



University of Kentucky
UKnowledge

Theses and Dissertations--Epidemiology and
Biostatistics

College of Public Health


2017

STATISTICAL ANALYSES TO DETECT AND REFINE GENETIC ASSOCIATIONS WITH NEURODEGENERATIVE DISEASES

Yuriko Katsumata

University of Kentucky, katsumata.yuriko@uky.edu

Author ORCID Identifier:

 <https://orcid.org/0000-0002-0188-8094>

Digital Object Identifier: <https://doi.org/10.13023/ETD.2017.486>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

Recommended Citation

Katsumata, Yuriko, "STATISTICAL ANALYSES TO DETECT AND REFINE GENETIC ASSOCIATIONS WITH NEURODEGENERATIVE DISEASES" (2017). *Theses and Dissertations--Epidemiology and Biostatistics*. 17. https://uknowledge.uky.edu/epb_etds/17

This Doctoral Dissertation is brought to you for free and open access by the College of Public Health at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Epidemiology and Biostatistics by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Yuriko Katsumata, Student

Dr. David W. Fardo, Major Professor

Dr. Steve R. Browning, Director of Graduate Studies

STATISTICAL ANALYSES TO DETECT AND REFINE GENETIC ASSOCIATIONS
WITH NEURODEGENERATIVE DISEASES

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Public Health
at the University of Kentucky

By
Yuriko Katsumata

Lexington, Kentucky

Director: Dr. David W. Fardo, Associate Professor of Biostatistics

Lexington, Kentucky

2017

Copyright © Yuriko Katsumata 2017

ABSTRACT OF DISSERTATION

STATISTICAL ANALYSES TO DETECT AND REFINE GENETIC ASSOCIATIONS WITH NEURODEGENERATIVE DISEASES

Dementia is a clinical state caused by neurodegeneration and characterized by a loss of function in cognitive domains and behavior. Alzheimer's disease (AD) is the most common form of dementia. Although the amyloid β ($A\beta$) protein and hyperphosphorylated tau aggregates in the brain are considered to be the key pathological hallmarks of AD, the exact cause of AD is yet to be identified. In addition, clinical diagnoses of AD can be error prone. Many previous studies have compared the clinical diagnosis of AD against the gold standard of autopsy confirmation and shown substantial AD misdiagnosis. Hippocampal sclerosis of aging (HS-Aging) is one type of dementia that is often clinically misdiagnosed as AD. AD and HS-Aging are controlled by different genetic architectures. Familial AD, which often occurs early in life, is linked to mainly mutations in three genes: *APP*, *PSEN1*, and *PSEN2*. Late-onset AD (LOAD) is strongly associated with the $\epsilon 4$ allele of apolipoprotein E (*APOE*) gene. In addition to the *APOE* gene, genome-wide association studies (GWAS) have identified several single nucleotide polymorphisms (SNPs) in or close to some genes associated with LOAD. On the other hand, *GRN*, *TMEM106B*, *ABCC9*, and *KCNMB2* have been reported to harbor risk alleles associated with HS-Aging pathology. Although GWAS have succeeded in revealing numerous susceptibility variants for dementias, it is an ongoing challenge to identify functional loci and to understand how they contribute to dementia pathogenesis.

Until recently, rare variants were not investigated comprehensively. GWAS rely on genotype imputation which is not reliable for rare variants. Therefore, imputed rare variants are typically removed from GWAS analysis. Recent advances in sequencing technologies enable accurate genotyping of rare variants, thus potentially improving our understanding the role of rare variants on disease. There are significant computational and statistical challenges for these sequencing studies. Traditional single variant-based association tests are underpowered to detect rare variant associations. Instead, more powerful and computationally efficient approaches for aggregating the effects of rare variants have become a standard approach for association testing. The sequence-kernel association test (SKAT) is one of the most powerful rare variant analysis methods. A recently-proposed scan-statistic-based test is another approach to detect the location of rare variant clusters influencing disease.

In the first study, we examined the gene-based associations of the four putative risk genes, *GRN*, *TMEM106B*, *ABCC9*, and *KCNMB2* with HS-aging pathology. We analyzed haplotype associations of a targeted *ABCC9* region with HS-Aging pathology and with *ABCC9* gene expression. In the second study, we elucidated the role of the non-coding SNPs identified in the International Genomics of Alzheimer's Project (IGAP) consortium GWAS within a systems genetics framework to understand the flow of biological information underlying AD. In the last study, we identified genetic regions which contain rare variants associated with AD using a scan-statistic-based approach.

KEYWORDS: Alzheimer's Disease, Hippocampal Sclerosis of Aging, Region-based Associations, Rare Variant Associations, Systems Genetics

Yuriko Katsumata

December 7, 2017

Date

STATISTICAL ANALYSES TO DETECT AND REFINE GENETIC ASSOCIATIONS
WITH NEURODEGENERATIVE DISEASES

By

Yuriko Katsumata

David W. Fardo, PhD
Director of Dissertation

Steve R. Browning, PhD
Director of Graduate Studies

December 7, 2017
Date

ACKNOWLEDGEMENTS

I would first like to express my deepest gratitude to my chair, Dr. David Fardo, for the continuous support of my Ph.D. study and related research. This dissertation would not have been finished without his guidance, patience, motivation, and immense knowledge. I would also like to thank Dr. Peter Nelson for his insightful directions and comments which incited me to widen my research.

I would like to thank Dr. Richard Kryscio and Dr. Erin Abner for being in my committee and providing valuable questions for my research that will keep me on the right track. My gratitude is also extended to the outside examiner Dr. Gia Mudd-Martin.

My sincere thanks goes to Dr. Frederick Schmitt, Dr. Allison Caban-Holt, and Ms. Roberta Davis, who always provided encouragement. I had wonderful time with them in the past four years during my Ph.D. study.

Finally, I would like to thank my sister's family, close relatives, and friends for their constant support.

