Fall 2023

Identifying Potential Alzheimer's Disease Biomarkers Beyond Amyloid-Beta and Tau

Frank Cai

Identifying Potential Alzheimer's Disease Biomarkers Beyond Amyloid-Beta and Tau

A Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Frank Cai

October 2023

The Designated Project Committee Approves the Project Titled

Identifying Potential Alzheimer's Disease Biomarkers Beyond Amyloid-Beta and Tau

by

Frank Cai

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

October 2023

Dr. Leonard Wesley	Department of Computer Science
Dr. Wendy Lee	Department of Computer Science
Dr. William Andreopoulos	Department of Computer Science

ABSTRACT

Identifying Potential Alzheimer’s Disease Biomarkers Beyond Amyloid-Beta and Tau

By Frank Cai

Alzheimer's Disease (AD) and other forms of Mild Cognitive Impairment (MCI) affect millions of people around the world. The buildup of Amyloid-Beta (Aβ) and Tau proteins in the brain produced by amyloid precursor protein (APP) has been identified as an important cofactor in the onset and progression of AD. However, although patients diagnosed with AD exhibit Aβ and Tau buildup, about 40% of the subjects with Aβ and Tau buildup are not diagnosed with AD. In this project, we hypothesize the involvement of other epigenetic interactions between APP and related genes in addition to the buildup of Aβ and Tau that might explain the onset and progression of AD. A robust and systematic methodology is applied to identify potential epigenetic biomarkers of AD. Single Nucleotide Polymorphisms (SNP) mutated proteins are considered in this study. A novel integrated epigenetic computational pipeline is implemented for SNP protein sequence generation, protein structural-functional change prediction, statistical analysis, and identification of significant SNPs associated with AD. These significant SNPs warrant further investigation as potential biomarkers linked to AD.

Keywords: Amyloid-Beta and Tau; Alzheimer's disease; significant single nucleotide polymorphisms; epigenetic interactions; biomarker; pipeline.

ACKNOWLEDGMENTS

The author wishes to express his sincere appreciation to Dr. L. Wesley for his encouragement, expert guidance, and supervision throughout the course of this project. Special thanks are due to Dr. W. Lee and Dr. W. Andreopoulos, members of his Supervisory Committee, for their valuable suggestions and continuing support. The author is very grateful to his classmates Shamika Majmudar and Rheya Mirani for their valuable suggestions and discussions as well as their contributions to the epigenetic pipeline used in this project.

The data for this study was provided by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine at the National Institute of Health. The Alzheimer's patient data was provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI), a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer’s disease. The Python programming language used to construct the integrated pipeline is provided by the Python Software Foundation. Additional data was provided by the NHGRI-EBI Catalog of human genome-wide association studies, or GWAS.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

AD affects more than 6 million Americans in 2023 reported by Alzheimer Association [1]. Its symptoms include a gradually worsening memory, to the point where the person affected can even forget his or her own identity. In recent years the number of people having AD is increasing and this number is projected to rise to an estimated 14 million people by 2060 in United States of America (USA) [2]. The cause of this disease is long thought to be a slow buildup of Aβ proteins [3, 4] and Tau protein [5, 6] within certain regions of the brain, which affects how the neurons in these regions communicate with each other [7, 8]. However, recent researchers have discovered that AD is a complex multifunctional disease and there are many bio-epigenetic relationships between other genes and AD beyond Aβ and Tau [9 - 13].

AD is only one of the many neurodegenerative diseases, such as Parkinson's disease, Huntington's disease, amyotrophic lateral sclerosis (ALS), and motor neuron disease. All these diseases are resulted from functional degradation or death of the nerve cells in the brain or peripheral nervous system. The possibility of developing a neurodegenerative disease rises dramatically with age. As life expectancy increases due to modern healthcare systems and medications more and more people will be affected by neurodegenerative diseases.

According to the special report from Alzheimer Association [1], 1 in 3 American seniors die with AD or another related dementia in 2023 [1]. Dementia is an overall term for a particular group of symptoms including problems with memory, language, problem-solving and other thinking skills. These symptoms are common in neurodegenerative diseases. AD and other dementias kill more people than breast cancer and prostate cancer combined [1]. During the last decade, deaths from heart disease have decreased 7.3% while deaths from AD have increased 145% [1]. In 2023, AD and other related dementias will cost the USA estimated $345 billion [1]. Over 11 million Americans provided unpaid care for people with AD or other dementias, which valued at nearly $340 billion [1]. It is projected that the cost will increase to $1 trillion in 2050 in the USA [1].

There are many more people who suffer from AD and related dementias globally [14]. According to the statistics reported by Alzheimer's Disease International (ADI) [14], over 55 million people worldwide are currently living with AD or another dementia [14]. More than 10 million new cases of dementia are developed each year worldwide with a new case every 3 seconds

[14]. The total number of cases are projected to reach 139 million in 2050 [14]. The total estimated worldwide cost of dementia was US$ 818 billion in 2015, which was equivalent to 1.09% of global GDP at that time [14]. The annual global cost of dementia is now above $ 1.3 trillion and is expected to rise to $ 2.8 trillion by 2030 [14].

Research on AD and other dementias is urgent since there are many people who are affected by such disease and the economic impacts are enormous. To address this, the National Alzheimer's Project Act (NAPA) (Public Law 111-375) was signed into law in the USA on January 4, 2011. The goals of NAPA are to create plans, improve early diagnosis, coordinate research and services, accelerate the development of treatments, and coordinate with international bodies to fight AD globally [15]. The World Health Organization (WHO) had also created a global action plan on the public health response to AD and related dementias in 2017 [16]. The plan comprises seven action areas including dementia research and innovation to improve the lives of people with dementia.

Although certain treatments may help relieve some of the physical or mental symptoms associated with neurodegenerative diseases, no cures exist [1]. The U.S. Food and Drug Administration (FDA) recently approved two drugs (aducanumab and lecanemab) for the treatment of AD [17, 18]. However, these mediations may not have benefits for some individuals; could result in serious potential side effects; and require complicated monitoring procedures [1]. We must improve our understanding of what causes AD and other neurodegenerative diseases to develop new approaches for treatment and prevention.

One of the important research areas is the identification of biomarkers in AD and related dementia. Biomarkers play a critical role in drug development, diagnosis assistance, target engagement, disease modification support, and safety monitoring [19]. The identification and validation of biomarkers for AD and related dementia are increasingly important with the advancement in biotechnology [19, 20]. The correct identification of biomarkers can accelerate the development of new diagnosis methods, medications, treatment therapies, and prevention tools.

Although AD and MCI are typically caused by the excessive buildup of Aβ and Tau proteins produced by the APP gene in the brain [11], excessive levels of Aβ and Tau proteins are not sufficient, in and of itself, for AD diagnoses. There are up to 40% of normal individuals showing high levels of Aβ and Tau [21, 22]. There must be other mechanisms, such as genetic and epigenetic interactions between other genes and the genes producing Aβ and Tau proteins,

which are additional factors causing AD. On the other hand, having a buildup of both of Aβ and Tau protein is necessary for diagnosing AD [23]. AD and MCI patients do not exhibit just a buildup of only Aβ or only tau protein, both are needed [23]. MCI is a condition in which people have more memory or thinking problems than other people their age. The symptoms of MCI are not as severe as those of AD or related dementia. However, people with MCI are at a greater risk of developing AD or a related dementia than people without MCI [24].

The focus of this study is on a systematic scheme to identify other mechanisms or biomarkers in AD and related dementia. Software tools are developed to facilitate large-scale bioinformatic data processing. A substantial number of SNPs from a few hundred genes related to AD and MCI are studied.

CHAPTER 2 BACKGROUND AND RELATED WORK

2.1 Background

2.1.1 Plaque Buildup

The Aβ and Tau protein buildup is caused by a cascade of mutated proteins. This process is incredibly complicated. A study by Aswathy et. al [25] goes into depth about how the plaque buildup happens in an AD patient. The formation of Aβ protein is a 2-step process, with the cleavage of APP by the BACE1 enzyme to form a β-secretase derived fragment of APP, followed by an action of γ-secretase to generate Aβ isoforms ranging from 37 to 42 amino acid residues. Although $A\beta_{40}$ isoform is the most abundant isoform, aggregated $A\beta_{42}$ isoform are associated with AD [25]. The BACE1 gene is just one of the many genes involved in the $A\beta_{42}$ production. The pathway of Aβ protein synthesis which creates the $A\beta_{42}$ is very complex. LaFerla et al [26] presented a cellular Aβ production site graphically as shown in Figure 1. From Figure 1, we can clearly see that there are many genes and pathways which have contributions in Aβ production within the endoplasmic reticulum (ER) and Golgi system. APP affects the plasma membrane by releasing sAPPα into the extracellular space, which results in C83 within the membrane. Reacting with SORL1, APP is recycled back to the Golgi in retromer endosomes. BACE1 cleavage of APP produces C99 which can turn into Aβ within the endosome/lysosome system. Aβ can be bound to cell surface receptors such as LRP, LDL, RAGE, FPRL1, NMDA receptors and α7nAChR, which are adopted into early endosomes. Aβ accumulates in the multivesicular body, lysosomes,

mitochondria, ER, Golgi and the cytosol. This is a complicated procedure involving a number of genes and through different mechanisms.

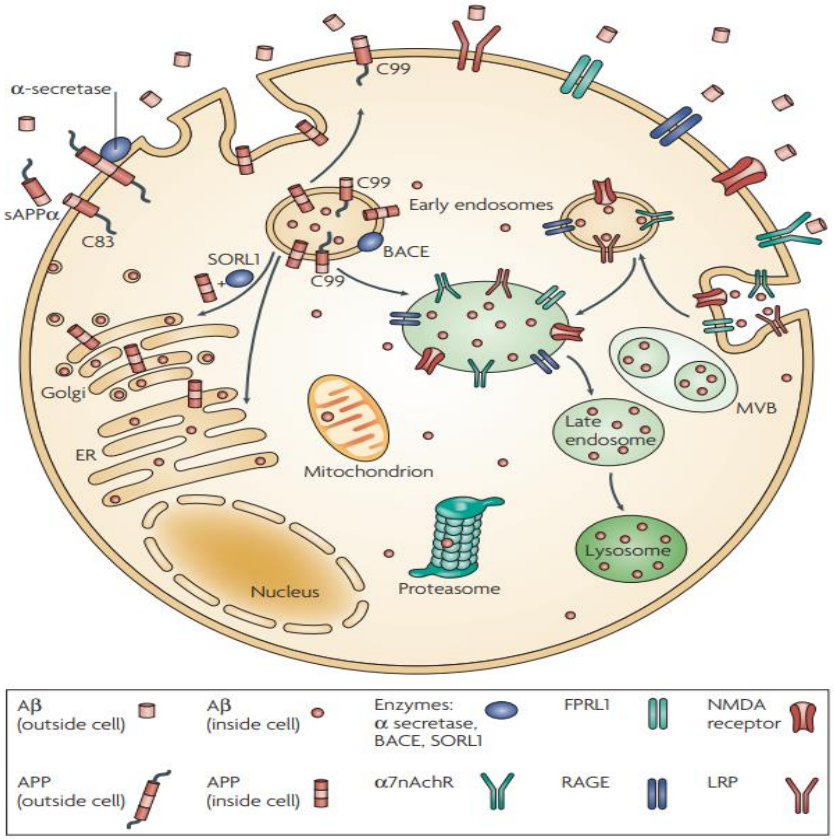

Figure 1. Sites of cellular Aβ production [26].

Furthermore, Aβ produced intracellularly or taken up from extracellular sources, has various pathological effects on cell and organelle function [26]. The intracellular Aβ can assemble itself into oligomers, which mediate pathological events in the cell. The oligomers will further assemble into plaques as shown in Figure 2 [26]. These plaques build up in nerve cells and in the extracellular matrix, blocking the chemical signals needed to transfer information through the brain as shown in Figure 3 [27]. The plaque buildup with excessive Aβ could facilitate Tau hyper-phosphorylation, disrupting proteasome and mitochondria function, and triggering calcium and synaptic dysfunction of the neuron and may cause the neuron to die out completely [26]. Eventually, with a certain amount of dysfunctional and dead brain nerve cells, it leads to MCI and AD.

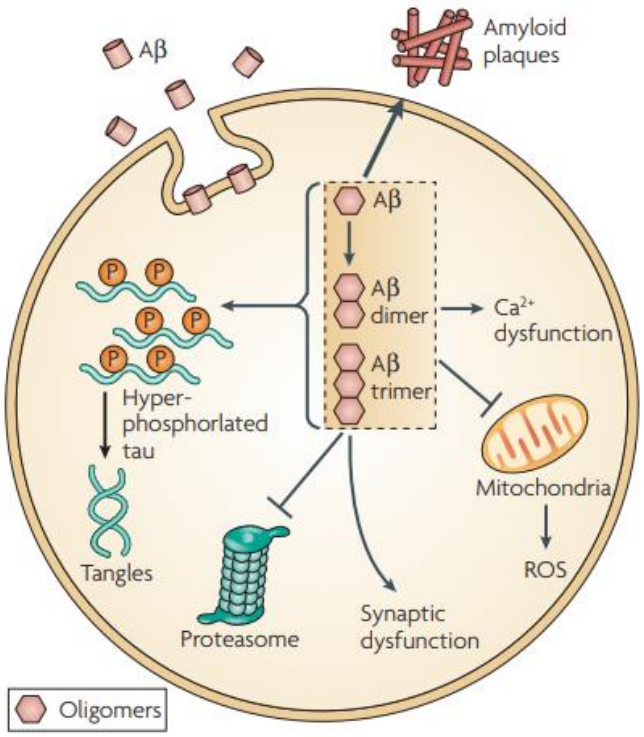

Figure 2. Pathological effects of intraneuronal Aβ [26].

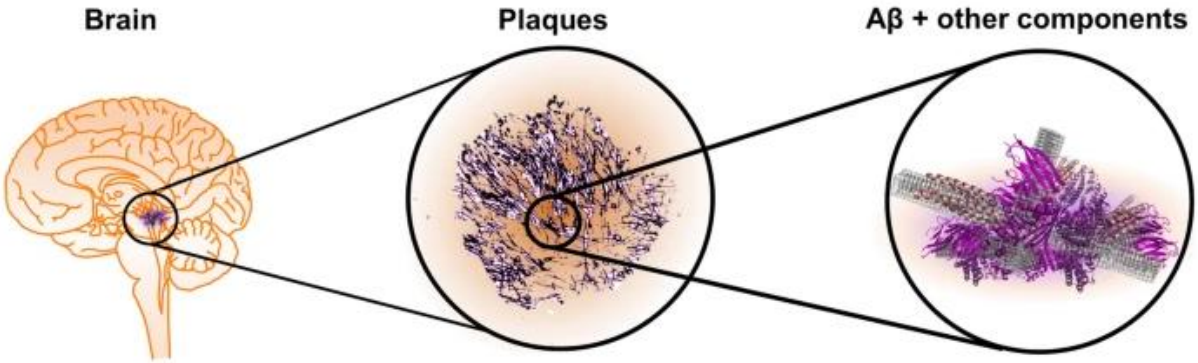

Figure 3. Plaque buildup in the brain from the overproduction of Aβ protein [27].

2.1.2 Epigenetic Mechanisms

Studying epigenetic mechanisms can provide insight into how different genes relate to each other and can uncover previously unknown factors in genetic relationships. The buildup of Aβ and Tau proteins in the brain is the primary cause of AD, but it is not the only cause [9 – 13]. The many

genes that have a relationship with the AD metabolic pathway contribute to the complexity of identifying possible causes of MCI and AD. Furthermore, irregularities with the regulation of these genes can contribute, along with or in combination with direct mutations, presents a formidable challenge to understand MCI and AD pathologies. There are many epigenetic mechanisms that can link to AD as shown in Figure 4 [28].

