIIG_Results<-data.frame(TIIG<-as.numeric(),TIIG_p_value<-as.numeric())
  for (i in 1:1000)
  {
    caseData<-genotypeData[1:case_cont[1],]
    controlData<-genotypeData[(case_cont[1]+1):nrow(tmpData), ]
    X<-c()
    Y<-c()
    for (j in 1:(ncol(controlData)-1))
    {
      X<-cbind(X,sum(controlData[,j]))
    }
    for (k in 1:(ncol(caseData)-1))
    {
      Y<-cbind(Y,sum(caseData[,k]))
    }
    T_IIG_pvalue <- p_value_func_T_TIIG(X, Y)
    temp_results<-data.frame(TIIG<-T_IIG_pvalue[1],
                             TIIG_p_value<-T_IIG_pvalue[2])
    TIIG_Results<-rbind(TIIG_Results,temp_results,stringsAsFactors = FALSE)
    rows <- sample(nrow(tmpData))
    genotypeData <- tmpData[rows, ]
  }
    perm_pvalue<-(1 + sum (TIIG_Results$TIIG....T_IIG_pvalue.1. >
TIIG_Results$TIIG....T_IIG_pvalue.1.[1])) / 1000
  print(paste(fileName," ",perm_pvalue))
  #write.csv(TIIG_Results, outpath)

}
```

Appendix B Results

Table 10 SNPs Filtered By PLINK-RF-RF

ADNI	GenADA		NCRAD		
rs10017010	rs2273570	rs10022344	rs1489449	rs2588478	rs6494944
rs2131006	rs11810899	rs1004677	rs1530498	rs2781092	rs6506593
rs557098	rs1879019	rs10056788	rs1552673	rs2792752	rs6604921
rs6705017	rs1553136	rs10108019	rs1558139	rs2821338	rs6645551
rs9307850	rs17793957	rs10240034	rs1560964	rs2860001	rs6743559
rs9896368	rs12493090	rs1028723	rs16822383	rs2953130	rs6789329
rs10788181	rs16863803	rs1033686	rs16826325	rs2966220	rs6789642
rs10807701	rs10024098	rs10401458	rs16841113	rs3104398	rs6859
rs11006011	rs11743813	rs10435761	rs16841178	rs31726	rs6887169
rs11749731	rs10050568	rs10445686	rs16861648	rs332847	rs6904295
rs2824808	rs17067596	rs10458404	rs16863430	rs3744587	rs6942330
rs6856771	rs4895529	rs10498554	rs16985682	rs3753306	rs6948502
rs7091014	rs1834615	rs10520536	rs17043361	rs3763696	rs6990287
rs717840	rs9314604	rs10784286	rs17080902	rs3775162	rs7110148
rs9313264	rs11785331	rs10804966	rs17148940	rs3777171	rs7181139
rs2207851	rs2978012	rs10860105	rs17149283	rs3785113	rs7212762
rs462074	rs4073002	rs10877627	rs17247661	rs3790630	rs7247886
rs6751810	rs7045548	rs10877894	rs17519749	rs3806448	rs7309474
rs12056012	rs10795618	rs10922743	rs17576289	rs3816620	rs7314903
rs3967101	rs472186	rs10941112	rs17742907	rs3827466	rs7428284
rs4954943	rs1608169	rs10945826	rs17830067	rs3843725	rs743026
rs7157639	rs1795977	rs11015383	rs1793773	rs3888795	rs7518469