TABLE OF CONTENTS

Acknowledgements	iii
List of Tables	vi
List of Figures	viii
Chapter One: Introduction	1
HS-Aging	2
Brief history of genetic risks of AD and HS-Aging	3
Rare variants and next-generation sequencing (NGS)	7
Dissertation Outline	10
Chapter Two: Gene-based association study of genes linked to hippocampal sclerosis of aging neuropathology: <i>GRN</i> , <i>TMEM106B</i> , <i>ABCC9</i> , and <i>KCNMB2</i>	
Abstract	16
Introduction	16
Material and methods	19
Study subjects	19
Quality control of the ADGC genotype data	20
Identifying ethnic outliers	20
Statistical analysis	21
Gene-based association analysis	21
Haplotype-based association analysis for HS-Aging	22
Haplotype-based expression Quantitative Trait Locus (eQTL) analysis for <i>ABCC9</i> gene expression	22
Results	24
Single variant-based association	24
Gene-based association	25
Haplotype-based association with HS-Aging	25
Haplotype-based expression Quantitative Trait Locus (eQTL) association with <i>ABCC9</i> gene expression	26
Discussion	26
Funding	31
Chapter Three: Translating Alzheimer’s disease-associated polymorphisms into functional candidates: a survey of IGAP genes	
Abstract	42
Introduction	43
Material and methods	46
Genetic datasets	46
Gene expression datasets	47
Statistical analysis	49
Hypothesis 1: identified IGAP SNPs are proxies of coding SNPs	49
Hypothesis 2: identified IGAP SNPs are regulatory SNPs	49

Results.....	50
Hypothesis 1: identified IGAP SNPs are proxies of coding SNPs.....	50
Hypothesis 2: identified IGAP SNPs are regulatory SNPs	51
Discussion.....	53
Hypothesis 1: identified IGAP SNPs are proxies of coding SNPs.....	53
Hypothesis 2: identified IGAP SNPs are regulatory SNPs	57
Funding.....	60
Chapter Four: Identifying regions harboring Alzheimer’s disease related rare variants using scan-statistic-based analysis of exome sequencing data	
Abstract.....	79
Introduction.....	80
Material and methods.....	83
Study subjects.....	83
ADSP WES data.....	84
Statistical analysis	84
Results.....	87
Scan-statistic-based analysis by a sliding window approach in <i>TREM2</i> and <i>TOMM40</i>	87
Scan-statistic-based analysis by an optimized window approach in <i>TREM2</i> and <i>TOMM40</i>	88
Genome-widely scan-statistic-based analysis by an optimized window approach	88
Discussion.....	89
Funding	93
Supplementary method 1	107
Supplementary method 2	109
Supplementary method 3	111
Chapter Five: Conclusion	
Summary.....	112
Strengths and Limitations	117
Future Research	118
References.....	120
Vita.....	133

LIST OF TABLES

Table 1.1. <i>APOE</i> isoforms encoded by two single nucleotide polymorphisms.....	12
Table 1.2. Single nucleotide polymorphisms for Alzheimer’s Disease identified in previous study.....	13
Table 2.1. Quality control filters for single nucleotide polymorphism.....	32
Table 2.2. Comparison of selected characteristics between hippocampal sclerosis of aging cases and controls who died at age 60 years or older	33
Table 2.3. Most associated variant with hippocampal sclerosis of aging in four genes using a logistic regression model assuming a recessive/additive/dominant mode of inheritance in people who died at age 60 years or older.....	34
Table 2.4. Gene-based associations of the target four genes with hippocampal sclerosis of aging assuming a recessive/additive/dominant mode of inheritance in people who died at age 60 years or older.....	35
Table 2.5. Gene-based associations of the target four genes with hippocampal sclerosis of aging assuming a recessive/additive/dominant mode of inheritance in people who died at age 80 years or older.....	36
Table 2.6. Haplotype association with <i>ABCC9</i> gene expression in human brain assuming an additive mode of inheritance.....	37
Table 3.1. ADGC and CHARGE studies in ADSP	61
Table 3.2. Characteristics of the individual in ADSP.....	62
Table 3.3. Exonic single nucleotide polymorphism correlated with the IGAP SNP	63
Table 3.4. Association results for exonic SNPs correlated with the IGAP SNPs.....	65
Table 3.5. Pathogenic nature of nonsynonymous single nucleotide polymorphism associated with AD	66
Table 3.6. Significant <i>cis</i> - and <i>trans</i> -association of the IGAP SNPs with blood gene expressions in ADNI.....	67
Table 3.7. Significant <i>cis</i> - and <i>trans</i> -association of the IGAP SNPs with brain gene expressions in NABEC and UKBEC	68
Table 3.8. Associations between gene expressions identified in NABEC and UKBEC and AD status in GSE5281 dataset, entorhinal cortex (EC), hippocampus (HIPP), and medial temporal gyrus (MTG).....	69
Table 3.9. Associations between gene expressions identified in NABEC and UKBEC and AD status in GSE5281 dataset, posterior cingulate (PC), superior frontal gyrus (SFG), and primary visual cortex (VCX).....	71
Table 3.10. Associations between gene expressions identified in NABEC and UKBEC and AD status in Allen Institute dataset.....	73
Table 4.1. Summary of the major findings in rare variants associated with late-onset Alzheimer’s disease.....	94

Table 4.2. Number of cases/controls and age range in each case-control study of ADSP	95
Table 4.3. Significantly associated rare coding variants with Alzheimer’s disease assuming a dominant mode of inheritance.....	96
Table 4.4. Characteristics of ADSP study subjects.....	97
Table 4.5. Scan-statistic (logLRw) for windows determined by a sliding window approach in several different settings, <i>TREM2</i> and <i>TOMM40</i>	98
Table 4.6. Scan-statistic (logLRw) for windows determined by an optimized window approach in each large genetic region, <i>TREM2</i> and <i>TOMM40</i>	100
Table 4.7. Optimized windows for top 10 genes with a large scan-statistic in risk and protective effects	101
Table 4.8. Single-variant-based association within the optimized windows	102