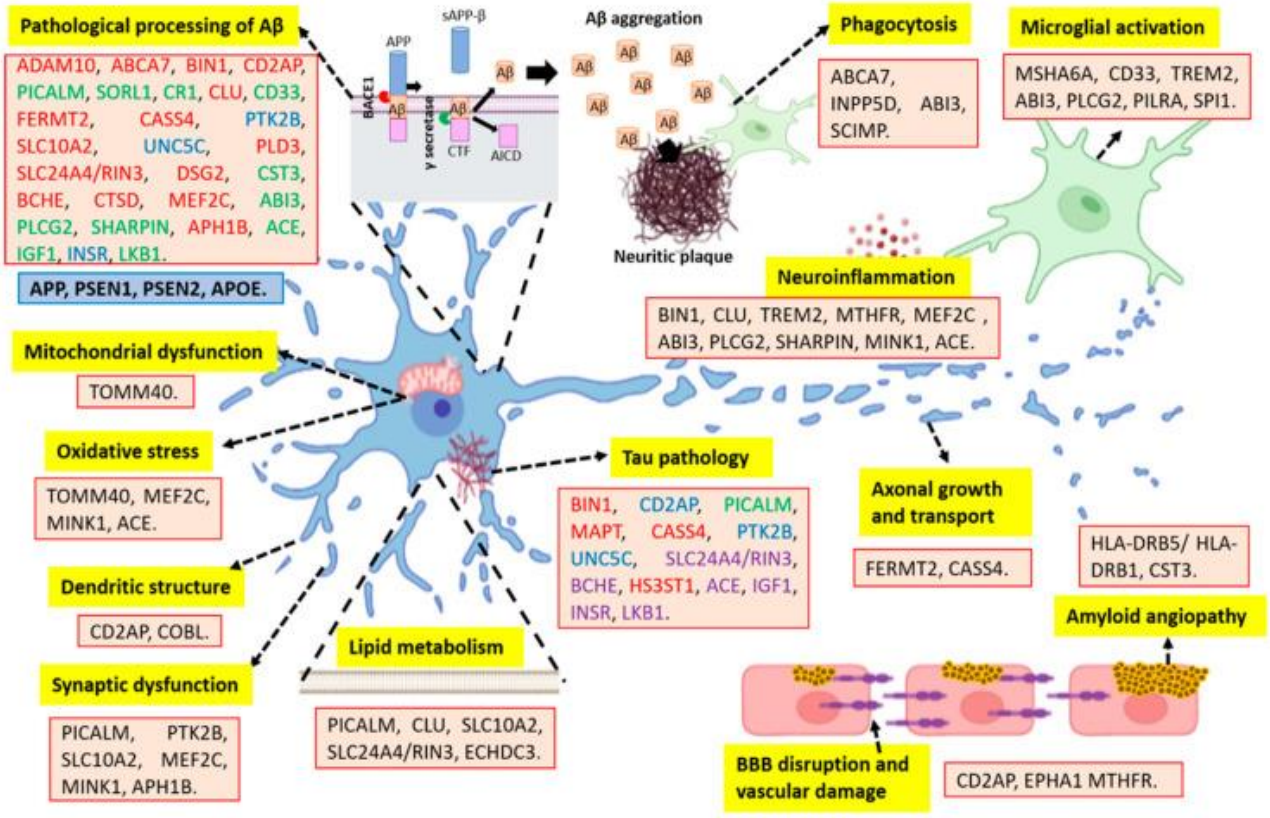

Figure 4. Epigenetic mechanisms linked to AD [28]. Epigenetic factors are in the yellow boxes. Classical AD genes are in the blue box. Genes related to sporadic AD are in the red boxes. Genes increasing production/aggregation of Aβ or Tau are colored in red. Genes cleaning Aβ or Tau are colored in green. Genes for Aβ- or Tau-mediated neuronal damage are colored in blue. Genes for Tau phosphorylation are colored in purple.

The major metabolic pathways of Aβ and Tau can be shown as Figure 5 [29]. Mutations in BACE1 and BACE2 change the regulation of APP to cause Aβ generation. Changes in GSK3B, CDK5, and MARK1 can result in Tau kinases and contribute to Tau pathology. The Aβ and Tau buildup will eventually lead to neuronal death and cause AD.

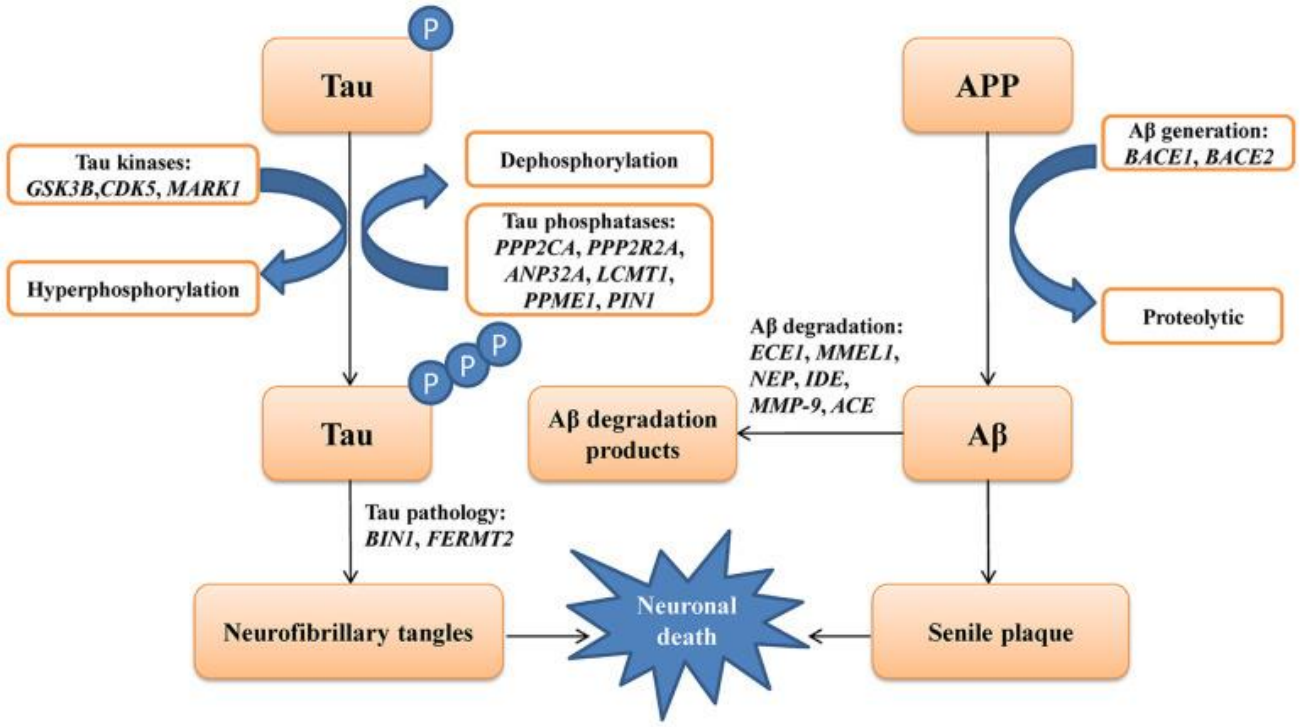

Figure 5. Genes involving the metabolic pathways of Aβ and Tau protein in AD [29].

The work in this project focus on exploration of genes epigenetically associated with classical AD genes based on "Significant" SNP (SSNP) mutations. The meaning of "significant" SNP mutations is mutations that are predicted to change the shape and function of the gene versus SNPs that do not change a protein's function. The study involves investigating if and the extent that SSNP mutations of genes having an epigenetic relationship to the APP, Tau, and other AD-related genes can become biomarkers to the onset and progression of MCI or AD.

There are many types of mutations that change the composition of a given sequence, from small scale changes that affect only a few nucleotides to duplications of entire chromosomes. These mutations can have a positive, negative, or no impact on the structure and function of the expressed protein. The most basic type of mutation is the SNP which is the change of a single nucleotide in the sequence. Many SNPs in the human genome occur in non-coding regions of DNA. Non-coding regions make up about 90 percent of the human genome [30]. Non-coding DNA is not translated into protein or other molecular structures, so they are usually ignored when studying mutations [30]. However, there are many SNPs that do lie within the coding regions of DNA. These SNPs have the chance of affecting the structure and function of the proteins that the

sequence codes, and these are the most studied. There are many other types of mutations that are also important, including deletion, insertion, and translocation mutations. These are currently ignored for the purpose of this study. However, these types of mutation can certainly be used in future studies.

There might be many explanations why 100% of AD subjects exhibit a buildup of both Aβ and Tau. On the other hand, there might also be many explanations why ~40% of subjects that exhibit a buildup of both Aβ and Tau appear and function normally and are not diagnosed with AD or MCI. One possible explanation might be epigenetic related. In order to explore such potential epigenetic explanations, an important part of the research is to identify the genes that are involved with or related to the APP and APOE physiology through biochemical pathways and networks. Thus, it is critical to investigate if, which, and how "significant" mutations of related genes are consistent with observed clinical results. Identification of potential AD/MCI SNP-related biomarkers using statistical significance tests is indispensable in AD and MCI research.

2.1.3 SNPs and Protein Function Change Prediction

It is a well-established fact that mutations, alongside other environmental factors, are a driving force of evolution. However, mutations in one's DNA can also lead to one of many debilitating genetic diseases and syndromes. One or more SNPs could change the shape or the function of a protein which play an important role in mutations. Mutations are essential to evolution and every genetic feature in every organism was, initially, the result of a mutation [31]. Therefore, understand the shape and functional changes of the protein are the key to explore the evolution changes in any organism.

The amino acid sequence defines the protein's shape [32] which is the main source for its function. An alteration in the amino acid sequence of the protein can lead to changes in its folding and stability property [33, 34], its interaction with other molecules [35, 36], its functional levels [37], and even its overall function. A mutation in the amino acid sequence may alter the protein shape but may not change its function [38]. Accurate prediction of the changes in protein shape and function from SNPs is critical in the research of biological evolution. Lee et al [39] had reviewed some of the methodologies in protein function prediction in the field of computational biology. Sequence-based function prediction is one of the important technologies in protein function prediction, which uses the gateways provided by the NCBI and the European Molecular

Biology Laboratory–European Bioinformatics Institute (EMBI–EBI). These resources provide crucial bioinformatic data including protein and domain family information, functional sites and function prediction methods. They also provide a list of links of related information corresponding to a protein accession code, gene name or similar term. There are many resources and tools have been developed for the application of the database, including Entrez, an integrated literature and molecular databases, BLAST, a sequence similarity search service, VAST, a structure similarity searches, Cn3D, a 3D structure viewer, and Genome, a workbench standalone sequence analysis annotation platform, used in NCBI. However, no sophisticated tools exist for predicting the protein change from its coding gene and SNPs.

2.1.4 Networks and Pathways

The $A\beta_{42}$ synthesis pathway is vast and complicated, with up to 430 genes being involved some way or the other as found out by Hu et. al [13]. Using the human genetic association studies deposited in the PubMed database, they were able to gain a collection of 430 genes that were associated with AD. From there, they used the WebGestalt and ToppGene software to analyze the biological relationships between the genes and built a crosstalk network by calculating the overlap coefficient and the Jaccard coefficient. A network diagram was plotted using the Steiner minimal tree algorithm as shown in Figure 6. The network contains 496 vertices and 1521 lines, with the vertex color designating its corresponding degree under the background of the human protein interactome. The circular genes correspond to the Alzgset database, and the triangular vertices are the genes that are expanded in their study. It shows visually the interconnects between the genes.

Another way to explore the genes related to AD is through the biological pathways. There are many pathways that relate to AD, as shown in Figure 7 [13]. These include apoptosis, oxidative phosphorylation, and ovarian steroidogenesis, amongst others. The genes from the network all belong to one of these pathways. For example, the genes BDNF, CAMK2D, GSK3B, and IRS1 belong to the Neurotrophic signaling pathway while the genes COL11A1, EFNA5, EIF4EBP1, FGF1 and GNB3 belong to the PI3K-Akt signaling pathway.

A few genes, including epidermal growth factor receptor (EGFR), nuclear respiratory factor 1 (NRF1), somatostatin receptor 2 (SSTR2), and sortilin 1 (SORT1), were already shown to be related to AD in previous studies.

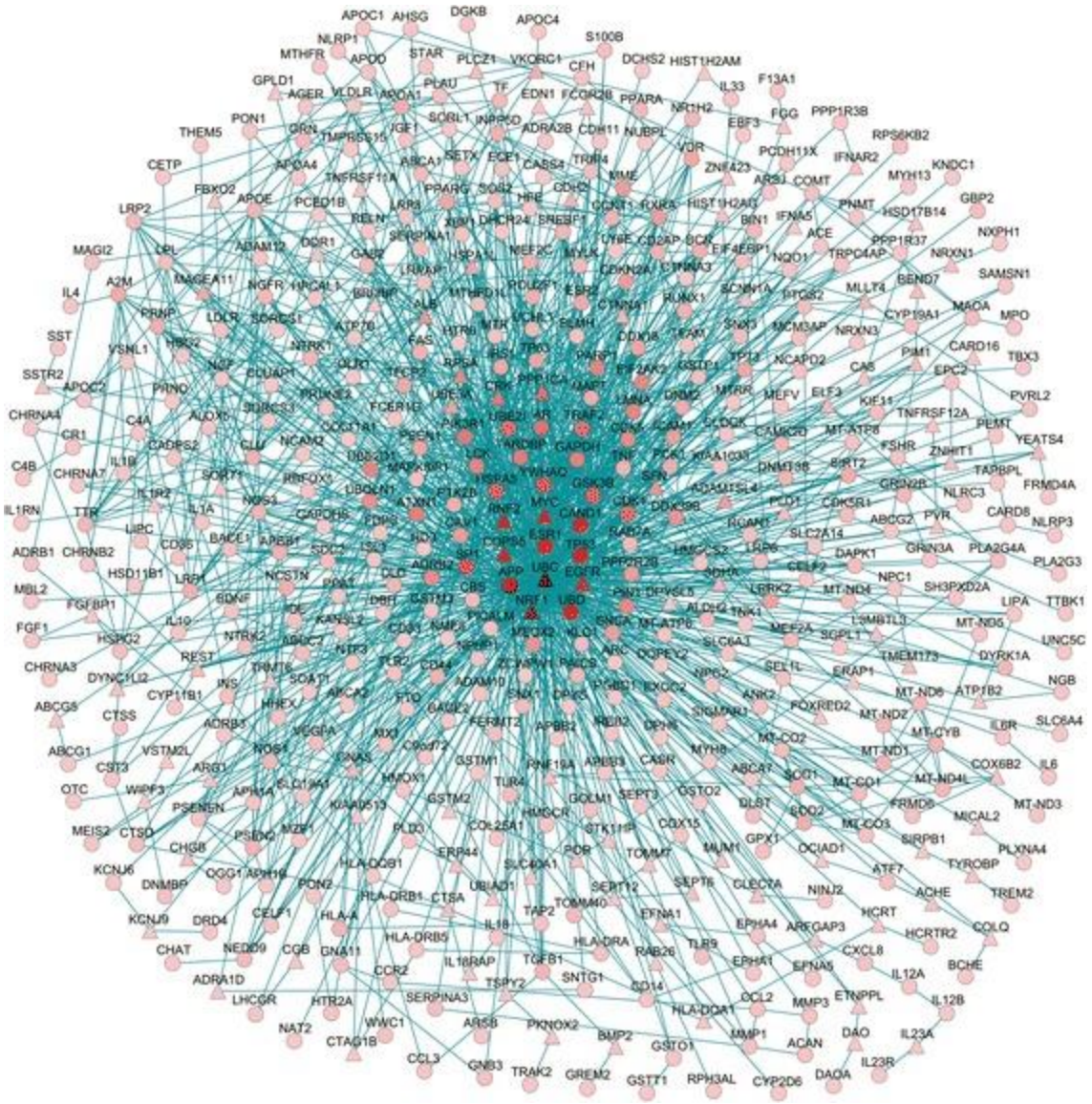

Figure 6. Network diagram of AD genes via the Steiner minimal tree algorithm. Circular vertices, genes of Alzgset; triangular vertices, expanding genes. Color of a typical vertex designates its corresponding degree under the background of the human protein interactome. Darkness of color for a vertex is directly proportional to the corresponding degree value [13].

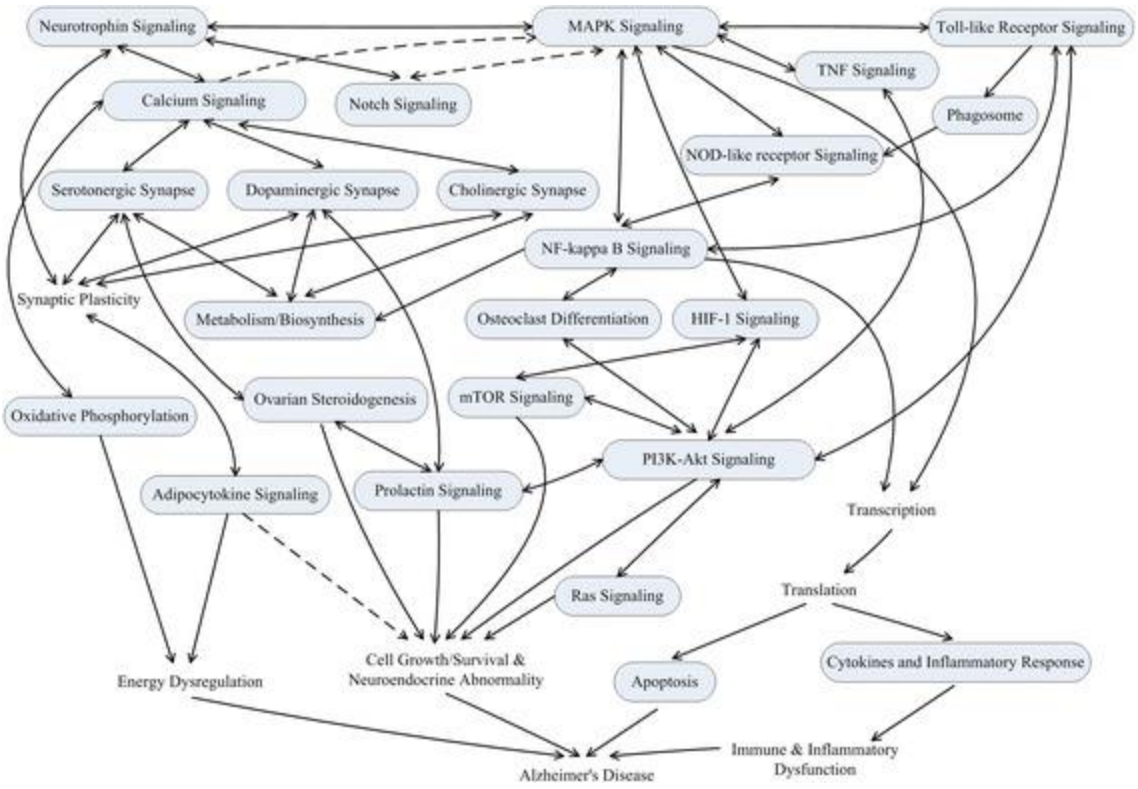

Figure 7. Pathways and processes of AD [13].