rs9366664	rs1519959	rs11023775	rs1842076	rs4076290	rs7531902
rs1023276	rs17081694	rs11049103	rs1872240	rs4255200	rs761009
rs10960174	rs605928	rs11049111	rs1904616	rs4307816	rs766787
rs1150360	rs2028389	rs11235796	rs1920045	rs4450019	rs7683253
rs2548032	rs7204799	rs1149938	rs1920086	rs4478858	rs7723920
rs2633466	rs11862388	rs11645986	rs19334	rs4498832	rs7726567
rs324389	rs11652714	rs11782215	rs2040761	rs4535265	rs7743868
rs3780792	rs9911460	rs11889431	rs2045896	rs4634829	rs7752524
rs4409091	rs52911	rs11924077	rs2049969	rs4648489	rs7761785
rs4561856	rs10402361	rs12135108	rs2060802	rs4668565	rs7769144
	rs12974182	rs12236795	rs2075650	rs4750306	rs7876155

rs6098412	rs12379827	rs2088746	rs4755903	rs7971558
rs16993582	rs12468084	rs2105806	rs4760479	rs8030257
rs136687	rs12483892	rs2108392	rs4770244	rs837957
	rs12549680	rs2118844	rs4774371	rs839933
	rs12663008	rs2151641	rs4778582	rs8802
	rs12677921	rs2166001	rs4789127	rs882299
	rs12683681	rs2223541	rs4803586	rs920633
	rs12978451	rs2244263	rs4841920	rs931774
	rs13030842	rs2256331	rs495959	rs9367597
	rs13224531	rs2268408	rs4978508	rs943724
	rs1322669	rs2283821	rs544166	rs9454454
	rs1333190	rs2297741	rs5752338	rs9490248
	rs13386681	rs2299776	rs5989589	rs9516391
	rs13416635	rs2301718	rs6074067	rs9580739
	rs1372173	rs2310878	rs6099457	rs961848
	rs1382794	rs2325652	rs6102926	rs9670249
	rs1407469	rs2385507	rs6151630	rs9691167
	rs1437635	rs2420516	rs6449986	rs991974
	rs1439279	rs2432762	rs6455400	rs9957722
	rs1447830	rs2436111	rs6455403	rs9966557
	rs1450913	rs2467261	rs6474161	
	rs1459241	rs248471	rs6480813	

Table 12 Triplets that are found to involve SNP pairs with significant 2WI in NCRAD dataset

SNP1	SNP2	SNP3	TIG	p-value
rs1004677	rs4770244	rs9966557	10.96	0.0009
rs10445686	rs11924077	rs16822383	11.14	0.0008
rs10458404	rs11924077	rs839933	10.89	0.001
rs11924077	rs11235796	rs9670249	11.37	0.0007
rs11924077	rs7110148	rs9670249	11.68	0.0006
rs11924077	rs839933	rs10498554	12.08	0.0005
rs13416635	rs1028723	rs3888795	12.71	0.0004
rs13416635	rs1028723	rs9966557	11.84	0.0006
rs13416635	rs1028723	rs16985682	11.69	0.0006
rs13416635	rs11924077	rs839933	11.04	0.0009
rs13416635	rs11924077	rs10498554	14.52	0.0001
rs13416635	rs4770244	rs9966557	11.90	0.0006
rs16822383	rs11023775	rs1028723	10.98	0.0009
rs16822383	rs12663008	rs17830067	11.82	0.0006
rs2821338	rs6942330	rs9580739	10.86	0.001
rs3775162	rs9580739	rs9966557	10.84	0.001