LIST OF FIGURES

Figure 2.1. Flow diagram of the subjects included in the analyses	38
Figure 2.2. The first and second principal components plots along with 1000 genome reference samples.....	39
Figure 2.3. Proportion and 95% confidence interval of hippocampal sclerosis of aging cases	40
Figure 2.4. Estimation of haplotype frequencies and association using four tag single nucleotide polymorphisms on the <i>ABCC9</i> gene when assuming a recessive mode of inheritance	41
Figure 3.1. Possible causal relationships between single nucleotide polymorphisms (SNPs), mRNA, and phenotype	74
Figure 3.2. Flow diagram of the subjects included in the analyses	75
Figure 3.3. Correlation between gene expressions potentially regulated by IGAP SNPs in NABEC.....	76
Figure 3.4. Correlation between gene expressions potentially regulated by <i>CRI</i> SNPs in UKBEC.....	77
Figure 3.5. Plot for the associations of the <i>CRI</i> SNPs with the <i>CRI</i> expression in the average of 10 brain regions of UKBEC	78
Figure 4.1. Flow diagram of the subjects included in the analyses	104
Figure 4.2. First and second principal components plots along with 1000 genome reference samples	105
Figure 4.3. Manhattan plot of scan-statistic (logLRw) for the optimized windows in each gene.....	106

CHAPTER ONE

Introduction

Dementia is a clinical state caused by neurodegeneration and characterized by a loss of function in cognitive domains and undesirable changes in behavior. Alzheimer's disease (AD) is the most common form of dementia, accounting for over 50% of dementia cases [1]. In the US, it has been estimated that 5.2 million people have AD and total payments from health-care and long-care services for AD patients are \$214 billion in 2014 [2]. AD imposes a severe burden on patients themselves as well as caregivers and public health systems. Although it has been more than 100 years since Alois Alzheimer published "About a Peculiar Disease of the Cerebral Cortex" in 1907 [3], the exact cause of AD is yet to be identified. Amyloid β ($A\beta$) protein and hyperphosphorylated tau aggregates in the brain are considered to be the key pathological hallmarks in AD patients [4, 5]. A predominant mechanistic hypothesis for AD pathogenesis is the "amyloid cascade hypothesis" that suggests that AD is caused by lack of $A\beta$ clearance, which triggers downstream neuronal injury such as synaptic and neuronal loss, enhanced neuroinflammation, tau hyperphosphorylation, and eventually the clinical symptoms of AD [6].

$A\beta$ is a peptide of amino acids which is derived from amyloid precursor protein (APP) cleaved by β - and γ -secretases [7, 8]. The γ -secretase cleavage occurs at position 40 or 42 of APP, yielding two major species of $A\beta$: $A\beta_{40}$ ($A\beta$ ending at residue 40) and $A\beta_{42}$ ($A\beta$ ending at residue 42) peptides [7]. Although $A\beta_{40}$ is the most abundant form of $A\beta$, $A\beta_{42}$

is less soluble and more neurotoxic than A β ₄₀ as it produces higher levels of A β oligomers [9, 10].

In brains with AD, tau, a major neuronal microtubule-assembly-activator protein, is abnormally hyperphosphorylated in neurofibrillary tangles (NFTs). The function of tau is regulated by its degree of phosphorylation. Putatively, 85 phosphorylation sites have been identified at serine, threonine, and tyrosine residues [11, 12] with approximately 45 specific sites identified for AD pathogenesis [13]. The abnormally hyperphosphorylated tau reduces the binding affinity to microtubules, binds to normal tau to form insoluble oligomers, and eventually develop NFTs which cause neurodegenerative diseases (called tauopathy) [14].

HS-Aging

The clinical diagnosis of AD is a challenging process that requires to remove other potential types of dementia. Many previous studies have compared the clinical diagnosis of AD against the gold standard of autopsy confirmation and shown substantial AD misdiagnosis [15-18]. The accurate diagnosis of AD is crucial to provide optimal treatments for patients as well as to recruit participants in clinical trials for new therapies. Hippocampal sclerosis of aging (HS-Aging) is one type of dementia that is often clinically misdiagnosed as AD [19-21]. Clinical signs and symptoms of HS-Aging are similar to those of AD with amnesic memory deficits [20, 21]. AD is characterized by the accumulation of amyloid plaques and neurofibrillary tangles [22], while HS-Aging is pathologically characterized by neuronal cell loss and gliosis in the hippocampus

unilaterally (~50%) or bilaterally [20, 23]. HS-Aging is generally diagnosed postmortem. The large majority of cases with HS-Aging show bilateral TAR DNA-binding protein 43 (TDP-43) pathology in limbic structures [24, 25]. TDP-43 pathology had been considered to be a specific marker for frontotemporal lobar degeneration with ubiquitinated inclusions (FTLD-U). However, TDP-43 pathology is found in both HS-Aging and AD; in one study TDP-43 was detected in 71% of HS-Aging and 23% of AD cases [24]. There is no known treatment or preventive care for HS-Aging so far. Understanding its genetic architecture is important to reduce misdiagnosis with AD and to elucidate the aetiology of HS-Aging, yielding new insights into the molecular-based mechanisms of the underlying developmental process.

Brief history of genetic risks of AD and HS-Aging

Familial AD, which often occurs early in life, is linked mainly to mutations in three genes: *APP* and the presenilin proteins (*PSEN1* and *PSEN2*) [8], which generally cause a shift in A β production from A β ₄₀ to less soluble and more neurotoxic A β ₄₂ (e.g., Volga German mutation in *PSEN2* and Iberian mutation in *APP*) [26-29], an increased total A β levels (Swedish mutation in *APP*) [30], and an increased protofibril formation of A β (Arctic mutation in *APP*) [31]. On the other hand, late-onset AD (LOAD), which often occurs later in life and accounts for 95% of all AD cases [32], has more complex genetic architecture. The ϵ 4 allele of apolipoprotein E (*APOE*) gene is the major genetic risk factor for LOAD. There are three apoE isoforms, apoE2 (cys112, cys158), apoE3 (cys112, arg158), and apoE4 (arg112, arg158), determined by rs429358 (T/C) and rs7412 (C/T) located on chromosome 19q13 (Table 1.1). The risk of AD is increased in

individuals with the $\epsilon 4$ allele: 2 to 3 times in those with one $\epsilon 4$ allele, and more than 12-time in those with two $\epsilon 4$ alleles, whereas the $\epsilon 2$ allele has a protective effect: 0.6 times the odds compared to $\epsilon 3/\epsilon 3$ carriers. These isoforms have different effects on $A\beta$ metabolism, influencing age of onset of $A\beta$ deposition. It is suggested that the binding ability of the apoE isoforms to $A\beta$ follows the order of apoE2, apoE3, and apoE4, and therefore apoE2 and apoE3 inhibit the aggregation and enhance the clearance of $A\beta$ compared to apoE4 [33]. The *APOE* alleles is also reported to be associated with tau levels in CSF [34, 35]. This association, however, has not been established as thoroughly as the association between *APOE* alleles and $A\beta$ deposition [36].