Most of the genes are not directly related to AD, but some genes did have direct involvement. Those that are indirect are linked to AD by other genes in the same family. For example, ABCG5 is not directly related, but the genes ABCA1, ABCA2, ABCA7, ABCC2, ABCG1, and ABCG2 are in the same family. These genes are all directly related to AD. Other genes in the network, like HAMP1, are involved in the pathway, but are supplementary to the main genes.

2.2 Related Work

Pimplikar [9] pointed out that there are a significant number of normal individuals who have an excessive buildup of Aβ plaque. Morris et al. [10] found that APOE is a risk factor for AD but it is not a necessary factor. Some individuals clinically diagnosed with AD do not have Aβ pathology. By studying several SNPs on chromosome 6, Kim et al. [11] discovered that 15 genes associated with the glutathione pathway play an important glutathione role in AD pathogenesis. These genes include TMEM14A, RPS16P5, MLIP-IT1, MIR5685, MCM3, LRRC1, GSTA7P,

GSTA5, GSTA3, GSTA2, GFRAL, GCM1, FBXO9, FAM83B, and ELOVL5. Among them, the above GSTA<X> genes express glutathione proteins. They suggested that the mutations produced by epigenetic mechanisms can negate glutathione function and could be a significant factor in AD.

Giri et al. [12] studied the genes associated with late-onset AD and analyzed the relationship between these risk genes and the neuropathologic features of AD. They reviewed the revolutionizing bioinformatic technologies, such as Genome-Wide Association Study (GWAS) and Next-Generation Sequencing (NGS), which play important roles in the study of AD. 28 genes with their corresponding SNPs are studied for their relationships to AD. These genes were categorized into 3 pathways: inflammatory response, lipid metabolism, and Endocytosis.

Hu et al. [13] performed a comprehensive and systematic analysis of 430 human genes reported to be associated with AD from 823 publications. They focused on the biological function and interactions of these genes in the context of AD using network and pathway-based methodology. They systematically explored the pathogenetic mechanism underlying AD. A framework they proposed can be applied for investigating the pathological molecular network and genes relevant to other complex diseases or phenotypes.

A recent two-stage genome-wide association study had been performed by Bellenguez et al. [21] on 111,326 clinically diagnosed 'proxy' AD cases and 677,663 controls. The information used is from GWAS catalog including the location of the gene, the pathways related to the gene, and all the mutations that can be found in the gene. 42 new loci among the 75 risk ones were found related to AD. The study also found some new genes that previously did not have a relationship with AD, like PRKD3.

All these prior and recent studies have suggested that Aβ and Tau are not sufficient to cause AD, but it is necessary. Rather, the pathology of AD is linked to many other genes associated through different pathways. Current research results have discovered many genes and SNPs that are linked to AD in a variety of ways. To find an effective early detection and prevention treatment schemes for AD, epigenetic mechanisms underlying AD pathogenesis and the potentially implicated pathways appear to warrant investigation. In addition, advanced technologies and software tools are indispensable in future research which need to process huge amounts of information from gigantic databases.

CHAPTER 3 APPROACH

3.1 Essential SNP Investigation

Most of the prior studies were performed on a small number of SNPs which limited the scope of research [9-13, 40]. The work presented here focuses on an approach that involves identifying as many "Significant" SNPs (SSNPs) as possible in epigenetic-related genes that are linked to AD and MCI. The SSNP is defined as a SNP which results in Protein Structure and Function Change (PSFC) due to its nucleotide change in the DNA sequence.

SSNP will be used as an important ID to identify epigenetic-related genes that might be associated with AD and MCI along with the necessary data. Genes used in this project are collected from related research and development publications. SNPs associated with each gene are obtained from the NCBI database. These SNPs are processed by the software modules implemented to detect SSNPs. Statistical analyses will then be performed on these SSNPs against the ADNI database to identify differentiating genes and SNPs.

To study the deeper relationships of the selected AD genes, more research needs to be done involving the SNPs that can potentially affect those genes. For a specific gene, hundreds of thousands of SNPs could be found associated with it. All these SNPs should be studied to explore the epigenetic relationship between the gene and AD. This results in a large amount of data that needs to be processed. Therefore, a systematic approach is essential in this study. Well-established biological databases consisting of enough information are also critical. Established biological databases are also more trustworthy as well. Furthermore, effective software tools which can automatically perform many heavy identification, complex manipulation, and large number of calculations tasks are indispensable to the success of this project.

3.2 Data

There are a few databases, including ADNI, NCBI, and Protein Data Bank (PDB), that are used to get information about AD related genes. The ADNI database contains a wealth of information about the SNPs that appear in AD patients. The information includes the position of the SNP and what changed it had occurred, amongst other data. The names of the patients and the SNP ID can be extracted and used to make comparisons between the wild-type and mutated sequences of the proteins that are coded by the genes. To do this, a large database of genes and

their respective proteins will need to be accessed. That is the place where the NCBI database comes into play. The NCBI database contains huge amounts of genes, SNPs, DNA and protein sequences, and much other information as needed for the study. To verify the prediction accuracy from the pipeline, it is necessary to have a database which can generate the biological identity for a specific protein sequence. The PDB database is applied for this purpose. The PDB database is used to compare the physical properties of the wildtype and modified SNP sequences of the protein. All of these three databases are well-established and widely used by the researchers in the bioinformatic community. They are also essential in this project.

3.3 Data Collection

Based on the genes presented in [13] together with some other genes from a few publications [11, 12], 219 genes related to AD in various ways are selected for this study. These genes are listed as follows.

A2M, ABCA1, ABCA2, ABCA7, ABCC2, ABCG1, ABCG2, ADAM10, ADAM17, ADRB1, ADRB2, AKAP9, ALOX5, ANK3, APH1A, APH1B, APOA1, APOE, APP, ATP8B3, BDNF, BIN1, BLNK, BMPR1B, C4A, C4B, CAMK2D, CARD8, CASS4, CCDC6, CCL2, CCL3, CCR2, CD14, CD2AP, CD33, CD36, CDK1, CDK10, CDK5, CDK5R1, CDKN2A, CEBPZOS, CELF1, CFH, CHRNA7, CLOCK, CLU, COL11A1, COMT, COX10, COX15, CR1, CTSB, CTSH, CTSS, CXCL8, CYP19A1, CYP2D6, DBH, DOC2A, DRD4, DSG2, EFNA5, EGFR, EIF2AK2, EIF4EBP1, EPHA1, ESR1, ESR2, FAM114A1, FAS, FCER1G, FERMT2, FGF1, FOXF1, FSHR, GAB2, GAPDH, GNA11, GNB3, GPX1, GRIN2B, GRN, GSK3B, GSTM1, GSTO1, GSTO2, GSTP1, GSTT1, HLA-A, HLA-DQB1, HLA-DRA, HLA-DRB5, HMOX1, HS3ST5, HTR2A, HTR6, ICA1, ICA1L, ICAM1, IGF1, IL10, IL12A, IL18, IL1A, IL1RN, IL23R, IL4, IL6, IL6R, INPP5D, INS, IQCD, IRS1, ITGB1BP1, JAZF1, KLF16, LCK, LDLR, LHCGR, LILRB2, LIME1, LLGL1, LPL, LRRC1, MAF, MAGI2, MAOA, MAPK8IP1, MAPT, MBL2, MCM3, MEF2A, MEF2C, MEFV, MIR5685, MME, MS4A1, MT-ATP6, MT-ATP8, MT-CO1, MYADM, MYLK, MYO15A, NCK2, NCSTN, NGF, NGFR, NLRP1, NLRP3, NME8, NOS1, NOS3, NTF3, NTRK1, NTRK2, OLR1, OTULIN, PCK1, PICALM, PIK3R1, PLA2G3, PLA2G4A, PLAU, PLD3, PLEKHA1, PNMT, PPARA, PPARG, PPP2R2B, PRDM7, PRKD3, PSEN1, PSEN2, PSENEN, PTK2B, RAB7A, RBCK1, RELN, REXO1, RHOH, RITA1, RPS6KB2, RXRA, SERPINA1, SERPINF2, SHARPIN, SHROOM3, SIGLEC11, SLC24A4, SNX1, SOD1, SOD2, SORL1, SORT1, SOS2, SP1, STK32B, TAP2, TF, TGFB1, TLR2, TLR4, TMEM106B, TMEM14A, TNF, TNIP1, TOP3A, TP53, TRAF2, TREM2, TSPAN14, UBE2I, UMAD1, UNC5C, VEGFA, WDR81, ZCWPW1

Among them, some of the genes are well known as the causes of AD if certain SNP occurs in them such as APP, APOE, PSEN1, PSEN2 etc., which are the main genes for Aβ and Tau. Some of the genes, such as APOA1, CHRNA7, ADRB1, HTR2A, COMT, IL1A etc., are identified as AD associated by biological function enrichment analysis of different biological process such as lipid and/or lipoprotein-related processes, metabolism, the immune system, and neural development [13]. Some of the genes, such as A2M, ABCA7, SORT1, NOS1, PTK2B etc., are collected as AD related genes through biochemical enriched pathways and pathological protein networks using crosstalk analysis [13]. These genes can react with the main Aβ and Tau genes through human protein–protein interaction pathways or networks [13]. The protein structure and function changes of these genes can affect the pathway and eventually result in AD. For example, ABCG2 is a gene in the ABC transporters pathway. It codes an Aβ transporter protein. Any changes in this protein can potentially result in AD [41]. BDNF codes for a protein called a brain derived neurotrophic factor, which maintains synaptic plasticity in the neurons and the cognitive functions in the brain. Since AD is a synaptic disease, the gene is considered as a potential factor for AD [42].

For each gene, 3 sets of data including the DNA sequence, coding region, and list of SNP information are collected from the NCBI database. The data are retrieved by searching the NCBI SNP database using the phrase "all[sb] AND gene[gene name] AND SNV." The results were filtered out by selecting only the synonymous variants. The SNP file was created and downloaded by clicking the file link on the "send to" dropdown menu. The SNP file contains a list of all SNPs and their corresponding nucleotide change information including position and the nucleotide change. In order to get the coding regions, the following processes need to be performed. First, by clicking on one of the SNPs to go to the dbSNP page, then scrolling down to the "Genomic regions, transcripts, and products" section of the page, where the SNPs of the genes are visually listed. The gene itself appeared at the top of the table as a green bar. By clicking the gene and selecting the Genbank record from the drop-down menu to go to the Genbank record of the gene. The Genbank record contains all the information about the gene. The CDS section is where the coding regions are. It is also necessary to get the start and end positions for the gene in respect to its position on the chromosome as well as the sequence direction: forward or complementary for the SNP Protein Generation Module (SNPPGM) in the pipeline to generate the SNP protein sequences. In the same page, by clicking the FASTA tab on the top-left corner, the fasta file for the gene DNA sequence

can be obtained and downloaded by clicking on the “send to” button and clicking the file button on the drop-down menu.

CHAPTER 4 METHOD

For a given gene, there could be hundreds or even thousands of SNPs associated with it, which can be obtained from the NCBI database. SNPs are mutations of a single nucleotide in the genetic sequence of a certain gene that could potentially change the shape and function of the protein it codes. These changes can accumulate over time, leading to large scale effects on the organism that it is affecting. One of the important tasks in bioinformatics is to predict whether the SNPs will result in such changes and be considered as SSNPs. In addition, statistical analysis need to be performed on the results and the corresponding info in the ADNI to explore the correlations between them. A novel and robust python-based Integrated Epigenetic Computational Pipeline (IECP) has been implemented to facilitate systematic and effective process of such large amount of SNP data. The IECP consists of three main modules: an SNP Protein Generation Module (SNPPGM), a Protein Structure and Function Change Prediction Module (PSFCPM), and a Statistical Analysis Module (SAM). SNPPGM is created to generate the SNP protein sequence automatically for many SNPs. PSFCPM is implemented for PSFC prediction to obtain SSNPs. SAM is developed for statistical analysis between the SSNPs predicted and the SNPs in the ADNI database. The SNPs in the ADNI database are classified into three groups: control, AD, and MCI. Statistical analyses including chi-square test and general hypothesis test are performed on the comparison results to obtain the correlations between these three groups. The flow chart of IECP is shown in Figure 8.

4.1 Integrated Epigenetic Computational Pipeline

4.1.1 SNP Protein Generation Module

In the SNP list for each gene collected from the NCBI database, there are a lot of SNPs. It is necessary to generate their corresponding protein sequences which are used in PSFCPM. SNPPGM is implemented for this purpose. The flow chart of this module is shown in Figure 9.

The module inputs three files: the gene DNA sequence fasta file, the coding region file, and the SNP information file. It outputs the SNP protein sequence file in the format that PSFCPM takes.

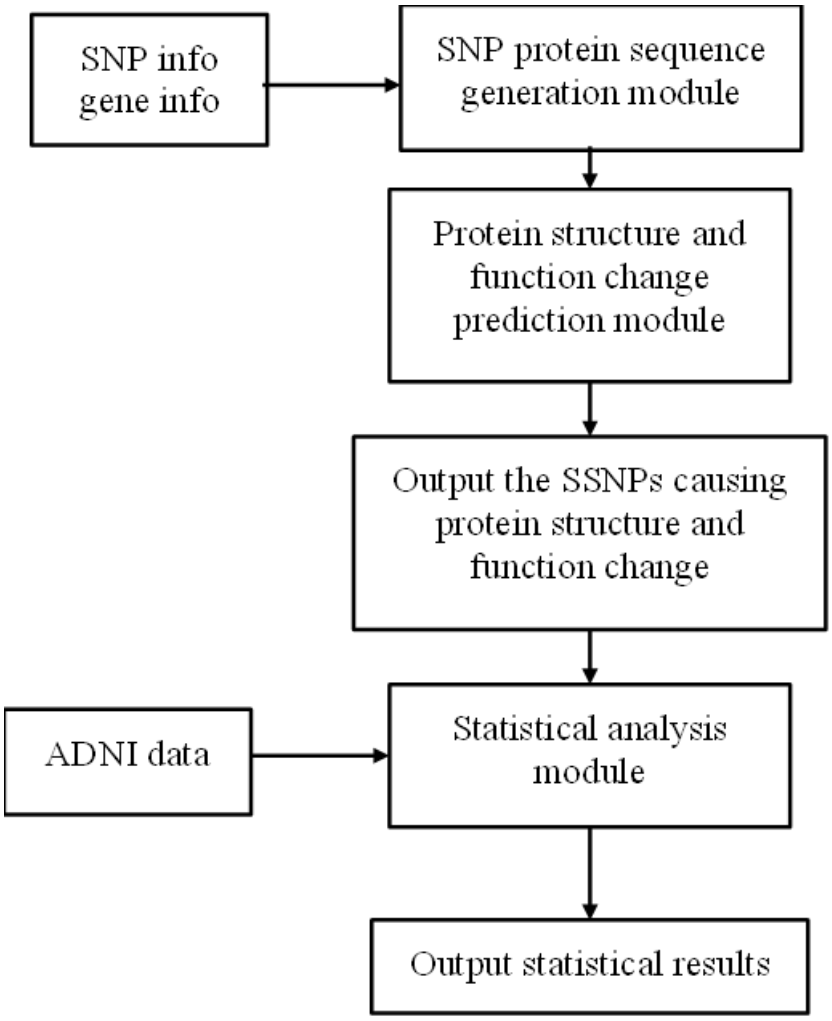

Figure 8. Flow chart of the IECP. The SNP and gene info are inputted into SNPPGM to generate the wildtype and SNP protein sequences which are processed in the protein structure and function prediction module. The output is analyzed in the statistical analysis module with the ADNI data to obtain the statistical results.