rs3806448	rs991974	rs16985682	13.60	0.0002
rs4450019	rs991974	rs16985682	12.16	0.0005
rs4770244	rs9580739	rs9966557	12.04	0.0005
rs6449986	rs9670249	rs3888795	11.49	0.0007
rs9580739	rs1028723	rs16985682	11.87	0.0005
rs991974	rs839933	rs16985682	11.49	0.0007

Table 11 SNP Pairs with Significant Interactions in NCRAD Dataset

SNP_1	SNP_2	2WII	p_value
rs4770244	rs9966557	9.901943	0.001651
rs10445686	rs11924077	10.68177	0.001082
rs10458404	rs839933	11.22902	0.000805
rs11924077	rs11235796	12.89142	0.00033
rs11924077	rs7110148	8.756695	0.003085
rs11924077	rs10498554	10.55483	0.001159
rs13416635	rs1028723	9.727701	0.001815
rs13416635	rs11924077	14.65575	0.000129
rs16822383	rs1028723	8.872948	0.002894
rs16822383	rs12663008	11.86114	0.000573
rs6942330	rs9580739	6.885878	0.008688
rs3775162	rs9580739	8.020484	0.004625
rs991974	rs16985682	10.05218	0.001522
rs9580739	rs9966557	9.361782	0.002216
rs6449986	rs9670249	7.873453	0.005017
rs9580739	rs16985682	8.520132	0.003512
rs991974	rs839933	8.748702	0.003098

Table 11 Test Statistics and Permutation Testing Results for NCRAD Dataset

SNP1	SNP2	SNP3	TIG	p-value	Permuation p-value	Gene1	Gene2	Gene3	3WII Sign
rs10056788	rs10108019	rs931774	12.67	0.0003	0.001	NIM1K	ADAM18	SOX6	Positive
rs11924077	rs2108392	rs7971558	12.21	0.0004	0.001	AADACL2-AS1	LYRM7	REP15	Positive
rs12379827	rs837957	rs11645986	10.88	0.0009	0.001			LCMT1	Positive
rs12468084	rs2420516	rs3763696	10.92	0.0009	0.001				Positive
rs12663008	rs7247886	rs4634829	13.63	0.0002	0.001				Negative
rs17043361	rs13030842	rs7309474	11.12	0.0008	0.001	DPP10	RAPGEF4		Negative
rs17148940	rs2297741	rs4307816	10.97	0.0009	0.001	ZNF474	LAMA2		Positive
rs17830067	rs7247886	rs4634829	13.73	0.0002	0.001				Negative
rs19334	rs11023775	rs2299776	11.40	0.0007	0.001	PPP1R3B		PCP4	Negative
rs2385507	rs12379827	rs3888795	17.15	3.45E-05	0.001	ANXA13		GNAL	Positive
rs2420516	rs1560964	rs3888795	10.89	0.0009	0.001		RYR3	GNAL	Negative
rs2860001	rs12683681	rs17742907	12.15	0.0004	0.001		C9orf84	DGCR6	Positive
rs2953130	rs2432762	rs12379827	12.48	0.0004	0.001		FARS2		Positive
rs4076290	rs3775162	rs2060802	12.60	0.0003	0.001	TPO	SLC4A4	ADAM18	Positive
rs6474161	rs839933	rs9966557	14.41	0.0001	0.001	ADAM18			Positive
rs6789329	rs7761785	rs4978508	12.58	0.0003	0.001		HCRTR2	SNX30	Positive
rs9367597	rs1028723	rs4803586	11.11	0.0008	0.001	FAM83B	CHORDC2P	PSG4	Negative