The microtubule-associated protein tau (*MAPT*) gene on chromosome 17q21, encoding tau and containing 16 exons, is also a candidate gene playing an important role in AD development. The tau primary transcript contains 13 exons without exons 4A, 6 and 8 in human brain. Exons 2, 3, and 10 are alternatively spliced, resulting in six different tau isoforms with the range from 352 to 441 amino acids. These isoforms differ by the presence of 0, 1, or 2 N-terminal inserts (0N = exons 2-3-; 1N = exons 2+3-; 2N = exons 2+3+) and either three (3R) or four (4R) microtubule binding repeats located at the C terminus [37]. In the normal brain, the levels of 3R-tau and 4R-tau are approximately equal. The mutations in *MAPT* alter the balance by increasing the ratio of 4R to 3R, and disruption of the normal 4R to 3R ratio is associated with neurodegeneration by accelerating phosphorylation of tau. However, the ratio is approximately 1 in the AD brain with NFTs, and no mutations in *MAPT* have been found to be associated with AD so far. Instead, the *MAPT* haplotypes associated with AD have been found. Two

haplotypes exist in *MAPT*, directly oriented H1 and the inverted H2, which cover the entire *MAPT* gene. These haplotypes are tagged by a 238bp H1 insertion/H2 deletion polymorphism in intron 9 (del-In9). Many researchers reported that the H1 haplotype was associated with risk of LOAD [38-40]. Since the H1 and H2 haplotypes do not alter amino acid sequence, this pathogenic effect of the H1 haplotype may be due to differences in the gene expression rather than tau protein structure [38].

In addition to the *APOE* alleles and *MAPT* haplotypes, a series of genome-wide association studies (GWAS) have identified AD-associated single nucleotide polymorphisms (SNPs) in or close to genes that include *CRI*, *BINI*, *INPP5D*, *MEF2C*, *CD2AP*, *NME8*, *EPHA1*, *PTK2B*, *PICALM*, *SORL1*, *FERMT2*, *SLC24A4-RIN3*, *DSG2*, *CASS4*, *HLA-DRB5-DBR1*, *CLU*, *MS4A6A*, *ABCA7*, *CD33*, *ZCWPW1*, and *CELF1* (Table 1.2) [41-45]. The report with the largest numbers of cases and controls was the International Genomics of Alzheimer's Project (IGAP), a consortium to discover the genetic landscape of AD that included 74,046 individuals to show significant AD-associations with 19 SNPs by meta-analyzing GWAS from four component consortia [41]. Although GWAS have succeeded in revealing numerous susceptibility SNPs for AD, it is an ongoing challenge to identify functional loci and to understand how they contribute to dementia pathogenesis.

Unlike AD, the *APOE* ϵ 4 allele is not a genetic risk factor for HS-Aging [19, 21, 25, 46, 47]. The following four genes (in the chronological order they were so identified) have been reported to harbor risk alleles associated with HS-Aging pathology: *GRN* on

chromosome 17q, *TMEM106B* on chromosome 7p, *ABCC9* on chromosome 12p, and *KCNMB2* on chromosome 3q [21, 48-53]. The T-alleles of *GRN* rs5848 and *TMEM106B* rs1990622 were shown to have a risk of HS-Aging using an allele test, following the known relationship of those two genes to frontotemporal lobar degeneration with TDP-43 inclusions (FTLD-TDP). The connections of the *ABCC9* and *KCNMB2* genes to HS-Aging risk were discovered via GWAS. The association of *ABCC9* SNP rs704180 with HS-Aging pathology was demonstrated using a recessive mode of inheritance (MOI) [51]. Beecham and colleagues reported the *KCNMB2* SNP rs9637454 as the top SNP for HS pathology [48].

Genetic variants are located with much less frequency in coding regions than in non-coding regions (about only 1% are within a protein-coding sequence) [54]. However, it is estimated that about 85% of the mutations with large effects on diseases are located in protein-coding functional regions [55]. To understand disease development mechanisms that underlie disease-associated genetic variants, identifying functional genes and/or variants is an important challenge. Functional variants may be located in nonsynonymous/synonymous coding regions, alternative splice region, and regulatory regions such as promoter, operator, insulator, enhancer and silencer. A nonsynonymous substitution includes a missense and nonsense mutations. The former alters the amino acid sequence of a protein, and the later introduces a premature termination codon resulting in a truncated protein. Many Mendelian diseases are due to nonsynonymous mutations causing deleterious amino acid substitutions. Synonymous mutation occurs in the coding region, but it does not change the amino acid sequence. These variants were

referred to as “silent mutation” until recently [56]. However, several synonymous mutations have been reported to affect mRNA splicing and stability, gene expression, and protein folding and function [56]. Other disease-associated genetic variants are located in the intronic and intergenic regions (i.e., non-coding regions) which may contain the regulatory or splice sites. They may have an important role in regulating expression level of disease-associated genes and modulating translation efficiency and stability [57].