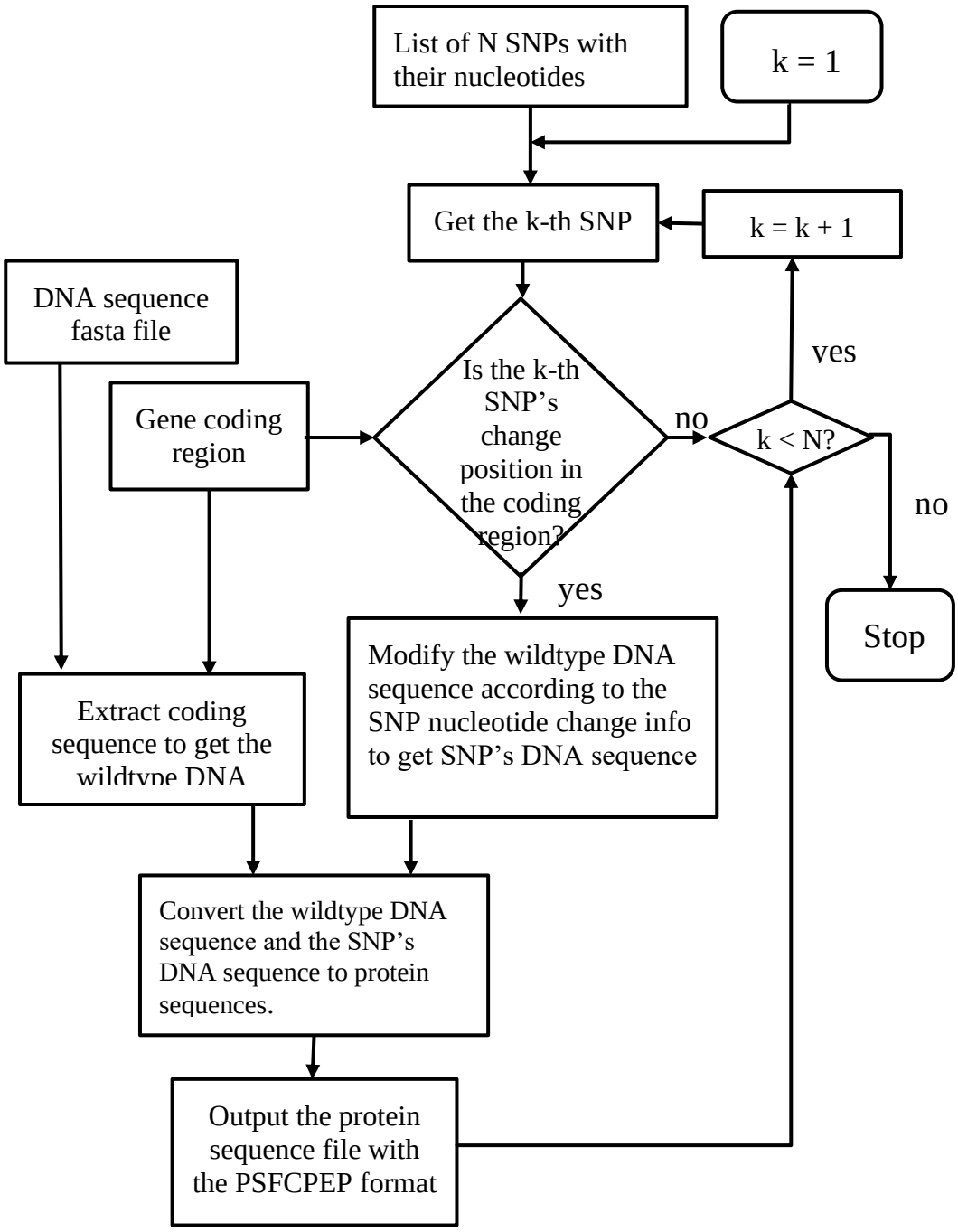

Figure 9. Flow chart of the SNPPGM module. The coding region is used for extracting both wildtype DNA and the SNP DNA. A list of SNPs with their nucleotides are processed according to their change position. If the change position is in the coding region, the SNP DNA sequence will be extracted from the wildtype sequence by modifying the corresponding nucleotide. Then, both wildtype and SNP DNA sequences are converted to protein sequences and outputted to a file.

4.1.2 Protein Structure and Function Change Prediction Module

PSFCPM takes the protein sequence files produced by SNPPGM as input. It will compare the SNP protein sequence with its corresponding wild-type protein sequence to see if there are any changes that are significant. This is done by comparing the protein sequences using the Biopython package [43, 44]. The metric parameters obtained from Biopython for SNP protein sequences are compared with the metric parameters for the wildtype protein sequence. The differences of these metric parameters are used for PSFC prediction. Several thresholds are used to control the accuracy of the prediction. The prediction results are inputted into the statistical calculation module for statistical processing.

The ProtParam module in the SeqUtils package from Biopython is used for analyzing protein sequences. The ProteinAnalysis class is used as the object for the protein analysis. The metric parameters obtained from the analyzing results are listed in Table 1.

Table 1. The metric parameters for protein change prediction

Metric Parameters	Type	Description
MOLWT	float	Molecular mass
GRAVY	float	Grand Average of Hydropathy
AACOUNT	dictionary	Counts of standard amino acids
SECSTRUCT	list	Fraction of helix, turn and sheet
SCALE	list	Profile by any amino acid scale
AROMATICITY	float	Aromaticity value
INSTABILITY	float	Instability index
FLEXIBILITY	list	Flexibility
ISOELECTRIC	float	Isoelectric point

All the metric parameters obtained using the wildtype protein sequence are compared to the corresponding ones obtained using the SNP sequence. A score is assigned to each parameter. The scores for the parameters with float type are their absolute difference. For the parameters with

dictionary type and list type, the scores are the sum of the absolute differences for the corresponding item in the dictionary or list. In addition, a total score is calculated with the sum of the scores from all the parameters. A total score for the wildtype itself is also calculated as the maximum score for total relative comparison. After many experiments and tunings, the following criteria are defined for PSFC prediction.

If a SNP's biometric scores satisfy one of the following conditions, it is considered as a SNP causing PSFC.

Criteria 1. Total score > T_t

Criteria 2. $Total\ relatve\ score\ = \frac{Total\ score}{Maximum\ score} > T_r$

Criteria 3. (Total score > T_1) AND (SECSTRUCT score > S_1)

Criteria 4. (Total score > T_2) AND (GRAVY score > S_2)

Criteria 5. [(SECSTRUCT score > S_3) OR (GRAVY score > S_4)] AND ($\frac{SECSTRUCT\ score}{Total\ score} > S_5$) AND ($\frac{GRAVY\ score}{MOLWT\ score} > S_6$)

Where T_t, T_r, T_1, T_2, S_1, S_2, S_3, S_4, S_5 and S_6 are thresholds and their values are given in Table 2.

Table 2. Values for thresholds

threshold	T_t	T_r	T_1	T_2	S_1	S_2	S_3	S_4	S_5	S_6
value	5000	0.001	75	18.06	0.00135	0.007	0.001	0.009	6e-5	2e-4

4.1.3 Validation of PSFCPM

The results predicted by PSFCPM are compared with PDB database predictions to validate the prediction accuracy of the PSFCPM. In order to see whether the SNP changes the structure and function of the protein using the PDB database, the following steps are performed.

Step 1. For each gene, the wildtype protein sequence is used to get the reference PDB identity from the PDB database.

Step 2. For each SNP, its PDB identity is obtained using its modified protein sequence from the PDB database.

Step 3. The SNP's PDB identity is compared to the reference PDB identity using the

thresholds above. If the two identities are different, the SNP will be considered as a SNP which changes the structure and function predicted by the PDB database.

When getting the identity from the PDB database using the protein sequence, a list of metrics is returned. For accurate prediction results, only the first identity, the total score, is used for comparison. This is used since it is the overall score of each of the other metrics of the protein. Heat maps are used to help find the correlations between the parameter scores and the correctness of prediction. One of them is shown in Figure 10 where SCORE is the total score and RSCORE is the total relative score. From Figure 10, both pipeline prediction and PDB prediction are the same and point to PSFC (value = 1) when the total score is a large value.

A total of 1228 SNPs are used for verifying the prediction accuracy. The results are shown in Table 3. The correct prediction percentages are about 70% compared with the predictions using the PDB database.

Table 3. PSFC prediction results

Total number of SNPs causing PSFC predicted by IEPL	590
Total number of SNPs causing PSFC predicted by PDB	589
Total number of SNPs causing PSFC predicted by both IEPL and PDB	408
Correct prediction percentage compared to PDB	69.3%

4.1.4 Statistical Analysis Module

The SAM module takes the prediction results from the PSFCPM module to perform statistical calculations. The SSNPs are compared to a list of SNPs that have been saved in the ADNI database. For each SSNP predicted by PSFCPM, statistical analysis is performed between three groups: control (CN) group, AD group, and MCI group. SAM will search the ADNI database to collect the statistical data as shown in Table 4. In the table, N_{CN} is the number of times that the SNP appears in the CN group, N_{AD} is the number of times that the SNP appears in the AD group, and N_{MCI} is the number of times that the SNP appears in the MCI group. These statistical data will be used in chi-square and hypothesis tests to obtain the statistical results.

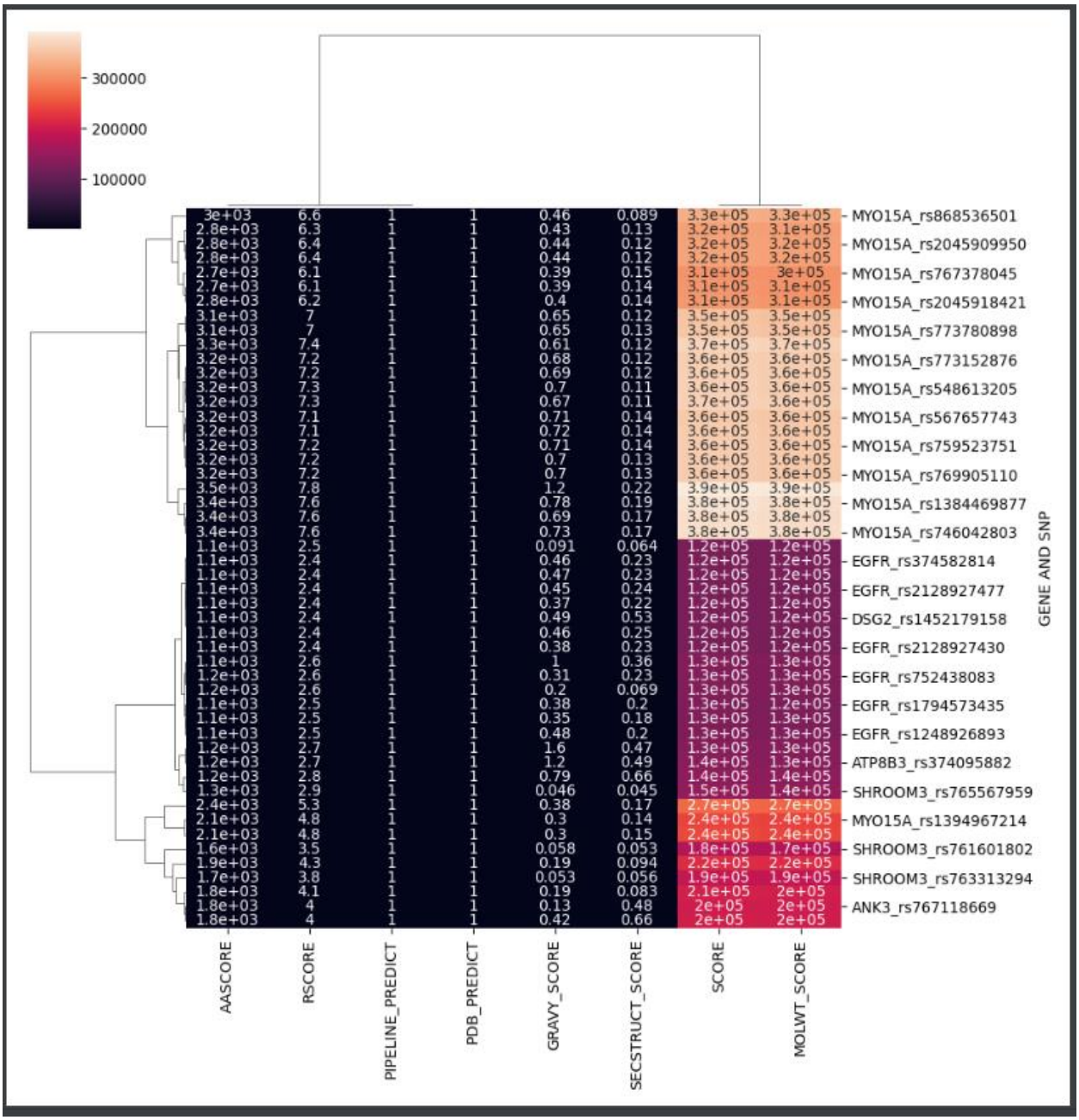

Figure 10. Heat map of PSFC prediction correlations to the parameter scores.

Table 4. Statistical data obtained by SAM

SNP	CN	AD Group	MCI Group
SNP ID	N_{CN}	N_{AD}	N_{MCI}

4.2 Application of IECP

Although IECP is used in this project for analyzing the genes and SNPs related to AD and MCI, it is a general epigenetic computational pipeline which can be applied to process genes related to any type of disease or symptoms. Given the DNA sequence, coding region, and a list of SNPs, SNPPGM will automatically generate the protein sequences for all the SNPs and output the protein sequence files in the PSFCPM format. The SNP protein sequence files can be inputted into PSFCPM for PSFC prediction to obtain SSNPs. The SSNPs are then inputted into SAM for statistical analysis together with a reference database such as ADNI for AD. The results generated by IECP can be used for epigenetic research to explore the genetic relationships of genes, SNPs, and related diseases and symptoms. It can also be used in pharmaceutical research to find medications or therapeutic treatments for the disease.

4.3 Statistical Analysis

There are many statistical tests that can be used to compare groups. The most common statistical test to compare whether the data from 2 or more groups is significantly different enough is the chi-square test. Chi-square tests are used in many different applications where the data from 2 or more groups is compared. The chi-square test creates a contingency table, counting the row and column totals for the data. The contingency table is used for the calculation of the chi-square statistic, the degrees of freedom, and the p-value.

Hypothesis tests using the two-tail F-test and the t-test are applied in this study to determine the significance of SSNPs between two groups. For each SSNP, the following procedures are performed.

Step 1. Select two groups of the total samples from CN, MCI, AD and name them as Group-1 and Group-2.

Step 2. For each group, randomly select 70% of the samples and get the count and percentage (%) of the number of the selected samples that have the SSNP and save them.

Step 3. Repeat Step 2 100 times.

Step 4. Compute the mean and standard deviation (std) values for the count and % using the results from the 100 experiments above.

Step 5. Use the two-tail F-test to determine if the variances (variance = std^2) of Group-1

and Group-2 are the same for both count and %.

Step 6. Use the t-test, equal variance or different variance based on the result of Step 5, to determine if the counts and % are significant.

The hypothesis is defined below for the F-test and t-test.

For F-test:

The null hypothesis H0: Variance of Group-1 = Variance of Group-2

The alternative hypothesis HA: Variance of Group-1 != Variance of Group-2

For t-test:

The null hypothesis H0: Mean of Group-1 = Mean of Group-2

The alternative hypothesis HA: Mean of Group-1 != Mean of Group-2

An alpha value of 0.05 ($\alpha = 0.05$) with 95% confidence level is used to compare the p-value to determine if the H0 will be rejected. When p-value is less than α, H0 will be rejected. Here the p-value is a statistical measurement used to validate a hypothesis against observed data. It measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the observed difference.

CHAPTER 5 RESULTS

In this project, a total number of 202,827 SNPs in 219 genes were processed. These SNPs are first processed by SNPPGM to generate the protein sequences for SNPs which are different from the wildtype protein sequence. The generated SNP protein sequences and their corresponding wildtype sequences were inputted into the PSFCPEP for SSNP prediction. All the SSNPs predicted by PSFCPEP were compared to the sampling data obtained from ADNI database. Finally, the comparison data was used by SAM for statistical analysis. The results are shown in the following sections.

5.1 SNPPGM Result

The results from SNPPGM are shown in Table 5. There are 112,342 SNPs whose protein sequences are different from their corresponding wildtype protein sequences.

Table 5. Results from SNPPGM

Total number of genes processed	219
Total number of SNPs	202,827
Total number of SNPs with different protein sequence from wildtype	112,342

5.2 PSFCPM Result

The 112,342 SNPs are inputted into PSFCPM for SSNP prediction. The results are shown in Table 6. There are 29,523 SSNPs predicted by PSFCPM.

Table 6. Results from PSFCPM

Total number of SNPs with different protein sequence from wildtype	112,342
Total number of SSNPs predicted by PSFCPEP	29,523

5.3 SAM Result

The 29,523 SSNPs from 219 genes were compared to the sampling data obtained from the ADNI database. There are a total of 644 samples in the data as shown in table 7.

Table 7. Samples from ADNI database

Total number of samples in the control group	150
Total number of samples in the AD group	109
Total number of samples in MCI group	385
Total number of samples	644

By comparison, among 29,523 SSNPs from 219 genes, there are 77 SNPs from 54 genes in the ADNI samples as shown in Table 8. The genes and their SSNPs are shown in Table 9.