rs10108019	rs839933	rs9966557	14.98	0.0001	0.002	ADAM18			Positive
rs11924077	rs7314903	rs9957722	15.08	0.0001	0.002	AADACL2-AS1	FAM19A2	RNA5SP458	Positive
rs11924077	rs9454454	rs31726	12.40	0.0004	0.002	AADACL2-AS1	FGF12	PLEKHG2	Positive

rs13224531	rs6990287	rs1149938	12.65	0.0004	0.002				Positive
rs17080902	rs17148940	rs991974	12.61	3.82E-04	0.002		ZNF474	LMBRD1	Positive
rs2108392	rs2792752	rs3785113	10.88	0.001	0.002	LYRM7	GJC1	PRMT7	Positive
rs2268408	rs3763696	rs10860105	11.26	0.0008	0.002	MDFI		NEDD1	Positive
rs2301718	rs2420516	rs839933	12.40	0.0004	0.002				Positive
rs3753306	rs2045896	rs6480813	11.55	0.0007	0.002	SELP	C1orf112	C10orf11	Positive
rs4535265	rs1004677	rs839933	12.34	0.0004	0.002	FYCO1	LINC-PINT		Positive
rs4535265	rs1920045	rs4760479	11.64	0.0006	0.002	FYCO1	CBX5/ HNRNPA1		Negative
rs6887169	rs10108019	rs931774	12.42	0.0004	0.002	NIM1K	ADAM18	SOX6	Positive
rs6904295	rs7110148	rs1920086	11.11	0.000858	0.002		SLC6A5	FAM19A2	Negative
rs7428284	rs2953130	rs4770244	12.42	0.0004	0.002	ROBO2		MTND3P1	Positive
rs10804966	rs9966557	rs332847	11.60	0.0006	0.003	EVC		SIPA1L3	Positive
rs16841178	rs7683253	rs5989589	17.80	2.45E-05	0.003				Negative
rs16863430	rs11049103	rs6859	11.02	0.0009	0.003		MRPS35	PVRL2	Positive
rs16863430	rs991974	rs31726	12.27	0.0005	0.003		LMBRD1	PLEKHG2	Positive
rs2420516	rs7726567	rs1920086	12.19	0.0005	0.003			FAM19A2	Positive
rs6455403	rs6990287	rs7876155	11.64	0.0006	0.003			FRMPD4	Negative
rs7181139	rs2966220	rs2223541	11.65	0.0006	0.003	LINGO1		PTPRT	Negative
rs16822383	rs19334	rs12978451	11.81	0.0006	0.004		PPP1R3B	ZNF331	Positive
rs10941112	rs4307816	rs1028723	11.66	6.38E-04	0.005	AMACR		CHORDC2P	Positive
rs9490248	rs931774	rs1033686	11.04	0.0009	0.005		SOX6	RAD51B	Positive
rs10445686	rs17576289	rs2075650	11.90	0.0006	0.006	RAB3GAP1 / SNORA40/ ZRANB3	LARS2	TOMM40	Negative
rs17148940	rs1552673	rs16985682	11.90	0.0006	0.006	ZNF474	ARRDC4		Positive
rs1842076	rs12683681	rs17742907	12.15	0.0005	0.006	-	C9orf84	DGCR6	Positive
rs7743868	rs6990287	rs7876155	11.09	0.0009	0.006			FRMPD4	Negative

rs9490248	rs1437635	rs1033686	11.69	0.0006	0.006		SOX6	RAD51B	Positive
rs11889431	rs6887169	rs3763696	12.10	0.0005	0.009	THSD7B	NIM1K		Positive
rs13386681	rs7181139	rs9957722	11.91	0.0005	0.009	ATOH8	LINGO1	RNA5SP458	Positive
rs16861648	rs3888795	rs9966557	11.18	0.0008	0.009	IGSF21	GNAL		Negative