As shown in Table 1.2, all 21 variants identified in IGAP are non-coding. To elucidate the role of these SNPs, we hypothesized that each of the SNPs is: (1) a proxy of a coding variant or (2) a regulatory variant. One frequently used approach for the first hypothesis is to search coding variants in strong linkage disequilibrium (LD) with the variant identified by GWAS. LD is generally measured using the squared correlation coefficient (r^2) between two variants, and the most widely used threshold is $r^2 \geq 0.8$. For the second hypothesis, expression quantitative trait locus (eQTL) analysis can be used. eQTL is a genetic locus that contributes to variation in gene expression. By mapping eQTL, we could investigate how the SNPs regulate gene expression.

Rare variants and next-generation sequencing (NGS)

Rare variants have become a focus in the recent past. Although GWAS have been successful in interrogating genetic variants for association with disease, GWAS are performed under the “common disease – common variant” hypothesis positing that common traits are caused by the combination of common variants with a small to moderate effect [58]. GWAS rely on genotyping preselected SNPs and imputing

ungenotyped variants based on local linkage disequilibrium (LD) of a set of some haplotypes from reference population. Imputation approaches have continually improved and are quite accurate for common variants [59, 60] but not as reliable for rare variants [61]. Therefore, imputed rare variants are typically removed from GWAS analysis.

Recent advances in sequencing technologies have allowed to move toward comprehensive genome-wide approaches, enabling to accurately genotype rare variants generally defined as a variant with minor allele frequency (MAF) $< 1-5\%$. These NGS technologies have the potential to improve our understanding the role of both common and rare variants in the underlying biological mechanisms of developing a disease. Whole-exome sequencing (WES) and whole-genome sequencing (WGS) are ideal approaches to identify novel variants and genes associated with complex traits. Most coding variants, however, are very rare, and thus an extremely large sample size is required to identify a single variant associated with a disease. There are significant computational and statistical challenges for these sequencing studies. Traditional single variant-based association tests, typically used for analysis of common variants, are underpowered to detect rare variants unless sample size and/or effect size is very large [62]. The disease-variant associations may be less accurate if computed by standard regression method for evaluating the effect. Instead of testing single variant individually, more powerful and computationally efficient approaches by aggregating the effects of rare variants have become a standard approach for association testing.

Many such approaches for testing association between rare variants within a pre-specified region and a disease have been proposed. Sequence-kernel association test (SKAT) is one of the most powerful rare variants analysis methods [63, 64]. The SKAT aggregates score test statistics. It is powerful when both risk and protective variants are mixed and when a small proportion of variants are causal [63]. Instead of summing up the square of weighted score test statistics, burden test treats the square of the sum of weighed score test statistics. The burden test is more powerful than SKAT when most of the variants are causal and have the same direction of effect [63].

A recently-proposed scan-statistic-based test is another approach to detect the location of rare variant clusters influencing disease. Scan-statistic-based test was introduced into human genetics by Hoh et al [65] to locate susceptibility genes. Ionita-Laza et al. adapted this test to identify clusters of rare risk variants based on a likelihood ratio under a Bernoulli model proposed by Kulldorff [66] for disease association [67]. Variants within a functional protein-coding domain may be located in close proximity and may play a similar role in genetic mechanisms of a disease. Unlike association tests or other cluster detection analyses, the scan-statistic-based test can both detect the location of clusters and examine the association under the null hypothesis that probability of being a risk variant within a certain scan window equals to that outside the window. This approach is powerful when the disease risk variants significantly make a cluster in the window.

Dissertation Outline

This dissertation work presents studies on gene-based association of genes with hippocampal sclerosis of aging, translation of AD-associated polymorphisms into functional candidates, and genetic regions containing rare variants associated with AD identified by scan statistic-based approach. In Chapter two, gene-based association tests of *GRN*, *TMEM106B*, *ABCC9*, and *KCNMB2* and haplotype-based test of *ABCC9* were performed. The major findings from this study were that the significant gene-based association between the *ABCC9* gene and HS-Aging appeared to be driven by a region with a significant haplotype-based association. In addition, the haplotype of *ABCC9* was associated with decreased *ABCC9* expression. In Chapter three, the role of the non-coding SNPs identified in the IGAP consortium GWAS were elucidated within a systems genetics framework. Systems genetics is a global approach to understand how genetic information flows from DNA to transcripts, proteins, metabolites, and ultimately diseases. Focusing on a causal relationship model in which SNP affects phenotype through mRNA, each IGAP SNP was evaluated whether it is a proxy of a coding variant or whether it is a regulatory variant. In Chapter four, genetic regions which contained risk or protective rare variants associated with AD were identified using a scan-statistic-based approach. The scan statistics with different settings were evaluated in *TREM2* and *TOMM40* as highly-replicated positive controls. Very similar scan statistic values were obtained when we specified the whole genome and chromosome as a large genetic region. The optimized window approach captured almost the entire gene in *TREM2* and the single variant in *TOMM40* as a meaningful cluster. Applying the optimized window approach across the genome, clusters harboring risk or protective variants for AD were

detected including within *MUC6*, *NXNLI*, and *BCAM*. The conclusion of the dissertation and future research interests are discussed in Chapter Five.

Table 1.1. *APOE* isoforms encoded by two single nucleotide polymorphisms

	rs429358		rs7412	
	Codon	Amino acid	Codon	Amino acid
apoE2	T G C	Cysteine	T G C	Cysteine
apoE3	T G C	Cysteine	C G C	Arginine
apoE4	C G C	Arginine	C G C	Arginine

A bold letter represents a nucleotide of each polymorphism

Table 1.2. Single nucleotide polymorphisms for Alzheimer's Disease identified in the previous studies