Table 8. Number of genes and SSNPs in the ADNI samples

Total number of genes in ADNI samples	54
Total number of SNPs in ADNI samples	77

Table 9. Genes and SSNPs in the ADNI samples

Gene	SSNPs
NME8	rs10250905
TP53	rs1042522, rs17882252
ADRB2	rs1042713, rs1800888
SORT1	rs10458463
CCDC6	rs1053266
FAM114A1	rs11096964, rs11944159
MEFV	rs11466045
NGF	rs11466110, rs11466112, rs6330
GSTO1	rs11509439, rs4925
LIME1	rs1151625
OTULIN	rs11953822, rs9312870
GSTM1	rs12068997
MBL2	rs12260094, rs1800450, rs5030737, rs8191995, rs8191996
COL11A1	rs12735019
COMT	rs13306279, rs13306281, rs4680, rs5031015, rs6267
COX10	rs16948978
IL1A	rs17561
ICAM1	rs1799969, rs1801714
PPARA	rs1800234
IRS1	rs1801278
PPARG	rs1801282
NTF3	rs1805149
SERPINF2	rs2070863
PRDM7	rs2078478
PLAU	rs2227564
LDLR	rs2228671
PLA2G3	rs2232176, rs2232183

TREM2	rs2234255
CYP19A1	rs2236722, rs2304462
GAB2	rs2279374
PCK1	rs28359542
CYP2D6	rs28371717
LLGL1	rs28523978
ABCG2	rs3116448
IL1RN	rs315952
FAS	rs3218611
EPHA1	rs34372369
TNF	rs35131721, rs4645843
IGF1	rs35767
CDKN2A	rs3731249
STK32B	rs3733182
ATP8B3	rs3764606
NOS3	rs3918234
CCL2	rs4586
SOD2	rs4880, rs4987023, rs5746096, rs5746129
TLR4	rs4986791
GNB3	rs5444
FSHR	rs6165
BDNF	rs6265, rs8192466
SERPINA1	rs6647
APOE	rs7412
IL23R	rs7530511
IL6R	rs8192284
CLU	rs9331940

5.3.1 Counting Results

The counts of genes and SSNPs in each group are shown in Table 10 and 11. The counts in Table 10 are the sums of counts of SSNPs associated with the gene appearing in each group.

Table 10. Counts of genes in each group

Gene	Control	MCI	AD	Row total
NME8	116	256	24	396
TP53	274	698	188	1160
ADRB2	274	698	188	1160
SORT1	137	349	94	580
CCDC6	137	349	94	580
FAM114A1	274	698	188	1160
MEFV	137	349	94	580
NGF	411	1047	282	1740
GSTO1	274	698	188	1160
LIME1	137	349	94	580
OTULIN	274	698	188	1160
GSTM1	137	349	94	580
MBL2	664	1652	400	2716
COL11A1	137	349	94	580
COMT	685	1745	470	2900
COX10	137	349	94	580
IL1A	137	349	94	580
ICAM1	274	698	188	1160
PPARA	137	349	94	580
IRS1	137	349	94	580
PPARG	116	256	24	396
NTF3	137	349	94	580
SERPINF2	137	349	94	580
PRDM7	137	349	94	580

PLAU	137	349	94	580
LDLR	137	349	94	580
PLA2G3	274	698	188	1160
TREM2	137	349	94	580
CYP19A1	274	698	188	1160
GAB2	137	349	94	580
PCK1	137	349	94	580
CYP2D6	137	349	94	580
LLGL1	137	349	94	580
ABCG2	137	349	94	580
IL1RN	137	349	94	580
FAS	137	349	94	580
EPHA1	137	349	94	580
TNF	274	698	188	1160
IGF1	137	349	94	580
CDKN2A	137	349	94	580
STK32B	137	349	94	580
ATP8B3	137	349	94	580
NOS3	137	349	94	580
CCL2	137	349	94	580
SOD2	548	1396	376	2320
TLR4	137	349	94	580
GNB3	116	256	24	396
FSHR	137	349	94	580
BDNF	274	698	188	1160
SERPINA1	137	349	94	580
APOE	137	349	94	580
IL23R	137	349	94	580
IL6R	137	349	94	580
CLU	137	349	94	580

Column total	10465	26501	6958	43924

Table 11. Counts and percentages of SSNPs in each group

	Control	MCI	AD	Row_total	Control %	MCI %	AD %
rs10250905	116	256	24	396	77.33	66.49	22.02
rs1042522	137	349	94	580	91.33	90.65	86.24
rs1042713	137	349	94	580	91.33	90.65	86.24
rs10458463	137	349	94	580	91.33	90.65	86.24
rs1053266	137	349	94	580	91.33	90.65	86.24
rs11096964	137	349	94	580	91.33	90.65	86.24
rs11466045	137	349	94	580	91.33	90.65	86.24
rs11466110	137	349	94	580	91.33	90.65	86.24
rs11466112	137	349	94	580	91.33	90.65	86.24
rs11509439	137	349	94	580	91.33	90.65	86.24
rs1151625	137	349	94	580	91.33	90.65	86.24
rs11944159	137	349	94	580	91.33	90.65	86.24
rs11953822	137	349	94	580	91.33	90.65	86.24
rs12068997	137	349	94	580	91.33	90.65	86.24
rs12260094	137	349	94	580	91.33	90.65	86.24
rs12735019	137	349	94	580	91.33	90.65	86.24
rs13306279	137	349	94	580	91.33	90.65	86.24
rs13306281	137	349	94	580	91.33	90.65	86.24
rs16948978	137	349	94	580	91.33	90.65	86.24
rs17561	137	349	94	580	91.33	90.65	86.24
rs17882252	137	349	94	580	91.33	90.65	86.24
rs1799969	137	349	94	580	91.33	90.65	86.24
rs1800234	137	349	94	580	91.33	90.65	86.24
rs1800450	137	349	94	580	91.33	90.65	86.24
rs1800888	137	349	94	580	91.33	90.65	86.24
rs1801278	137	349	94	580	91.33	90.65	86.24

rs1801282	116	256	24	396	77.33	66.49	22.02
rs1801714	137	349	94	580	91.33	90.65	86.24
rs1805149	137	349	94	580	91.33	90.65	86.24
rs2070863	137	349	94	580	91.33	90.65	86.24
rs2078478	137	349	94	580	91.33	90.65	86.24
rs2227564	137	349	94	580	91.33	90.65	86.24
rs2228671	137	349	94	580	91.33	90.65	86.24
rs2232176	137	349	94	580	91.33	90.65	86.24
rs2232183	137	349	94	580	91.33	90.65	86.24
rs2234255	137	349	94	580	91.33	90.65	86.24
rs2236722	137	349	94	580	91.33	90.65	86.24
rs2279374	137	349	94	580	91.33	90.65	86.24
rs2304462	137	349	94	580	91.33	90.65	86.24
rs28359542	137	349	94	580	91.33	90.65	86.24
rs28371717	137	349	94	580	91.33	90.65	86.24
rs28523978	137	349	94	580	91.33	90.65	86.24
rs3116448	137	349	94	580	91.33	90.65	86.24
rs315952	137	349	94	580	91.33	90.65	86.24
rs3218611	137	349	94	580	91.33	90.65	86.24
rs34372369	137	349	94	580	91.33	90.65	86.24
rs35131721	137	349	94	580	91.33	90.65	86.24
rs35767	137	349	94	580	91.33	90.65	86.24
rs3731249	137	349	94	580	91.33	90.65	86.24
rs3733182	137	349	94	580	91.33	90.65	86.24
rs3764606	137	349	94	580	91.33	90.65	86.24
rs3918234	137	349	94	580	91.33	90.65	86.24
rs4586	137	349	94	580	91.33	90.65	86.24
rs4645843	137	349	94	580	91.33	90.65	86.24
rs4680	137	349	94	580	91.33	90.65	86.24
rs4880	137	349	94	580	91.33	90.65	86.24

rs4925	137	349	94	580	91.33	90.65	86.24
rs4986791	137	349	94	580	91.33	90.65	86.24
rs4987023	137	349	94	580	91.33	90.65	86.24
rs5030737	137	349	94	580	91.33	90.65	86.24
rs5031015	137	349	94	580	91.33	90.65	86.24
rs5444	116	256	24	396	77.33	66.49	22.02
rs5746096	137	349	94	580	91.33	90.65	86.24
rs5746129	137	349	94	580	91.33	90.65	86.24
rs6165	137	349	94	580	91.33	90.65	86.24
rs6265	137	349	94	580	91.33	90.65	86.24
rs6267	137	349	94	580	91.33	90.65	86.24
rs6330	137	349	94	580	91.33	90.65	86.24
rs6647	137	349	94	580	91.33	90.65	86.24
rs7412	137	349	94	580	91.33	90.65	86.24
rs7530511	137	349	94	580	91.33	90.65	86.24
rs8191995	116	256	24	396	77.33	66.49	22.02
rs8191996	137	349	94	580	91.33	90.65	86.24
rs8192284	137	349	94	580	91.33	90.65	86.24
rs8192466	137	349	94	580	91.33	90.65	86.24
rs9312870	137	349	94	580	91.33	90.65	86.24
rs9331940	137	349	94	580	91.33	90.65	86.24
Col_total	10465	26501	6958	43924	6976.67	6883.38	6383.49

5.3.2 Chi-Square Test Results

The chi-square tests are performed with the results in Table 10 and 11 on the number of counts. The results are shown in tables 12 and 13. From Table 12 and Table 13, the group of CN and AD and group of AD and MCI show significant differences since the p-values are less than α (0.05), but the group of CN and MCI doesn’t. The results could be due to the selections of the ADNI data.

Table 12. Chi-square test results with genes for each two groups

	CN and AD	CN and MCI	AD and MCI
statistic	97.20337	4.926837	81.64699
dof	53	53	53
p-value	0.000205	1	0.006963

Table 13. Chi-square test results with SSNPs for each two groups

	CN and AD	CN and MCI	AD and MCI
statistic	125.313	6.305509	105.129
dof	76	76	76
p-value	0.00032	1	0.015121

5.3.3 Hypothesis Test Results

The hypothesis tests are performed on the data in Table 11 for each SSNPs using the procedures described in Section 3.6. The two-tail F-test results are shown in Table 14 where T represents True and F for False. It shows that the variances between the CN and AD groups are statistically the same except for 3 SSNPs, while the variances between the CN and CMI groups and between AD and CMI groups are not.

Table 14. F-test results for SSNPs between two groups

	CN and AD			CN and MCI			AD and MCI		
SNP	p value count	p value %	var same	p value count	p value %	var same	p value count	p value %	var same
rs10250905	5.0E-04	4.9E-01	T	3.6E-10	1.2E-06	F	2.2E-16	4.8E-04	F
rs1042522	1.6E-01	1.5E-05	T	1.7E-11	2.1E-05	F	4.8E-06	2.2E-16	F
rs1042713	9.0E-01	2.3E-03	T	5.2E-06	1.2E-12	F	1.1E-04	2.2E-16	F
rs10458463	3.5E-01	1.2E-04	T	4.9E-08	2.6E-09	F	1.1E-04	2.2E-16	F
rs1053266	8.9E-02	4.1E-06	T	2.8E-12	9.3E-05	F	5.8E-06	2.2E-16	F
rs11096964	4.8E-01	2.9E-04	T	1.2E-10	3.4E-06	F	6.4E-07	2.2E-16	F
rs11466045	2.1E-01	2.9E-05	T	1.4E-13	8.1E-04	F	1.1E-07	2.2E-16	F
rs11466110	7.1E-01	1.0E-03	T	2.5E-06	4.8E-12	F	1.7E-04	2.2E-16	F
rs11466112	1.7E-01	1.9E-05	T	6.2E-10	6.2E-07	F	3.9E-05	2.2E-16	F

rs11509439	6.5E-01	7.2E-04	T	3.9E-06	2.2E-12	F	3.3E-04	2.2E-16	F
rs1151625	3.3E-01	9.8E-05	T	1.2E-06	1.8E-11	F	1.0E-03	2.2E-16	F
rs11944159	2.6E-01	5.3E-05	T	6.1E-06	8.8E-13	F	4.7E-03	2.2E-16	F
rs11953822	2.3E-01	3.7E-05	T	1.1E-12	2.0E-04	F	3.1E-07	2.2E-16	F
rs12068997	2.2E-01	3.3E-05	T	5.4E-11	7.4E-06	F	4.6E-06	2.2E-16	F
rs12260094	9.4E-01	4.5E-03	T	5.1E-09	5.3E-08	F	2.3E-07	2.2E-16	F
rs12735019	3.5E-01	1.2E-04	T	8.4E-09	2.9E-08	F	3.3E-05	2.2E-16	F
rs13306279	8.7E-01	5.9E-03	T	2.9E-03	2.2E-16	F	5.3E-03	2.2E-16	F
rs13306281	3.0E-01	7.8E-05	T	1.8E-05	1.0E-13	F	6.7E-03	2.2E-16	F
rs16948978	9.3E-01	2.7E-03	T	6.5E-09	4.0E-08	F	6.0E-07	2.2E-16	F
rs17561	1.4E-01	1.1E-05	T	7.5E-11	5.4E-06	F	1.7E-05	2.2E-16	F
rs17882252	8.7E-01	6.0E-03	T	1.3E-11	2.6E-05	F	1.4E-09	1.2E-13	F
rs1799969	6.1E-02	1.7E-06	T	9.8E-10	3.8E-07	F	3.9E-04	2.2E-16	F
rs1800234	9.0E-01	2.3E-03	T	7.0E-07	4.3E-11	F	2.5E-05	2.2E-16	F
rs1800450	5.2E-02	1.3E-06	T	3.2E-13	4.7E-04	F	4.5E-06	2.2E-16	F
rs1800888	1.4E-01	1.2E-05	T	2.0E-08	9.0E-09	F	4.9E-04	2.2E-16	F
rs1801278	7.9E-01	1.5E-03	T	6.6E-14	1.3E-03	F	3.0E-10	2.0E-12	F
rs1801282	7.9E-03	8.5E-01	T	7.4E-04	2.2E-16	F	2.1E-08	6.7E-16	F
rs1801714	1.3E-01	1.8E-01	T	8.3E-05	3.1E-15	F	7.0E-07	2.2E-16	F
rs1805149	4.5E-01	2.4E-04	T	6.3E-09	4.1E-08	F	1.3E-05	2.2E-16	F
rs2070863	2.4E-01	4.2E-05	T	1.4E-10	2.9E-06	F	6.7E-06	2.2E-16	F
rs2078478	5.6E-01	4.6E-04	T	6.3E-10	6.0E-07	F	1.2E-06	2.2E-16	F
rs2227564	3.6E-01	1.3E-04	T	3.3E-09	9.0E-08	F	1.7E-05	2.2E-16	F
rs2228671	1.8E-01	2.0E-05	T	6.2E-10	6.0E-07	F	3.7E-05	2.2E-16	F
rs2232176	7.5E-01	9.4E-03	T	1.6E-09	2.1E-07	F	2.7E-08	4.4E-16	F
rs2232183	8.2E-01	7.2E-03	T	9.4E-07	2.6E-11	F	6.4E-06	2.2E-16	F
rs2234255	3.1E-01	8.5E-05	T	2.5E-10	1.6E-06	F	4.8E-06	2.2E-16	F
rs2236722	7.9E-02	3.1E-06	T	7.1E-15	4.8E-03	F	1.8E-07	2.2E-16	F
rs2279374	6.4E-01	1.5E-02	T	4.4E-14	1.7E-03	F	3.1E-12	2.3E-09	F
rs2304462	3.4E-02	5.3E-07	F	8.9E-14	1.1E-03	F	4.7E-06	2.2E-16	F
rs28359542	4.6E-02	9.8E-07	F	3.6E-08	4.1E-09	F	4.3E-03	2.2E-16	F
rs28371717	6.6E-01	7.8E-04	T	2.1E-08	8.9E-09	F	7.3E-06	2.2E-16	F
rs28523978	7.7E-01	9.0E-03	T	3.3E-10	1.2E-06	F	8.6E-09	3.8E-15	F
rs3116448	3.9E-01	4.2E-02	T	1.2E-07	7.0E-10	F	5.1E-08	2.2E-16	F
rs315952	5.9E-01	5.5E-04	T	6.1E-13	3.0E-04	F	6.7E-09	6.2E-15	F
rs3218611	8.0E-01	1.5E-03	T	5.2E-09	5.2E-08	F	1.1E-06	2.2E-16	F