rs17080902	rs9580739	rs7181139	13.60	0.0002	0.009			LINGO1	Positive
rs2860001	rs9580739	rs6102926	15.13	0.0001	0.009			PTPRT	Positive
rs495959	rs7181139	rs12483892	11.79	0.0006	0.009	SLC6A17	LINGO1		Positive
rs17080902	rs9670249	rs10401458	11.21	0.0008	0.013	SPTBN4			Positive
rs2060802	rs7181139	rs3785113	11.77	0.0006	0.016	ADAM18	PRMT7	LINGO1	Positive
rs13386681	rs7181139	rs3888795	13.98	0.0002	0.018	ATOH8	GNAL	LINGO1	Positive
rs16841113	rs10520536	rs9516391	12.06	0.0005	0.018	TENM3			Positive
rs4478858	rs6474161	rs7181139	10.9	0.001	0.018	SERINC2	ADAM18	LINGO1	Positive
rs16841113	rs10520536	rs4307816	12.4	0.0004	0.02	TENM3			Positive
rs9670249	rs7247886	rs12483892	14.03	0.001	0.021				Positive
rs12135108	rs991974	rs7181139	11.05	0.0009	0.022	MLK4	LINGO1	LMBRD1	Positive
rs17519749	rs6474161	rs7181139	11.09	0.0009	0.022	ADAM18	LINGO1		Positive
rs7428284	rs1530498	rs4255200	11.37	0.0007	0.022	DNAH5	ROBO2		Positive
rs16822383	rs9490248	rs9516391	11.01	0.0009	0.042				Positive
rs13224531	rs6990287	rs2792752	10.91	0.001	0.067				Positive

Table 12 Pathway List in Reactome Analysis of The Triplets

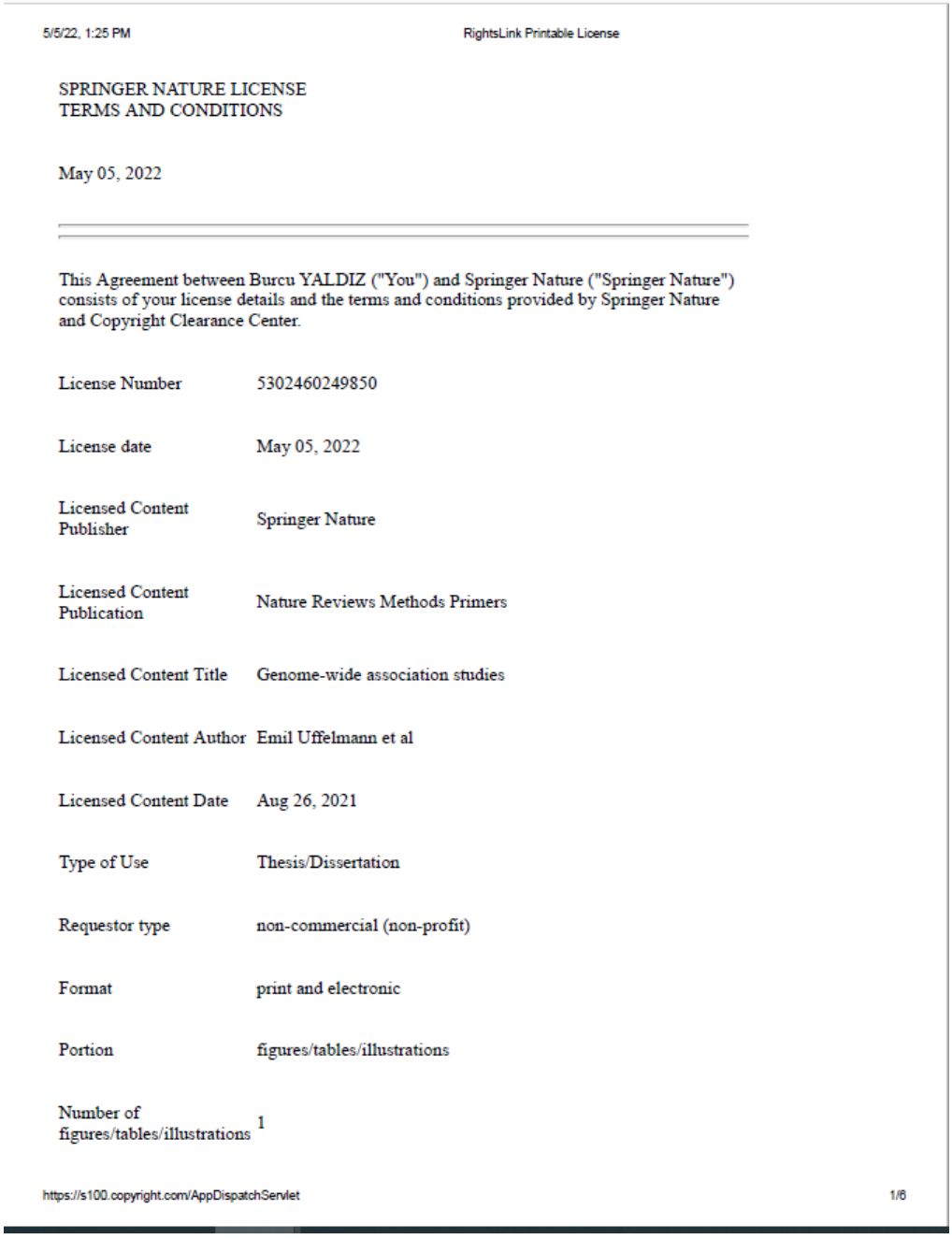
Pathway name	#Entities total	Entities ratio	Entities pValue	Entities FDR	#Reactions found	#Reactions total	Reactions ratio	Submitted entities found	Dataset
Extracellular matrix organization	301	0.026	0.002	0.020	6	319	0.024	ADAM10;FBLN2	GenADA
Post-translational protein modification	1427	0.123	0.042	0.041	4	526	0.039	NPLOC4;ADAM10	GenADA
Metabolism of proteins	1950	0.168	0.075	0.075	7	793	0.059	NPLOC4;ADAM10	GenADA
RAC3 GTPase cycle	94	0.008	1.96E-04	0.010	2	6	4.47E-04	FERMT2;VAV2	ADNI
RAC1 GTPase cycle	185	0.016	7.54E-04	0.019	2	6	4.47E-04	FERMT2;VAV2	ADNI
RHO GTPase cycle	452	0.039	0.004431	0.030	10	91	0.007	FERMT2;VAV2	ADNI
Signaling by Rho GTPases	678	0.058	0.009838	0.030	10	203	0.015	FERMT2;VAV2	ADNI
Signaling by Rho GTPases, Miro GTPases and RHOBTB3	694	0.060	0.010298	0.030	10	212	0.016	FERMT2;VAV2	ADNI
Signal Transduction	2588	0.223	0.126976	0.127	19	2461	0.183	FERMT2;VAV2	ADNI
Metabolism of lipids	752	0.065	0.012049	0.022	5	826	0.061	PI4K2B;ALDH3B1	ADNI
Metabolism	2143	0.185	0.089677	0.090	5	2012	0.150	PI4K2B;ALDH3B1	ADNI
Signal Transduction	2588	0.223	0.126976	0.127	12	2461	0.183	LINGO1;GNAL	NCRAD

This table only represents the pathways which gene pairs of the triplets were enriched in.