Gene	Chr	SNP	Position	Annotation	1000 Genomes			References
					MAF	r ²	D'	
<i>CRI</i>	1	rs6656401	207,692,049	Intron	0.17			[41, 43]
		rs3818361	207,784,968	Intron	0.18	0.83	0.94	[42, 45]
		rs6701713	207,786,289	Intron	0.18	0.83	0.94	[44, 45, 68]
		rs1408077	207,804,141	Intron	0.18	0.83	0.93	[45, 69]
<i>BINI</i>	2	rs6733839	127,892,810	Regulatory region variant	0.38			[41]
		rs744373	12,789,4615	Intergenic variant	0.27	0.49	0.90	[42, 45, 70]
		rs7561528	127,889,637	Intergenic variant	0.31	0.35	0.69	[44, 45, 69]
<i>INPP5D</i>	2	rs35349669	234,068,476	Intron	0.46			[41]
<i>MEF2C</i>	5	rs190982	88,223,420	Intron	0.37			[41]
<i>HLA-DRB5-DBR1</i>	6	rs9271192	32,578,530	Intergenic variant	0.26			[41]
<i>CD2AP</i>	6	rs10948363	47,487,762	Intron	0.25			[41]
		rs9296559	47,452,270	Intron	0.25	1	1	[42]
		rs9349407	47,453,378	Intron	0.25	1	1	[42, 44, 68]
<i>NME8</i>	7	rs2718058	37,841,534	Intron	0.37			[41]
<i>ZCWPWI</i>	7	rs1476679	100,004,446	Intron	0.30			[41]
<i>EPHA1</i>	7	rs11771145	143,110,762	Intron	0.36			[41]
		rs11767557	143,109,139	Intron	0.22	0.25	0.71	[42, 44]
<i>PTK2B</i>	8	rs28834970	27,195,121	Intron	0.34			[41]

A bold SNP ID represents the SNP identified the International Genomics of Alzheimer's Project reported by Lambert et al [41]. Chr = chromosome; SNP = single nucleotide polymorphism; 1000 Genomes = 1000 Genomes Project Phase 3 in individuals of European ancestry; MAF = minor allele frequency

Table 1.2. (Continued)

Gene	Chr	SNP	Position	Annotation	1000 Genomes			References
					MAF	r ²	D'	
<i>CLU</i>	8	rs9331896	27,467,686	Intron	0.40			[41]
		rs11136000	27,464,519	Intron	0.39	0.91	0.97	[43, 45]
		rs9331888	27,468,862	5 prime UTR	0.30	0.28	1	[43, 71]
		rs2279590	27,456,253	Non coding transcript exon	0.41	0.83	0.93	[43]
		rs7982	27,462,481	Missense	0.39	0.90	0.97	
		rs7012010	27,448,729	Downstream gene variant	0.28	0.18	0.83	
		rs1532278	27,466,315	Non coding transcript exon variant	0.39	0.91	0.97	[44]
<i>CELF1</i>	11	rs10838725	47,557,871	Intron	0.28			[41]
		rs1057233	47,376,448	3 prime UTR variant	0.32	0.17	0.97	[72]
<i>MS4A</i>	11	rs983392	59,923,508	Downstream intergenic	0.41			[41]
		rs670139	59,971,795	Intron	0.40	0.44	0.97	[42, 44]
		rs4938933	60,034,429	Intergenic	0.40	0.70	0.85	[44]
		rs610932	59,939,307	3 prime UTR variant	0.44	0.72	0.89	[42, 45]
		rs662196	59,942,757	Intron	0.44	0.69	0.88	[45]
		rs583791	59,947,252	Missense	0.48	0.68	0.87	[45]

A bold SNP ID represents the SNP identified the International Genomics of Alzheimer's Project reported by Lambert et al [41]. Chr = chromosome; SNP = single nucleotide polymorphism; 1000 Genomes = 1000 Genomes Project Phase 3 in individuals of European ancestry; MAF = minor allele frequency

Table 1.2. (Continued)

Gene	Chr	SNP	Position	Annotation	1000 Genomes			References
					MAF	r ²	D'	
<i>PICALM</i>	11	rs10792832	85,867,875	Downstream intergenic	0.37			[41]
		rs3851179	85,868,640	Downstream gene variant	0.37	0.99	0.99	[45, 73]
		rs541458	85,788,351	Intergenic	0.32	0.64	0.90	[43]
		rs561655	85,800,279	Upstream gene variant	0.35	0.77	0.92	[44]
<i>SORL1</i>	11	rs11218343	121,435,587	Intron	0.043			[41]
<i>FERMT2</i>	14	rs17125944	53,400,629	Intron	0.081			[41]
<i>SLC24A4-RIN3</i>	14	rs10498633	92,926,952	Intron	0.22			[41]
<i>DSG2</i>	18	rs8093731	29,088,958	Intron	0.012			[41]
<i>ABCA7</i>	19	rs4147929	1,063,443	Intron	0.19			[41]
		rs3764650	1,046,520	Intron	0.11	0.77	0.92	[42, 68]
		rs72973581	1,043,103	Missense	0.053	0.012	1	[74]
<i>CD33</i>	19	rs3865444	51,727,962	Upstream gene variant	0.31			[41, 42, 44]
		rs3826656	51,726,613	Upstream gene variant	0.22	0.13	1	
		rs12459419	51,728,477	Missense	0.31	1	1	
<i>CASS4</i>	20	rs7274581	55,018,260	Intron	0.080			[41]

A bld SNP ID represents the SNP identified the International Genomics of Alzheimer's Project reported by Lambert et al [41]. Chr = chromosome; SNP = single nucleotide polymorphism; 1000 Genomes = 1000 Genomes Project Phase 3 in individuals of European ancestry; MAF = minor allele frequency

CHAPTER TWO

Gene-based association study of genes linked to hippocampal sclerosis of aging neuropathology: *GRN*, *TMEM106B*, *ABCC9*, and *KCNMB2*

Abstract

Hippocampal sclerosis of aging (HS-Aging) is a common neurodegenerative condition associated with dementia. To learn more about genetic risk of HS-Aging pathology, we tested gene-based associations of the *GRN*, *TMEM106B*, *ABCC9*, and *KCNMB2* genes, which were reported to be associated with HS-Aging pathology in previous studies. Genetic data were obtained from the Alzheimer's Disease Genetics Consortium (ADGC), linked to autopsy-derived neuropathological outcomes from the National Alzheimer's Coordinating Center (NACC). Of the 3,251 subjects included in the study, 271 (8.3%) were identified as an HS-Aging case. The significant gene-based association between the *ABCC9* gene and HS-Aging appeared to be driven by a region in which a significant haplotype-based association was found. We tested this haplotype as an expression Quantitative Trait Locus (eQTL) using two different public-access brain gene expression databases. The HS-Aging pathology protective *ABCC9* haplotype was associated with decreased *ABCC9* expression, indicating a possible toxic gain of function.