rs34372369	1.5E-01	1.5E-01	T	3.0E-07	1.9E-10	F	5.0E-09	1.1E-14	F
rs35131721	3.0E-01	6.2E-02	T	1.3E-09	2.8E-07	F	4.4E-10	1.0E-12	F
rs35767	3.5E-01	1.2E-04	T	9.9E-08	9.8E-10	F	1.7E-04	2.2E-16	F
rs3731249	1.8E-01	2.2E-05	T	3.5E-09	8.2E-08	F	1.0E-04	2.2E-16	F
rs3733182	3.4E-01	5.2E-02	T	1.3E-06	1.5E-11	F	2.5E-07	2.2E-16	F
rs3764606	3.8E-01	1.5E-04	T	4.2E-12	6.7E-05	F	1.5E-07	2.2E-16	F
rs3918234	2.9E-01	6.9E-05	T	5.1E-10	7.6E-07	F	9.3E-06	2.2E-16	F
rs4586	1.1E-01	2.0E-01	T	1.6E-07	4.5E-10	F	1.2E-09	1.7E-13	F
rs4645843	9.6E-01	3.0E-03	T	2.4E-09	1.3E-07	F	2.4E-07	2.2E-16	F
rs4680	9.6E-01	4.2E-03	T	9.1E-10	4.0E-07	F	6.7E-08	2.2E-16	F
rs4880	1.2E-01	7.6E-06	T	9.2E-11	4.4E-06	F	2.8E-05	2.2E-16	F
rs4925	9.6E-01	3.0E-03	T	3.9E-10	1.0E-06	F	6.1E-08	2.2E-16	F
rs4986791	7.4E-01	1.1E-03	T	1.6E-04	6.7E-16	F	2.9E-03	2.2E-16	F
rs4987023	4.3E-01	3.5E-02	T	1.3E-08	1.6E-08	F	1.2E-08	2.0E-15	F
rs5030737	3.9E-01	4.2E-02	T	6.8E-06	7.0E-13	F	1.7E-06	2.2E-16	F
rs5031015	1.8E-01	2.1E-05	T	1.6E-08	1.2E-08	F	2.7E-04	2.2E-16	F
rs5444	1.8E-01	1.2E-01	T	1.7E-10	2.6E-06	F	1.4E-11	2.9E-10	F
rs5746096	4.9E-01	2.6E-02	T	5.8E-05	7.1E-15	F	2.5E-05	2.2E-16	F
rs5746129	8.1E-01	1.6E-03	T	1.8E-07	3.7E-10	F	1.4E-05	2.2E-16	F
rs6165	2.4E-01	8.6E-02	T	1.1E-05	2.8E-13	F	5.7E-07	2.2E-16	F
rs6265	6.9E-01	1.2E-02	T	2.4E-08	7.2E-09	F	1.5E-07	2.2E-16	F
rs6267	8.9E-03	3.7E-08	F	1.3E-07	6.6E-10	F	4.8E-02	2.2E-16	F
rs6330	4.5E-01	2.3E-04	T	1.4E-06	1.4E-11	F	5.1E-04	2.2E-16	F
rs6647	4.4E-01	2.3E-04	T	1.3E-10	3.0E-06	F	9.4E-07	2.2E-16	F
rs7412	7.1E-01	1.0E-03	T	5.8E-10	6.6E-07	F	4.0E-07	2.2E-16	F
rs7530511	8.5E-02	3.7E-06	T	5.8E-09	4.5E-08	F	6.1E-04	2.2E-16	F
rs8191995	3.5E-01	4.9E-02	T	4.5E-14	1.8E-03	F	1.7E-13	9.3E-08	F
rs8191996	5.2E-01	3.8E-04	T	5.1E-07	7.5E-11	F	1.6E-04	2.2E-16	F
rs8192284	3.9E-01	1.6E-04	T	6.9E-11	5.8E-06	F	9.3E-07	2.2E-16	F
rs8192466	7.0E-01	9.3E-04	T	4.7E-10	8.2E-07	F	3.8E-07	2.2E-16	F
rs9312870	1.2E-01	7.8E-06	T	3.3E-10	1.2E-06	F	6.1E-05	2.2E-16	F
rs9331940	1.7E-01	1.3E-01	T	1.5E-06	1.2E-11	F	3.3E-08	2.2E-16	F

The t-test results are shown in Table 15. It shows that all SSNPs are significant for all the groups except rs1800450 and rs6647 for the groups between CN and MCI.

Table 15. T-test results for all SSNPs between two groups

	CN and AD			CN and MCI			AD and MCI		
SNP	p value count	p value %	SNP SIG	p value count	p value %	SNP SIG	p value count	p value %	SNP SIG
rs10250905	6.0E-214	3.0E-198	T	0.0E+00	3.0E-85	T	0.0E+00	9.1E-126	T
rs1042522	1.2E-174	2.2E-42	T	0.0E+00	1.1E-04	T	0.0E+00	3.9E-28	T
rs1042713	9.4E-174	1.6E-42	T	0.0E+00	4.7E-04	T	0.0E+00	8.5E-30	T
rs10458463	8.2E-173	4.4E-42	T	0.0E+00	2.7E-03	T	0.0E+00	3.1E-29	T
rs1053266	2.0E-177	3.6E-44	T	0.0E+00	1.5E-05	T	0.0E+00	1.6E-28	T
rs11096964	4.9E-173	6.6E-47	T	0.0E+00	3.4E-10	T	0.0E+00	7.4E-29	T
rs11466045	8.5E-174	2.8E-39	T	0.0E+00	3.2E-03	T	0.0E+00	2.7E-27	T
rs11466110	3.9E-174	4.6E-44	T	0.0E+00	3.2E-06	T	0.0E+00	3.3E-29	T
rs11466112	9.0E-176	1.1E-47	T	0.0E+00	2.1E-08	T	0.0E+00	1.3E-29	T
rs11509439	3.0E-175	5.5E-48	T	0.0E+00	3.9E-05	T	7.5E-307	3.5E-32	T
rs1151625	1.4E-172	2.4E-43	T	0.0E+00	3.2E-05	T	6.5E-294	1.0E-28	T
rs11944159	6.7E-167	4.1E-38	T	0.0E+00	9.9E-08	T	1.8E-276	7.3E-24	T
rs11953822	3.6E-184	7.1E-52	T	0.0E+00	1.5E-08	T	0.0E+00	2.0E-32	T
rs12068997	7.5E-179	4.7E-50	T	0.0E+00	6.7E-08	T	0.0E+00	1.8E-31	T
rs12260094	4.3E-179	1.0E-44	T	0.0E+00	2.3E-04	T	0.0E+00	2.2E-31	T
rs12735019	2.1E-175	2.8E-47	T	0.0E+00	4.4E-05	T	0.0E+00	2.3E-31	T
rs13306279	6.0E-171	9.9E-41	T	0.0E+00	6.4E-08	T	1.7E-284	1.6E-26	T
rs13306281	5.8E-176	5.7E-46	T	0.0E+00	1.7E-03	T	3.1E-282	2.9E-31	T
rs16948978	8.2E-177	2.6E-47	T	0.0E+00	2.6E-07	T	0.0E+00	4.0E-31	T
rs17561	1.6E-175	5.0E-43	T	0.0E+00	1.7E-06	T	0.0E+00	1.9E-27	T
rs17882252	3.0E-180	1.3E-54	T	0.0E+00	1.6E-13	T	0.0E+00	1.7E-33	T
rs1799969	1.6E-173	1.0E-39	T	0.0E+00	1.5E-04	T	3.4E-299	4.1E-26	T
rs1800234	3.5E-171	8.8E-44	T	0.0E+00	1.9E-04	T	0.0E+00	2.2E-30	T
rs1800450	3.8E-167	7.8E-35	T	0.0E+00	6.6E-02	F	0.0E+00	5.3E-25	T
rs1800888	4.7E-174	6.7E-42	T	0.0E+00	2.0E-03	T	1.5E-299	1.7E-28	T
rs1801278	1.9E-184	5.4E-49	T	0.0E+00	6.5E-05	T	0.0E+00	1.3E-33	T
rs1801282	1.4E-204	4.3E-188	T	0.0E+00	3.5E-76	T	0.0E+00	2.5E-106	T
rs1801714	1.2E-170	2.6E-43	T	0.0E+00	5.3E-06	T	0.0E+00	3.7E-31	T
rs1805149	2.5E-173	7.9E-42	T	0.0E+00	3.6E-06	T	0.0E+00	1.9E-27	T
rs2070863	1.1E-176	2.8E-44	T	0.0E+00	3.9E-07	T	0.0E+00	3.4E-28	T
rs2078478	3.6E-177	7.9E-47	T	0.0E+00	1.4E-06	T	0.0E+00	1.1E-30	T
rs2227564	2.2E-175	1.3E-42	T	0.0E+00	2.1E-06	T	0.0E+00	1.1E-27	T

rs2228671	5.7E-176	9.4E-42	T	0.0E+00	1.7E-04	T	0.0E+00	7.4E-28	T
rs2232176	2.1E-179	1.4E-47	T	0.0E+00	2.0E-05	T	0.0E+00	6.3E-33	T
rs2232183	2.8E-174	3.6E-45	T	0.0E+00	1.1E-06	T	0.0E+00	2.4E-30	T
rs2234255	2.0E-180	1.5E-50	T	0.0E+00	4.9E-06	T	0.0E+00	7.4E-33	T
rs2236722	1.2E-179	2.3E-51	T	0.0E+00	8.1E-05	T	0.0E+00	2.1E-33	T
rs2279374	2.6E-180	2.0E-47	T	0.0E+00	8.3E-04	T	0.0E+00	1.7E-34	T
rs2304462	1.0E-142	3.8E-31	T	0.0E+00	9.0E-07	T	0.0E+00	3.7E-25	T
rs28359542	2.8E-143	4.1E-33	T	0.0E+00	3.9E-07	T	3.5E-280	1.1E-26	T
rs28371717	8.9E-180	1.2E-44	T	0.0E+00	6.9E-04	T	0.0E+00	5.3E-31	T
rs28523978	2.2E-179	3.4E-52	T	0.0E+00	1.9E-08	T	0.0E+00	2.5E-34	T
rs3116448	1.5E-169	2.3E-39	T	0.0E+00	1.4E-03	T	0.0E+00	2.0E-29	T
rs315952	3.1E-180	1.4E-50	T	0.0E+00	4.0E-08	T	0.0E+00	1.6E-32	T
rs3218611	3.4E-176	4.5E-41	T	0.0E+00	1.1E-02	T	0.0E+00	6.4E-30	T
rs34372369	1.7E-176	1.3E-47	T	0.0E+00	5.7E-07	T	0.0E+00	1.5E-33	T
rs35131721	3.8E-175	6.0E-48	T	0.0E+00	1.5E-06	T	0.0E+00	1.5E-33	T
rs35767	1.7E-175	3.2E-43	T	0.0E+00	3.4E-09	T	0.0E+00	1.2E-26	T
rs3731249	3.2E-175	5.9E-41	T	0.0E+00	4.1E-06	T	0.0E+00	1.9E-26	T
rs3733182	7.3E-180	2.3E-50	T	0.0E+00	3.9E-04	T	0.0E+00	2.3E-36	T
rs3764606	7.8E-177	9.5E-45	T	0.0E+00	2.4E-04	T	0.0E+00	2.1E-30	T
rs3918234	4.3E-180	9.2E-49	T	0.0E+00	1.2E-05	T	0.0E+00	6.9E-32	T
rs4586	2.9E-176	2.0E-45	T	0.0E+00	2.7E-02	T	0.0E+00	5.1E-36	T
rs4645843	5.8E-179	1.4E-48	T	0.0E+00	9.4E-03	T	0.0E+00	6.9E-35	T
rs4680	1.7E-175	1.9E-46	T	0.0E+00	8.7E-06	T	0.0E+00	1.3E-31	T
rs4880	3.2E-177	7.8E-48	T	0.0E+00	2.0E-07	T	0.0E+00	5.5E-30	T
rs4925	7.9E-184	1.2E-47	T	0.0E+00	3.3E-06	T	0.0E+00	4.8E-32	T
rs4986791	1.1E-171	2.9E-40	T	0.0E+00	3.9E-07	T	5.4E-288	3.3E-26	T
rs4987023	1.1E-176	4.8E-47	T	0.0E+00	8.6E-04	T	0.0E+00	2.3E-34	T
rs5030737	2.9E-172	2.5E-46	T	0.0E+00	1.1E-05	T	0.0E+00	1.6E-32	T
rs5031015	5.3E-174	4.3E-44	T	0.0E+00	4.5E-08	T	1.8E-303	1.9E-27	T
rs5444	1.9E-210	3.6E-192	T	0.0E+00	6.7E-87	T	0.0E+00	7.9E-109	T
rs5746096	2.2E-168	1.6E-41	T	0.0E+00	3.3E-06	T	0.0E+00	1.2E-28	T
rs5746129	7.7E-172	2.4E-40	T	0.0E+00	3.0E-04	T	0.0E+00	3.5E-28	T
rs6165	1.7E-169	1.7E-40	T	0.0E+00	3.0E-04	T	0.0E+00	5.0E-30	T
rs6265	9.8E-174	3.0E-44	T	0.0E+00	8.7E-04	T	0.0E+00	6.5E-32	T
rs6267	1.2E-134	3.4E-30	T	0.0E+00	4.4E-05	T	9.2E-256	2.3E-25	T
rs6330	1.9E-173	1.3E-40	T	0.0E+00	2.1E-04	T	1.2E-301	1.1E-27	T

rs6647	1.9E-175	1.2E-39	T	0.0E+00	5.5E-02	F	0.0E+00	4.1E-29	T
rs7412	9.7E-172	2.3E-43	T	0.0E+00	1.3E-03	T	0.0E+00	1.6E-30	T
rs7530511	5.3E-172	1.5E-40	T	0.0E+00	1.5E-04	T	9.4E-295	9.8E-27	T
rs8191995	1.6E-212	3.9E-194	T	0.0E+00	2.2E-91	T	0.0E+00	1.7E-111	T
rs8191996	6.6E-177	1.6E-47	T	0.0E+00	1.7E-05	T	0.0E+00	1.5E-31	T
rs8192284	4.2E-177	2.5E-47	T	0.0E+00	2.5E-04	T	0.0E+00	5.2E-32	T
rs8192466	3.5E-175	2.4E-41	T	0.0E+00	4.8E-03	T	0.0E+00	1.2E-29	T
rs9312870	6.5E-175	2.0E-41	T	0.0E+00	2.3E-03	T	0.0E+00	3.3E-28	T
rs9331940	1.7E-177	2.4E-50	T	0.0E+00	1.8E-05	T	0.0E+00	3.5E-36	T

CHAPTER 6 DISCUSSION

6.1 SSNPs for Biomarkers of AD

The hypothesis test results from the t-test indicate that all the SSNPs predicted by IECP are significant between CN and AD groups. These SSNPs are linked to AD through different pathways. They act as biogenetic and epigenetic activators leading to AD or targets for therapies to treat AD. The genes associated with these SSNPs play certain important roles in AD diagnosis, treatments, and prevention. These SSNPs can be considered as biomarkers in the research of AD and related dementia.

6.2 Relations between Genes and AD

As shown in Table 8 and Table 9, 54 genes with 77 SSNPs predicted by PSPCPEP are in the ADNI database. To validate the result, these genes are checked with research publications to find their link to AD.

ABCG2: it is a gene in the ABC transporters pathway of AD. The gene codes an Aβ transporter protein and has been found to be significantly up regulated in Alzheimer's disease [41].

ADRB: this gene is found in the calcium signaling pathway. It codes for the beta-2 adrenergic receptors, which regulate cognitive function. Mutations in this gene can affect the signaling between neurons in the brain [45].

APOE: it is one of the main genes that is responsible for AD, since its codes for the apolipoprotein responsible for synthesizing the Aβ protein. More specifically, the ε4 allele of the gene increases the risk of AD [46].

ATP8B3: this gene is present in the vesicular pathway, which is used in many cellular functions. The gene was identified using a network modeling algorithm. The algorithm was developed to find overlapping candidate genes [47].

BDNF: it codes for a protein called a brain derived neurotrophic factor. These proteins maintain synaptic plasticity in the neurons in the brain, making it important for cognitive functions. Since AD is a synaptic disease, the gene is considered as a potential biomarker for AD [42].

CCDC6: it is a gene that was found to have a relationship to AD using the genome wide association study [48].

CCL2: it is a gene that codes for the Chemokine C-C motif ligand, which is upregulated in the case of AD. The gene was studied in mice, using the mouse model of tauopathy [49]. The protein produced by the gene produces microglia induced Aβ oligomerization.

CDKN2A: it codes for the Cyclin-dependent kinase inhibitor protein, which is important in cellular aging [50]. The protein was found to be a biomarker for AD, since aging neurons are a big cause for AD.

CLU: this is the gene that codes for clustering, a protein that is expressed under cellular stress. Its role in lipid transport and immune modulation makes it an important protein for cells. The protein is an important biomarker for AD [51].

COL11A1: this gene has been identified as a candidate AD gene in genetic association studies. It codes for a protein important to cell adhesion [52].

COMT: this gene codes for the catechol-O-methyltransferase protein and has a synergistic effect with APOE that makes it a biomarker for AD [53].

COX10: this codes the cytochrome C oxidase protein, which serves the mitochondrial electron transport chain. The gene was found to be significantly downregulated in AD patients [54].

CYP19A1: this gene is associated with the biosynthesis of estrogen, which is significantly associated with AD [55].

EPHA1: this gene was found in the genome wide association study of AD. It is specifically found in late onset AD patients [56].

FAM114A1: this gene was found to have a relationship by using gene set enrichment analysis [57].

FAS: this gene is associated with apoptosis or cell death. Apoptosis affects all cells, including neurons. Because of this, the gene is associated with AD [58].

FSHR: this gene codes for the follicle-stimulating hormone receptor. It determines human fertility. It was found in a gender specific AD study that links the disease with fertility [59].

GBA2: this gene codes for the β-glucocerebrosidase enzyme, which was found to have an association with Parkinson's disease [60].

GNB3: this gene codes for the G protein β3 subunit, part of the G-protein receptor. This protein was found to have a relationship with ADRB3. Both genes contribute to AD [61].

GSTM1: this gene codes for the protein glutathione, which protects cells from damage caused by oxidation. Underrepresentation of this gene is linked to AD [62].