Table 13 Genes involved in functional enrichment analysis

AADACL2-AS1	AC093840.1	ANXA13	GPAM	NIM1K	RP11-63E5.6
AC009117.1	AC096992.1	AP003398.2	HCRTR2	NPLOC4	RP11-968A15.2
AC009509.1	AC104662.2	APOOP2	HIF1AN	NPSR1	RPL12P4
AC009869.1	AC105252.1	BX119904.1	HMCN2	NPSR1-AS1	RUNX1
AC010255.3	AC105450.1	C9orf84	HSPD1P15	PCP4	RYR3
AC022784.1	AC106744.1	CBX5	LAMA2	PHF21B	SAMSN1
AC024909.1	AC106800.1	CDH13	LCMT1	PLAGL1	SEPSECS-AS1
AC025252.1	AC117382.2	CTC-210G5.1	LINC00430	PSG4	SLC4A4
AC026396.1	AC123767.1	CTC-441N14.4	LINC00507	PSG5	SNX30
AC026884.1	ADAM10	DGCR6	LINC01376	PSMC1P6	SOX6
AC034268.2	ADAM18	DNAJB8	LINC01440	RAPGEF4	SPARC
AC068787.1	AF127936.2	DNER	LINC02199	RF00017	SPOCK1
AC073592.2	AL024498.2	DPP10	LINC-PINT	RF00019	ST3GAL1
AC073592.8	AL137230.2	EEFSEC	LYRM7	RN7SKP190	SYCP2L
AC079362.1	AL138885.1	FAM19A2	MBD2	RN7SKP93	TAF15
AC079380.1	AL139317.5	FAM230F	MCPH1	RNU1-150P	TEMN3-AS1
AC087379.1	AL161716.1	FAM76B	MELK	RNU6-1323P	TMPRSS15
AC090515.3	AL353595.1	FAM83B	MGAT5	RNU6ATAC21P	TPO
AC091895.1	AL589740.1	FAM87A	MIR4475	RP11-166A12.1	TPST1
AC092100.1	AL589826.1	FARS2	MMP28	RP11-285A15.1	TRIP12
AC092364.1	AL713851.1	FBLN2	MRPS35P3	RP11-309P22.1	UBBP5
AC093277.1	ALDH3B1	FERMT2	NDFIP1	RP11-406O16.1	UGGT1
AC093462.1	ANGPT2	GNAL	NHSL1	RP11-427M20.1	VAV2

Appendix C Copyright Permissions



5/5/22, 1:25 PM RightsLink Printable License

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

May 05, 2022

This Agreement between Burcu YALDIZ ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5302460249850
License date	May 05, 2022
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Methods Primers
Licensed Content Title	Genome-wide association studies
Licensed Content Author	Emil Uffelmann et al
Licensed Content Date	Aug 26, 2021
Type of Use	Thesis/Dissertation
Requestor type	non-commercial (non-profit)
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1

https://s100.copyright.com/AppDispatchServlet 1/6

Figure 10 Copyright permission of the image used in Figure 1

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

May 13, 2022

This Agreement between Burcu YALDIZ ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5307090588006
License date	May 13, 2022
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Genetics
Licensed Content Title	Detecting epistasis in human complex traits
Licensed Content Author	Wen-Hua Wei et al
Licensed Content Date	Sep 9, 2014
Type of Use	Thesis/Dissertation
Requestor type	non-commercial (non-profit)
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1

Figure 11 Copyright permission of the image used in Figure 3

CURRICULUM VITAE

Surname, Name: Yaldız, Burcu

Nationality: Turkish

Date and Place of Birth: 01-02-1984 / Polatlı

Phone: +31 680211706

email: yaldizburcu@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Informatics Institute/ Bioinformatics	2014
BS	Ankara University/ Mathematics	2009
High School	Gazi Anatolian High School	2002

WORK EXPERIENCE

Year	Place	Enrollment
2020-2022	RadboudUMC	Research Scientist

- Coordination of exome/genome data for SolveRD Project
- Performing sequencing data analysis and reporting the results for SolveRD Project
- Authoring the publications

2018-2020	TÜSEB	Bioinformatics Specialist

- Tracking and inspecting bioinformatics analysis platform development and bioinformatics analysis progress of Turkish Genome Project (TGP)
- Taking part in Genome Project's bioinformatics infrastructure installation works

2016-2018	Akgün Software	Researcher & Software Specialist

- Machine learning and data mining techniques have been applied for a decision support system which will be used for the interpretation of results of biochemistry tests applied in hospitals. (TÜBİTAK Project)
- Supporting other projects on statistics and data mining issues
- Conducting detailed field research for proposing R&D projects and contributing to the preparation of project files.