Introduction

Hippocampal sclerosis of aging (HS-Aging) is a high-morbidity brain disease in people of advanced age [75]. The prevalence of HS-Aging pathology ranges from 5 to 30% in older people in large autopsy series [23, 47, 76, 77]. Clinical signs and symptoms of HS-

Aging are similar to those of Alzheimer's disease (AD) with amnesic memory deficits [20, 21]. Because of the overlapping symptomology, HS-Aging is often clinically misdiagnosed as AD [19-21]. AD is characterized by the accumulation of amyloid plaques and neurofibrillary tangles [22], while HS-Aging is pathologically characterized by neuronal cell loss and gliosis in the hippocampus seen by hematoxylin and eosin (H&E) stain, which can occur unilaterally (~50%) or bilaterally [20, 23]. Whatever the laterality on H&E stain, the large majority of cases with HS-Aging show bilateral TAR DNA-binding protein 43 (TDP-43) pathology in limbic structures [24, 25]. Awareness of this common cause of dementia is rapidly increasing, and we recently recommended a revision of the terminology for describing this disease to cerebral age-related TDP-43 with sclerosis (CARTS) [78]. However, here we will maintain use of the term HS-Aging because the neuropathologic databases we assessed did not include TDP-43 pathologic information until quite recently.

Genetic risk factors for HS-Aging have been recently identified. Unlike AD, the apolipoprotein E (*APOE*) ϵ 4 allele is not a risk factor for HS-Aging [19, 21, 25, 46, 47]. By contrast, the following four genes (in the chronological order they were so identified) have been reported to harbor risk alleles associated with HS-Aging pathology: Granulin (*GRN*) on chromosome 17q, Transmembrane protein 106B (*TMEM106B*) on chromosome 7p, ATP-binding cassette sub-family member 9 (*ABCC9*) on chromosome 12p, and potassium channel subfamily M regulatory beta subunit 2 (*KCNMB2*) on chromosome 3q [21, 48-53].

Alleles near the coding portions of the *GRN* and *TMEM106B* genes were shown to have an association with HS-Aging using an allele test, following the known relationship of those two genes to frontotemporal lobar degeneration with TDP-43 inclusions (FTLD-TDP). Specifically, HS-Aging pathology was associated with the T-allele of the *GRN* single nucleotide polymorphism (SNP) rs5848 [49, 53, 79, 80]. For the other FTLD-related gene, *TMEM106B*, persons with eventual autopsy-proven HS-Aging pathology were more likely to have the T-allele than controls [50, 53, 81]. We confirmed an increase in HS-Aging odds for each copy of the T-allele of *TMEM106B* rs1990622 [51].

The connections of the *ABCC9* and *KCNMB2* genes to HS-Aging risk were discovered via genome-wide association studies (GWAS), which are neither helped nor biased by prior mechanistic hypotheses. The association of *ABCC9* SNP rs704178 with HS-Aging pathology was demonstrated in a GWAS using a recessive mode of inheritance (MOI) [51]. The relationship of this locus with HS-Aging was subsequently tested in a different group of research subjects, and the association was replicated [52]. Beecham and colleagues reported the *KCNMB2* SNP rs9637454 as the top SNP for HS pathology, although this association was not genome wide significant [48], and has not been replicated to date.

In the present study, we examined the associations of these four putative risk SNPs with HS-aging pathology, using genetic data obtained from Alzheimer's Disease Genetics Consortium (ADGC) linked to neuropathological outcomes from the National Alzheimer's Coordinating Center (NACC) [51, 52]. Here we aggregated those data sets

to attain greater statistical power for gene-wide association analyses, for the purpose of understanding better the association of multiple (often co-inherited) gene variants with disease development. Thus, we tested *GRN*, *TMEM106B*, *ABCC9*, and *KCNMB2* for gene-based associations with HS-Aging pathology by aggregating SNPs and indels (small insertions or deletions) on each of those genes. In addition, we focused on the interesting region located around intronic SNP rs704178 on the *ABCC9* gene that was identified in the previous work, and analyzed haplotype associations of the region with HS-Aging pathology and *ABCC9* gene expression.

Material and methods

Study subjects

ADGC genotype data were linked to data from the National Institute on Aging (NIA)-funded 36 AD Centers (ADCs) and NACC registry phenotype information. Of 3,730 subjects with both genotype and autopsy information available to us, those who died at age 60 years or older were included in this study. Cases of HS-Aging were identified as patients who met at least one of the following criteria at autopsy; 1) the primary pathologic diagnosis was hippocampal sclerosis, 2) there was a contributing pathologic diagnosis of hippocampal sclerosis, or 3) medial temporal lobe sclerosis was present at autopsy. We then excluded 180 individuals who had FTLN with ubiquitin-positive inclusions, FTLN with no distinctive histopathology, FTLN-tau, or prion associated disease (Figure 2.1).

Quality control of the ADGC genotype data

Standard quality control (QC) procedures were performed on the ADGC genotype data using PLINK v1.90a [82]. Markers were excluded based on the following criteria: (1) minor allele frequency (MAF) < 1%; (2) call rate per variant (SNPs and indels) < 95%, (3) Hardy-Weinberg equilibrium test in controls < 10^{-5} . (Table 2.1). Samples were excluded based on the following criteria: (1) call rate per individual < 95%, (2) a high degree of relatedness per an estimated proportion of identical by descent (IBD) > 0.1875, (3) excess of ± 3.0 standard deviations of heterozygosity rate. Of the 3,407 individuals after the inclusion and exclusion criteria were applied, 3,330 passed the QC (Figure 2.1).