GSTO1: this gene codes the glutathione S-transferase protein, which is also part of the glutathione pathway [63].

ICAM1: this gene codes for the Inter-Cellular Adhesion Molecule, which is involved in cell-to-cell interactions. The gene has a peripheral role with AD [64].

IGF1: this gene is important in the synthesis of insulin. Lower levels of IGF1 were found to be a contribution to AD [65].

IL1A: this gene codes for the interleukin protein, which is important in the inflammatory pathway. It was found that this gene could increase the risk of AD [66].

IL1RN: this gene is associated with signal transduction through IL-1R. Significant transcriptional up regulation of this gene was found in AD patients which could play an important group-specific role in AD pathophysiology [67].

IL23R: this gene encodes Inhibition of interleukin-23 receptor and was found being associated AD in a Northern Han Chinese population [68].

IL6R: this genes codes the interleukin 6 (IL6) receptor protein whose polymorphisms could modify IL6 signaling and affect AD pathogenesis directly or indirectly [69].

IRS1: this gene is associated with neurotrophic signaling pathway and plays key roles in regulating growth and survival, metabolism, and aging. It was found that its phosphorylation is increased in the brains of AD patients [70].

LDLR: this gene is linked to ovarian steroidogenesis. Its variants were found to be significantly associated with AD [71].

LIME1: this gene encodes a transmembrane adaptor protein that links the T and B-cell receptor stimulation. It was found that this gene could modulate the metabolism of APP and link to AD [72].

LLGL1: this gene encodes a protein that is similar to a tumor suppressor in Drosophila which is part of a cytoskeletal network. It is one of the genes associated with most significantly enriched gene ontology terms potentially linked to AD [73].

MBL2: this gene encodes the soluble mannose-binding lectin or mannose-binding protein is an important element in the innate immune system. It was found that its two haplotypes, LXP and LYQ, were significantly associated with AD risk [74].

MEFV: this gene encodes a protein called pyrin (also known as marenostrin) that is an important modulator of innate immunity. It is one of the key genes associated with familial Mediterranean fever. Its variants are found to be associated with sporadic early-onset Alzheimer's disease in an Italian population [75].

NGF: this gene is a member of the NGF-beta family and encodes a secreted protein which homodimerizes and is incorporated into a larger complex. It is a protein that exerts pharmacological effects on a group of cholinergic neurons known to atrophy in AD [76].

NME8: this gene is known to be responsible for primary ciliary dyskinesia type 6. It was found to play a role in lowering the brain neurodegeneration related to AD [77].

NOS3: this gene acts as a biologic mediator in several processes, including neurotransmission and antimicrobial and antitumoral activities. It was found that the Glu298Asp polymorphism of this gene could be a genetic risk factor for late-onset AD, especially in Chinese population [78].

NTF3: the protein encoded by this gene controls survival and differentiation of mammalian neurons and is closely related to both nerve growth factor and brain-derived neurotrophic factor. Its polymorphism was found to be a relevant risk factor for AD [79].

OTULIN: this gene is associated with the peptidase C65 family of ubiquitin isopeptidases, which remove ubiquitin from proteins. It can regulate the linear ubiquitin chain assembly complex which was linked to AD and related dementias [80].

PCK1: this gene is a main control point for the regulation of gluconeogenesis. Its polymorphisms were found to be strongly associated with AD [81].

PLA2G3: this gene encodes a protein in the secreted phospholipase A2 family which functions in lipid metabolism. It is the most overexpressed gene in a human neuronal model of oxidative stress and found to be associated with AD [82].

PLAU: this gene encodes a serine protease that converts plasminogen to plasmin. Its variant was found to have a genetic and functional involvement in the pathogenesis of AD [83].

PPARA: this gene encodes the subtype Peroxisome Proliferator-Activated Receptor (PPAR)-alpha, which is a nuclear transcription factor. It is an important factor regulating autophagy in the clearance of Aβ and a potential therapeutic target for AD [84].

PPARG: this gene encodes a member of the PPAR subfamily of nuclear receptors regulating amyloidogenic pathways. Its Pro12Ala polymorphism may modify the age at onset of AD [85].

PRDM7: the protein encoded by this gene has a role in transcription and other nuclear processes. Its variants could be linked to several neurodegenerative diseases, including AD [86].

SERPINA1: this gene is associated with complement and coagulation cascades pathway. Its isoforms was found to be linked to AD and could be an interesting diagnostic supplement to the related dementia [87].

SERPINF2: this gene encodes a member of the serpin family of serine protease inhibitors. It has connections to the elements of the amyloid machinery and plays a role in the onset of AD [88].

SOD2: the protein encoded by this gene belongs to the iron/manganese superoxide dismutase family. Its polymorphism was found as a risk factor for AD in Polish population [89].

SORT1: this gene encodes a member of the VPS10-related sortilin family of proteins. Its genetic variant was identified as being associated with reduced risk of AD [90].

STK32B: this gene encodes a serine-threonine protein kinase which transfers phosphate molecules to the oxygen atoms of serine and threonine. It was found that this gene could have been associated with AD [91].

TRL4: the protein encoded by this gene belongs to the toll-like receptor family which plays a fundamental role in pathogen recognition and activation of innate immunity. It was identified as a promising therapeutic target in AD treatment [92].

TNF: this gene encodes a multifunctional proinflammatory cytokine that belongs to the tumor necrosis factor superfamily. Its involvement in the pathogenesis of AD has been classified [93].

TP53: this gene encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. Its mutations in exon 7 may be associated with pathogenesis of AD [94].

TREM2: the protein encoded by this gene forms a receptor signaling complex with the tyrosine kinase binding protein. This gene was suggested as a potential therapeutic target for AD [95].

6.3 Limitations

The ADNI database selected for this project consists of a total of 644 samples with 150 in CN group, 109 in AD group, and 385 in CMI group. These are not much from the statistical point of view. For more accurate results more samples should be collected for the study. In the hypothesis tests, they are performed with respect to each SSNPs which also limit the scope of the tests. The hypothesis should also be done with a combination of SSNPs associated with a gene or a number of genes in the same biological pathway. Such tests will provide more inside to the effect of SSNP combinations on AD and MCI. In addition, it can improve the accuracy of the results with more factors to be considered.

CHAPTER 7 CONCLUSIONS

A systematic approach for identification of biomarkers in AD and related dementia is developed in this project. A novel IECP software tool including SNPPGM, PSFCPM, and SAM is implemented for large-scale bioinformatic data processing. A total of 219 genes and 202,827 SNPs were processed and 54 genes and 77 SNPs were found to be significant between the control group and AD group with the samples obtained from the ADNI database. These SSNPs can be used as biomarkers for future study of AD and related dementia to explore more genetical and biological factors. It can also help to accelerate the development of new diagnosis methods, medications, treatment therapies, and prevention tools. IECP is a general epigenetic pipeline and can be used for studying the genes and SNPs related to any other type of diseases such as pancreatic cancer.

CHAPTER 8 FUTURE WORK AND RESEARCH

Immediate work that can be done is to implement a SNP collecting module and integrate it into the IECP. Given a list of genes, the IECP can automatically collect all the SNPs associated with these genes from a given database such as NCBI. All the collected SNPs will be automatically processed by IECP. The results of statistical analysis will be outputted by IECP at the final stage. There is no manual process required in the entire flow. The IECP will be an effective and efficient tool for researchers to study SNP's epigenetic relations to any disease.

The metric parameters used in significant SNP prediction should be further studied and tuned with more experiments to increase the accuracy of the prediction. The prediction results should be verified with more databases, such as UniProt, in addition to PDB to improve the prediction credibility. The number of samples for each group including control group and disease group in the study should be increased as many as possible to assure that the statistical results are correct and reliable.

Statistical analysis can be enhanced by extending the single SNP analysis to a group of SNPs or even multiple groups of SNPs together. Statistical analysis performed on multiple SNPs can explore inter correlations between different SNPs or genes on a disease and its related dementia.

REFERENCES

[1] Alzheimer's Association, "2023 Alzheimer's disease facts and figures," *Alzheimers Dement*, vol 19, no 4, pp. 1598-1695, 2023.

[2] K. A. Matthewsa, W. Xu, A. H. Gagliotic, J. B. Holta, J. B. Crofta, D. Mackc, and L. C. McGuire, "Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015-2060) in adults aged ≥65 years," *Alzheimers Dement,* vol 15, no 1, pp. 17-24, 2019.

[3] G. Glenner and C. W. Wong, "Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein," *Biochemical and Biophysical Research Communications,* vol 120, pp. 885-890, 1984.

[4] G. Glenner and C. W. Wong, "Alzheimer's disease and Down's syndrome: Sharing of a unique cerebrovascular amyloid fibril protein," *Biochemical and Biophysical Research Communications,* vol 122, pp. 1131-1135, 1984.

[5] H. Braak and E. Braak, “Evolution of neuronal changes in the course of Alzheimer’s disease,” *Journal of Neural Transmission Supplement*, vol 53, pp. 127-140, 1998.

[6] C. Ballatore, V. M.-Y. Lee, and J. Q. Trojanowski, “Tau-mediated neurodegeneration in Alzheimer’s disease and related disorders”, *Nature Reviews Neuroscience*, vol 8, pp. 663–672, 2007.

[7] D. L. Price, R. E. Tanzi, D. R. Borchelt, and S. S. Sisodia, “Alzheimer’s disease: genetic studies and transgenic models,” *Annu Rev Genet*, vol 32, pp. 461–493, 1998.

[8] J. Hardy and D. J. Selkoe, “The amyloid hypothesis of Alzheimer’s disease: progress and problems on the road to therapeutics,” *Science*, vol 297, pp. 353–356, 2002.

[9] S. W. Pimplikar, “Reassessing the amyloid cascade hypothesis of Alzheimer’s disease,” *Int J Biochem Cell Biol*, vol 41, pp. 1261–1268, 2009.

[10] G. P. Morris, I. A. Clark, and B. Vissel, “Questions concerning the role of amyloid-β in the definition, aetiology and diagnosis of Alzheimer’s disease,” *Acta Neuropathol*, vol 136, pp. 663–689, 2018.

[11] H. R. Kim, T. Lee, J. K. Choi, and Y. Jeong, “Genetic variants beyond amyloid and tau associated with cognitive decline,” *Neurology*, vol 95, pp. e2366-e2377, 2020.

[12] M. Giri, M. Zhang, and Y. Lü, “Genes associated with Alzheimer's disease: an overview and current status,” *Clinical Interventions in Aging*, vol 11, pp. 665–681, 2016.

[13] Y. S. Hu, J. Xin, Y. Hu, L. Zhang, and J. Wang, “Analyzing the genes related to Alzheimer’s disease via a network and pathway-based approach,” *Alzheimer's Research & Therapy*, vol 9, no. 29, 2017.

[14] Alzheimer's Disease International: Dementia statistics. Retrieved 30 April 2023.

[15] Office of The Assistant Secretary for Planning and Evaluation, “National Plan to Address Alzheimer’s Disease: 2022 Update,” *The U.S. Department of Health and Human Services*, 2022.

[16] The World Health Organization, “Global action plan on the public health response to dementia 2017 – 2025,” *The World Health Organization*, 2017.

[17] The U.S. Food and Drug Administration, “FDA Grants Accelerated Approval for Alzheimer’s Disease Treatment,” *FDA News Release*, June 07, 2021.

[18] The U.S. Food and Drug Administration, “FDA Grants Accelerated Approval for Alzheimer’s Disease Treatment,” *FDA News Release*, January 06, 2023.

[19] J. Cummings, “The Role of Biomarkers in Alzheimer's Disease Drug Development.” *Adv Exp Med Biol.* Vol 1118, pp. 29-61, 2019.

[20] C. Humpel, “Identifying and validating biomarkers for Alzheimer's disease.” *Trends in Biotechnology*, vol. 29, No. 1, 2011.

[21] D. A. Bennett, J. A. Schneider, Z. Arvanitakis, J. F. Kelly, N. T. Aggarwal, R. C. Shah, and R. S. Wilson, “Neuropathology of older persons without cognitive impairment from two community-based studies.” *Neurology*, vol 6, no 12, pp. 837-1844, 2006.

[22] M. A. Mintun, G. N. Larossa, Y. I. Sheline, C. S. Dence, S. Y. Lee, R. H. Mach, W. E. Klunk, C. A. Mathis, S. T. DeKosky, and J. C. Morris “[11C]PIB in a nondemented population: potential antecedent marker of Alzheimer disease.” *Neurology*, vol 67, no 3, pp. 446–452, 2006

[23] G. S. Bloom, “Amyloid-β and Tau: The Trigger and Bullet in Alzheimer Disease Pathogenesis.” *JAMA Neurol*., vol 71, no 4, pp. 505–508, 2014.

[24] Alzheimer’s Association, “Special report, More Than Normal Aging: Understanding Mild Cognitive Impairment.” *Alzheimer’s Association Alzheimer’s Disease Facts and Figures*, 2022.

[25] L. Aswathy, R. S. Jisha, V. H. Masand, J. M. Gajbhiye, I. G. Shibi, “Design of novel amyloid β aggregation inhibitors using QSAR, pharmacophore modeling, molecular docking and ADME prediction,” *In Silico Pharmacol*. vol 6, pp. 1-9, 2018.

[26] F. LaFerla, K. Green, and S. Oddo, “Intracellular amyloid-β in Alzheimer's disease.” *Nat Rev Neurosci*, vol 8, pp. 499–509, 2007.

[27] M.M. Rahman, C. Lendel, “Extracellular protein components of amyloid plaques and their roles in Alzheimer’s disease pathology.” *Mol Neurodegeneration*, vol 16, no 59, 2021.

[28] J. Andrade-Guerrero, A. Santiago-Balmaseda, P. Jeronimo-Aguilar, I. Vargas-Rodríguez, A.R. Cadena-Suárez, C. Sánchez-Garibay, G. Pozo-Molina, C.F. Méndez-Catalá, M.D. Cardenas-Aguayo, S. Diaz-Cintra, M. Pacheco-Herrero, “Alzheimer’s Disease: An Updated Overview of Its Genetics.” *International Journal of Molecular Sciences*, vol. 24, no. 4, 2023.

[29] X. Xiao, B. Jiao, X. Liao, W. Zhang, Z. Yuan, L. Guo, X. Wang, L. Zhou, X. Liu, X. Yan, B. Tang, “Association of genes involved in the metabolic pathways of amyloid-β and tau proteins with sporadic late-onset Alzheimer’s disease in the Southern Han Chinese population.” Frontiers in Aging Neuroscience, Vol 12, p. 584801, 2020.

[30] S. Dhamija and M. B. Menon, “Non-coding transcript variants of protein-coding genes - what are they good for?” *RNA biology*, vol. 15, no 8, pp. 1025–1031, 2018.

[31] J.L. Carlin, “Mutations Are the Raw Materials of Evolution.” *Nature Education Knowledge*, vol 3, no 2, pp. 10, 2011.

[32] C.B. Anfinsen, “Principles that govern the folding of protein chains.” *Science*, vol 181, pp. 223–230, 1973.

[33] M. Lorch, J.M. Mason, A.R. Clarke, and M.J. Parker, “Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the I-state.” *Biochemistry*, vol 38, pp. 1377–1385, 1999.

[34] M. Lorch, J.M. Mason, R.B. Sessions, and A.R. Clarke, “Effects of mutations on the thermodynamics of a protein folding reaction: implications for the mechanism of formation of the intermediate and transition states.” *Biochemistry*, vol 39, pp. 3480–3485, 2000.

[35] T.R. Rignall, J.O. Baker, S.L. McCarter, W.S. Adney, T.B. Vinzant, S.R. Decker, et al., “Effect of single active-site cleft mutation on product specificity in a thermostable bacterial cellulase.” *Appl. Biochem. Biotechnol*. vols 98-100, pp. 383–394, 2002.

[36] M.U. Ung, B. Lu, and J.A. McCammon, “E230Q mutation of the catalytic subunit of cAMP-dependent protein kinase affects local structure and the binding of peptide inhibitor. *Biopolymers*, vol 81, pp. 428–439, 2006.

[37] S. Tiede, M. Cantz, J. Spranger, and T. Braulke, “Missense mutation in the N-acetylglucosamine-1-phosphotransferase gene (GNPTA) in a patient with mucolipidosis II induces changes in the size and cellular distribution of GNPTG.” *Hum. Mutat.* vol 27, pp. 830–831, 2006.

[38] V.M. Prabantu, N. Naveenkumar, and N. Sriniva, “Influence of Disease-Causing Mutations on Protein Structural Networks,” *Front. Mol. Biosci., Sec. Biological Modeling and Simulation*, vol 7, 2021.

[39] D. Lee, O. Redfern, and C. Orengo, “Predicting protein function from sequence and structure,” *Nature Reviews Molecular Cell Biology*, vol 8, pp. 995-1005, 2007.

[40] C. Bellenguez, F. Küçükali, I. E. Jansen, et al. “New insights into the genetic etiology of Alzheimer’s disease and related dementias,” *Nature Genetics*, vol 54, pp. 412–436, 2022.