- Preparing project term reports for referee evaluation
- Participating maintenance work of Mobile Hospital Information System (HIS) platform as a developer
- Participating requirement development process in accordance with CMMI
- Performing functional testing and resolving errors

2013-2016	Genformatik	Bioinformatician

- Working as a bioinformatician in BeeMonitor Eurostars Project.
- Performing bioinformatics analysis for developing test kit for the rapid diagnosis of viral infections that cause colony collapse disorder in bees.

FOREIGN LANGUAGES

Native Turkish, Advanced English, Beginner Dutch

SKILLS

- Systems: Windows, Linux
- Languages: Python, R, Java, PL-SQL, JavaScript (ExtJS)
- Database: Oracle SQL Server
- Development Process Management: CMMI, IBM Jazz Tool
- Project Management Tool: JIRA
- Version Control Systems: SVN, Git
- Data manipulation and analysis
- NGS Analysis

PUBLICATIONS

- de Boer, E., Yaldiz B., Denommé-Pichon, A-S., Matalonga, L., Laurie S., Solve-RD SNV-indel working group, Solve-RD-DITF-ITHACA "Genome-wide

variant calling in reanalysis of exome sequencing data uncovered a pathogenic TUBB3 variant." European journal of medical genetics vol. 65,1 (2022): 104402. doi:10.1016/j.ejmg.2021.104402

- Yaldiz,B.,Erdogan,O.,Rafatov,S.,Iyigun,C.,Aydin Son,,Y., Integration of Machine Learning and Entropy Methods for Post-Genome Wide Association Studies Analysis, Biodatamining (Under Review)

PRESENTATIONS AND POSTERS

- Yaldız,B., Atasoy, Y., Karatas,H., 2017, Sağlık Hizmetlerinde Veri Madenciliği Önişleme Süreçlerinde Karşılaşılan Sorunlar ve Olası Çözüm Yolları, 10th International Congress of Medical Informatics, Antalya
- Yaldız, B., Aydın Son, Y., 2017, Structural Properties Of Methylated Human Promoter Regions In Terms Of DNA Helical Rise, 10th International Symposium on Health Informatics and Bioinformatics, Güzelyurt-KKTC
- Özkan, Ö., Aydın Son, Y., Aydınoğlu, A.U., Yalçın, A., Döm, A., Yaldız, B., Narcı, K., Baloğlu, O., 2017, Participatory Design Meetings: Genetic Information Included Personal Health Record Application, 10th International Symposium on Health Informatics and Bioinformatics, Güzelyurt-KKTC
- Yaldız, B., Son, Y.A. 2013. Investigation of Topological Properties of Human Promoter CpG Islands, 8th International Symposium on Health Informatics and Bioinformatics, Ankara-Turkey
- Hashemikhabir, S., Ersoy, G., Oguz, G., Yaldiz, B., Tuncel, Y., Budak, G., Karaaslan, S., Son, Y.A., Can, T. 2012. M4B: A Novel Method for Designing and Ordering of the Genetic Devices, 7th International Symposium on Health Informatics and Bioinformatics, Nevşehir-Turkey

PRIZES

- iGEM (The International Genetically Engineered Machine Competition) 2011, Silver Medal, VU University, Amsterdam
- iGEM (The International Genetically Engineered Machine Competition) 2011, "Best Use of Parts Registry" Prize, MIT, Boston