[41] Á. Fehér, A. Juhász, A. László, M. Pákáski, J. Kálmán, Z. Janka, “Association between the ABCG2 C421A polymorphism and Alzheimer's disease.” *Neurosci Lett.*, vol 550, pp. 51-54, 2013.

[42] L. Gao, Y. Zhang, K. Sterling, et al., “Brain-derived neurotrophic factor in Alzheimer’s disease and its pharmaceutical potential.” *Transl Neurodegener*, vol 11, no 4, pp. 1-34, 2022.

[43] P.A. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M.J.L. de Hoon, “Biopython: freely available

Python tools for computational molecular biology and bioinformatics." *Bioinformatics*, vol 25, pp. 1422-1423, 2009.

[44] B.A. Chapman and J.T. Chang, "Biopython: Python tools for computational biology." *ACM SIGBIO Newsletter*, vol 20, pp. 15-19, 2000.

[45] A.K. Evans, P.M. Ardestani, B. Yi, H.H. Park, R.K. Lam, and M. Shamloo, "Beta-adrenergic receptor antagonism is proinflammatory and exacerbates neuroinflammation in a mouse model of Alzheimer's Disease." *Neurobiology of Disease*, vol 146, pp. 1-16, 2020.

[46] A.C. Raulin, S.V. Doss, Z.A. Trottier, et al., "ApoE in Alzheimer's disease: pathophysiology and therapeutic strategies." *Mol Neurodegeneration*, vol 17, no 1, pp. 1-26, 2022.

[47] P. Talwar, Y. Silla, S. Grover, et al., "Genomic convergence and network analysis approach to identify candidate genes in Alzheimer's disease." *BMC Genomics* vol 15, pp. 1-6, 2014.

[48] J. Schwartzentruber, S. Cooper, J.Z. Liu, et al. "Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes." *Nat Genet*, vol 53, pp. 392–402, 2021.

[49] A. Joly-Amado, J. Hunter, Z. Quadri, F. Zamudio, P.V. Rocha-Rangel, D. Chan, A. Kesarwani, K. Nash, D.C. Lee, D. Morgan, M.N. Gordon, and M.B. Selenica, "CCL2 Overexpression in the Brain Promotes Glial Activation and Accelerates Tau Pathology in a Mouse Model of Tauopathy." *Frontiers in immunology*, vol 11, no. 997, pp. 1-17, 2020.

[50] H. Mori, Y. Funahashi, Y. Yoshino, H. Kumon, Y. Ozaki, K. Yamazaki, S. Ochi, A. Tachibana, T. Yoshida, H. Shimizu, T. Mori, J.I. Iga, and S.I. Ueno, "Blood CDKN2A Gene Expression in Aging and Neurodegenerative Diseases." *Journal of Alzheimer's Disease*, vol 82, no 4, pp. 1737–1744, 2021.

[51] E.M. Foster, A. Dangla-Valls, S. Lovestone, E.M. Ribe, and N.J. Buckley, "Clusterin in Alzheimer's Disease: Mechanisms, Genetics, and Lessons From Other Pathologies." *Frontiers in Neuroscience*, vol. 13, no 164, pp. 1-27, 2019.

[52] H. Gao, Y. Tao, Q. He, F. Song, and D. Saffen, "Functional enrichment analysis of three Alzheimer's disease genome-wide association studies identities DAB1 as a novel candidate liability/protective gene." *Biochemical and biophysical research communications*, vol 463, no 4, pp. 490–495, 2015.

[53] M.F. Martínez, X.E. Martín, L.G. Alcelay, et al., "The COMT Val158 Met polymorphism as an associated risk factor for Alzheimer disease and mild cognitive impairment in APOE 4 carriers." *BMC Neurosci*, vol 10, no 1, pp. 1-9, 2009.

[54] R. Bi, W. Zhang, D.F. Zhang, M. Xu, Y. Fan, Q.X. Hu, H.Y. Jiang, L. Tan, T. Li, Y. Fang, C. Zhang, and Y.G. Yao, "Genetic association of the cytochrome c oxidase-related genes

with Alzheimer's disease in Han Chinese." *Neuropsychopharmacology*, vol 43, no 11, pp. 2264–2276, 2018.

[55] Y. Song, Y. Lu, Z. Liang, et al., "Association between rs10046, rs1143704, rs767199, rs727479, rs1065778, rs1062033, rs1008805, and rs700519 polymorphisms in aromatase (CYP19A1) gene and Alzheimer's disease risk: a systematic review and meta-analysis involving 11,051 subjects." *Neurol Sci*, vol 40, pp. 2515–2527, 2019.

[56] M. Talebi, A. Delpak, M. Khalaj-kondori, et al., "ABCA7 and EphA1 Genes Polymorphisms in Late-Onset Alzheimer's Disease." *J Mol Neurosci*, vol 70, pp. 167–173, 2020.

[57] Q.S. Li and L. De Muynck, "Differentially expressed genes in Alzheimer's disease highlighting the roles of microglia genes including OLR1 and astrocyte gene CDK2AP1." *Brain, behavior, & immunity – health*, vol 13, article 100227, pp. 1-16, 2021.

[58] S.M. De la Monte, Y.K. Sohn, and J.R. Wands, "Correlates of p53- and Fas (CD95)-mediated apoptosis in Alzheimer's disease." *Journal of the neurological sciences*, vol 152, no 1, pp. 73–83, 1997.

[59] R.M. Corbo, G. Gambina, E. Broggio, and R. Scacchi, "Influence of variation in the follicle-stimulating hormone receptor gene (FSHR) and age at menopause on the development of Alzheimer's disease in women." *Dementia and geriatric cognitive disorders*, vol 32, no 1, pp. 63–69, 2011.

[60] R. Franco, G. Navarro, and E. Martínez-Pinilla, "Lessons on Differential Neuronal-Death-Vulnerability from Familial Cases of Parkinson's and Alzheimer's Diseases." *International Journal of Molecular Sciences*, vol 20, no 13, p. 3297, 2019.

[61] M.J. Bullido, M.C. Ramos, A. Ruiz-Gómez, A.S. Tutor, I. Sastre, A. Frank, F. Coria, P. Gil, F. Mayor Jr, and F. Valdivieso, "Polymorphism in genes involved in adrenergic signaling associated with Alzheimer's." *Neurobiology of aging*, vol 25, no 7, pp. 853–859, 2004.

[62] M. Wang, Y. Li, L. Lin, G. Song, and T. Deng, "GSTM1 Null Genotype and GSTP1 Ile105Val Polymorphism Are Associated with Alzheimer's Disease: a Meta-Analysis." *Molecular neurobiology*, vol 53, pp. 1355–1364, 2016.

[63] Y.J. Li, S.A. Oliveira, P. Xu, et al, "Glutathione S-transferase omega-1 modifies age-at-onset of Alzheimer disease and Parkinson disease." *Human molecular genetics*, vol 12, no 24, pp. 3259–3267, 2003.

[64] M. Reale, M.A. Kamal, L. Velluto, D. Gambi, M. Di Nicola, and N.H. Greig, "Relationship between inflammatory mediators, Aβ levels and ApoE genotype in Alzheimer disease." *Current Alzheimer research*, vol 9, no 4, pp. 447–457, 2012.

[65] A.J. Westwood, A. Beiser, C. Decarli, T.B. Harris, T.C. Chen, X.M. He, R. Roubenoff, A.

Pikula, R. Au, L.E. Braverman, P.A. Wolf, R.S. Vasan, and S. Seshadri, "Insulin-like growth factor-1 and risk of Alzheimer dementia and brain atrophy." *Neurology*, vol 82, no 18, pp. 1613–1619, 2014.

[66] P.L. McGeer and E.G. McGeer, "Polymorphisms in Inflammatory Genes and the Risk of Alzheimer Disease." *Arch Neurol.* vol 58, no 11, pp. 1790–1792, 2001.

[67] M.T. Mizwicki, G. Liu, M. Fiala, L. Magpantay, J. Sayre, A. Siani, M. Mahanian, R. Weitzman, E.Y. Hayden, M.J. Rosenthal, I. Nemere, J. Ringman, and D.B. Teplow, "1α,25-dihydroxyvitamin D3 and resolvin D1 retune the balance between amyloid-β phagocytosis and inflammation in Alzheimer's disease patients." *Journal of Alzheimer's disease*, vol 34, no 1, pp. 155–170, 2013.

[68] Y. Liu, J.T. Yu, W. Zhang, Y. Zong, R.C. Lu, J. Zhou, and L. Tan, "Interleukin-23 receptor polymorphisms are associated with Alzheimer's disease in Han Chinese." *Journal of neuroimmunology*, vol 271, no 1-2, pp. 43–48, 2014.

[69] J.S.K. Kauwe, M.H. Bailey, P.G. Ridge, R. Perry, M.E. Wadsworth, K.L. Hoyt, et al, "Genome-Wide Association Study of CSF Levels of 59 Alzheimer's Disease Candidate Proteins: Significant Associations with Proteins Involved in Amyloid Processing and Inflammation." *PLoS Genet*, vol 10, no 10, p. e1004758, 2014.

[70] D. Tanokashira, W. Fukuokaya, and A. Taguchi, "Involvement of insulin receptor substrates in cognitive impairment and Alzheimer's disease." *Neural regeneration research*, vol 14, no 8, pp. 1330–1334, 2019.

[71] R.K. Gopalraj, H. Zhu, J.F. Kelly, M. Mendiondo, J.F. Pulliam, D.A. Bennett, and S. Estus, "Genetic association of low density lipoprotein receptor and Alzheimer's disease." *Neurobiology of aging*, vol 26, no 1, pp. 1–7, 2005.

[72] C. Bellenguez, F. Küçükali, I.E. Jansen, et al, "New insights into the genetic etiology of Alzheimer's disease and related dementias." *Nat Genet*, vol 54, pp. 412–436, 2022.

[73] A.I. Bernstein, Y. Lin, R.C. Street, L. Lin, Q. Dai, L. Yu, H. Bao, M. Gearing, J.J. Lah, P.T. Nelson, C. He, A.I. Levey, J.G. Mullé, R. Duan, P. Jin, "5-Hydroxymethylation-associated epigenetic modifiers of Alzheimer's disease modulate Tau-induced neurotoxicity," *Human Molecular Genetics*, vol 25, no 12, pp. 2437–2450, 2016.

[74] Sjölander, A., Minthon, L., Nuytinck, L., Vanmechelen, E., Blennow, K., & Nilsson, S. (2013). Functional mannose-binding lectin haplotype variants are associated with Alzheimer's disease. *Journal of Alzheimer's disease*, vol 35, no 1, pp. 121–127, 2013.

[75] M. Arra, E. Emanuele, V. Martinelli, P. Minoretti, M. Bertona, and D. Geroldi, "The M694V variant of the familial Mediterranean fever gene is associated with sporadic early-onset Alzheimer's disease in an Italian population sample." *Dementia and geriatric cognitive*

disorders, vol 23, no 1, pp. 55–59, 2007.

[76] S.A. Scott and K.A. Crutcher, "Nerve growth factor and Alzheimer's disease." *Reviews in the neurosciences*, vol 5, no 3, pp. 179–211, 1994.

[77] Y. Liu, J.T. Yu, H.F. Wang, X.K. Hao, Y.F. Yang, T. Jiang, et al, "Association between *NME8* Locus Polymorphism and Cognitive Decline, Cerebrospinal Fluid and Neuroimaging Biomarkers in Alzheimer's Disease." *PLoS ONE*, vol 9, no 12, p. e114777, 2014.

[78] B. Wang, S. Tan, Z. Yang, Y.C. Xie, J. Wang, S. Zhou, S. Li, C. Zheng, and X. Ma, "Association between Alzheimer's disease and the NOS3 gene Glu298Asp polymorphism in Chinese." *Journal of molecular neuroscience*, vol 34, no 2, pp. 173–176, 2008.

[79] M. Liu, Y. R. Huo, J. Wang, S.L. Liu, S. Liu, C. Wang, J.H. Wang, and Y. Ji, "Polymorphisms of the neurotrophic factor-3 (NTF-3) in Alzheimer's disease: rs6332 associated with onset time and rs6489630 T allele exhibited a protective role," *Journal of Neurogenetics*, vol 29, no 4, pp. 183-187, 2015.

[80] C. Bellenguez, F. Küçükali, I.E. Jansen, et al, "New insights into the genetic etiology of Alzheimer's disease and related dementias." *Nat Genet*, vol 54, pp. 412–436, 2022.

[81] G. Hamilton, P. Proitsi, L. Jehu, A. Morgan, J. Williams, M.C. O'Donovan, M.J. Owen, J.F. Powell, and S. Lovestone, "Candidate Gene Association Study of Insulin Signaling Genes and Alzheimer's Disease: Evidence for *SOS2*, *PCK1*, and *PPARγ* as Susceptibility Loci." *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol 144, no 4, pp. 508-516, 2007.

[82] A. Martínez-García, I. Sastre, M. Recuero, J. Aldudo, E. Vilella, I. Mateo, P. Sánchez-Juan, T. Vargas, E. Carro, F. Bermejo-Pareja, E. Rodríguez-Rodríguez, O. Combarros, M. Rosich-Estrago, A. Frank, F. Valdivieso, and M.J. Bullido, "PLA2G3, a gene involved in oxidative stress induced death, is associated with Alzheimer's disease." *Journal of Alzheimer's disease*, vol 22, no 4, 1181–1187, 2010.

[83] M. Riemenschneider, L. Konta, P. Friedrich, et al, "A functional polymorphism within plasminogen activator urokinase (PLAU) is associated with Alzheimer's disease." *Human molecular genetics*, vol 15, no 16, pp. 2446–2456, 2006.

[84] R. Luo, L.Y. Su, G. Li, J. Yang, Q. Liu, L.X. Yang, D.F. Zhang, H. Zhou, M. Xu, Y. Fan, J. Li, and Y.G. Yao, "Activation of PPARA-mediated autophagy reduces Alzheimer disease-like pathology and cognitive decline in a murine model." *Autophagy*, vol 16, no 1, pp. 52–69, 2020.

[85] S. Wang, L. Guan, D. Luo, et al, "Gene- gene interaction between PPARG and APOE gene on late-onset Alzheimer's disease: A case- control study in Chinese han population." *J Nutr*

Health Aging, vol 21, pp. 397–403, 2017.

[86] J. Lake, C. Warly Solsberg, J.J. Kim, J. Acosta-Uribe, M.B. Makarious, Z. Li, K. Levine, P. Heutink, C.X. Alvarado, D. Vitale, and S. Kang, "Multi-ancestry meta-analysis and fine-mapping in Alzheimer's Disease." *Molecular Psychiatry*, pp.1-12, 2023.

[87] L. Barba, S. Halbgebauer, F. Paolini Paoletti, G. Bellomo, S. Abu-Rumeileh, P. Steinacker, F. Massa, L. Parnetti, and M. Otto, "Specific Cerebrospinal Fluid SerpinA1 Isoform Pattern in Alzheimer's Disease." *International journal of molecular sciences*, vol 23, no 13, p. 6922, 2022.

[88] P.L. De Jager, G. Srivastava, K. Lunnon, et al, "Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci." *Nature neuroscience*, vol 17, no 9, pp. 1156–1163, 2014.

[89] K. Spisak, A. Klimkowicz-Mrowiec, J. Pera, T. Dziedzic, G. Aleksandra, and A. Slowik, "rs2070424 of the SOD1 gene is associated with risk of Alzheimer's disease," *Neurologia i Neurochirurgia Polska*, vol 48, no 5, pp. 342-345, 2014.

[90] C.H. Andersson, O. Hansson, L. Minthon, N. Andreasen, K. Blennow, H. Zetterberg, I. Skoog, A. Wallin, S. Nilsson, and P. Kettunen, "A Genetic Variant of the Sortilin 1 Gene is Associated with Reduced Risk of Alzheimer's Disease." *Journal of Alzheimer's disease*, vol 35, no 4, pp. 1353–1363, 2016.

[91] A. Nazarian, A.I. Yashin, and A.M. Kulminski, "Genome-wide analysis of genetic predisposition to Alzheimer's disease and related sex disparities." *Alzheimer's research and therapy*, vol 11, no 1, p. 5, 2019.

[92] Y. Zhou, Y. Chen, C. Xu, H. Zhang, and C. Lin, "TLR4 Targeting as a Promising Therapeutic Strategy for Alzheimer Disease Treatment." *Frontiers in neuroscience*, vol 14, p. 602508, 2020.

[93] R.T. Perry, J.S. Collins, H. Wiener, R. Acton, and R.C. Go, "The role of TNF and its receptors in Alzheimer's disease." *Neurobiology of aging*, vol 22, no 6, pp. 873–883, 2001.

[94] J. Dorszewska, A. Różycka, A. Oczkowska, J. Florczak-Wyspiańska, M. Prendecki, M. Dezor, I. Postrach, P. Jagodzinski, and W. Kozubski, "Mutations of TP53 Gene and Oxidative Stress in Alzheimer's Disease Patients." *Advances in Alzheimer's Disease*, vol 3, pp. 24-32, 2014.

[95] S.E. Hickman and J. El Khoury, "TREM2 and the neuroimmunology of Alzheimer's disease." *Biochemical pharmacology*, vol 88, no 4, pp. 495–498, 2014.