Identifying ethnic outliers

We performed principal component analysis (PCA) in EIGENSTRAT [83] using a linkage disequilibrium (LD) pruned subset of markers (pairwise $r^2 < 0.2$) from our data merged to 1000 Genomes Project Phase 3 (1000 Genomes) [84] data after removing symmetric SNPs and flipping SNPs discordant for DNA strands between the two datasets. We then plotted the first and second principal components (PCs) for each individuals ($n = 5,834$: 2,504 from 1000 Genomes and 3,330 from the study) using the ggplot2 R package (version 2.2.0) [85] in R (version 3.2; <http://www.r-project.org>). Based on the PC plot, 79 study subjects were removed as ethnic outliers (Figure 2.1 and Figure 2.2). We reran the PCA for the remaining 3,251 European ancestries to derive orthogonal PCs which were used as covariates in the subsequent analyses.

Statistical analysis

Gene-based association analysis

Prior to gene-based association analyses, we performed the single variant association testing using logistic regression assuming each of the three most commonly used MOI (additive, dominant, and recessive) adjusted for age at death, sex and the top three PCs using PLINK v1.90a [82]. Gene-based association analyses were conducted using GATES (Gene-Based Association Test Using Extended Simes Procedure) [86] as implemented in the open-source software Knowledge-Based Mining System for Genome-wide Genetic Studies (KGG; version 3.5) [86]. GATES is a gene-based association test that combines the p-values of variants within a gene obtained from single variant association testing described above. We assigned variants to genes based on their physical positions at the UCSC Genome Browser GRCh37/hg19 human assembly (<https://genome.ucsc.edu/>) [87], and defined gene boundaries as $\pm 5\text{kb}$ from 5' and 3' untranslated regions (UTRs). This gene-based association test adjusts for LD in European super population genotype data from the 1000 Genomes (1000 Genomes EUR) [84]. The input data files to KGG contained four columns: chromosome number, marker ID, marker position, and single variant association p-value. We then obtained overall p-values for the associations of the target genes. Since those who live to advanced old age have a higher risk of HS-Aging pathology [25, 88], there is a possibility that those who died earlier would be always identified as a control even if they have a genetic risk. Therefore, for sensitivity analysis against these possible misclassifications, we further performed these gene-based association tests in cases and controls who died at age 80 years or older. For the gene-based association test, statistical significance level was

defined using the Bonferroni correction, yielding $\alpha = 0.05/(4 \text{ genes} \times 3 \text{ MOI} \times 2 \text{ age groups}) = 0.0021$ for the four examined genes and three MOI.

Haplotype-based association analysis for HS-Aging

After identifying the HS-Aging risk-associated region on the *ABCC9* gene by generating a regional association plot using LocusZoom software [89], we performed additional post hoc haplotype analysis for the variants on the region. First, we selected tag variants using a pairwise SNP tagging approach with $r^2 \geq 0.8$ based on the 1000 Genomes EUR in Haploview version 4.2 [90]. Maximum likelihood estimates of haplotype frequencies were computed using an expectation-maximization (EM) algorithm implemented in the functions `haplo.em` (for overall subjects) and `haplo.group` (for HS-Aging cases and controls) of the `haplo.stats` R package (version 1.7.7) [91] using R (version 3.2; <http://www.r-project.org>). The associations between common haplotypes (the estimated frequencies greater than 1% in entire subjects) and HS-Aging status assuming a recessive MOI were then tested with a haplotype score test adjusted for age at death, sex, and the top three PCs [92] implemented in the function `haplo.score`. The global and haplotype-specific empirical p-values were obtained via 10^7 Monte-Carlo simulations.

*Haplotype-based expression Quantitative Trait Locus (eQTL) analysis for *ABCC9* gene expression*

We examined the association of the haplotypes with *ABCC9* gene expression, focusing on the haplotypes that were identified in association analysis for HS-Aging pathology. We retrieved *ABCC9* gene expression values in human brain and genotype data from two

independent datasets: North American Brain Expression Consortium (NABEC) [93] and United Kingdom Brain Expression Consortium (UKBEC) [94].

In the NABEC dataset, the expression data were available at Gene Expression Omnibus (GEO) public repository (<http://www.ncbi.nlm.nih.gov/geo/>) under the GEO accession GSE36192, consisting of two brain regions (cerebellum and frontal cortex) from 228 neurologically normal donors. The genotype data were obtained from the database of Genotypes and Phenotypes (dbGaP: <http://www.ncbi.nlm.nih.gov/gap>) under the dbGaP study accession phs000249.v2.p1. After the QC procedure with the same settings as we did for the ADGC genotype data was applied, the genotype data were imputed using Michigan Imputation Server (<https://imputationserver.sph.umich.edu/start.html>) [95] with the following parameters: 1000 Genome Phase 3 v5 reference panel, Eagle v2.3 phasing [96], and EUR population. The imputed genotype with posterior probabilities < 0.9 were labeled as missing. Among the 228 NABEC subjects, 130 who died at age 30 years or older and passed the QC were included in the analysis (all of them were US Caucasians).

In the UKBEC dataset, gene expression for ten brain regions (cerebellar cortex, frontal cortex, hippocampus, medulla, occipital cortex, putamen, substantia nigra, thalamus, temporal cortex, and white matter) and genotype data from 134 “neuropathologically normal” individuals were obtained at BRAINEAC website (<http://www.braineac.org/>). The dosage files downloaded from the website (accessed 6/28/2016) were converted into PLINK file format using Genome-wide Complex Trait Analysis (GCTA) software version 1.24.4 [97]. The haplotype-based association analyses on *ABCC9* gene

expression were performed for the five haplotypes that were identified in the haplotype-based association analysis for HS-Aging assuming an additive MOI.

The analyses were carried out separately in the two datasets. We focused on *ABCC9* gene expression through Illumina probe ID ILMN_1751453 in frontal cortex of the NABEC and through Affymetrix transcript ID t3446919 in the average of all ten regions of the UKBEC dataset. Expression data were quantile normalized and log₂-transformed.

Results

Of the 3,251 included subjects from ADGC/NACC, 271 (8.3%) met at least one of the HS-Aging case criteria. Figure 2.3 shows the proportion of participants with HS-Aging pathology increased with age at death, from 3.1% (95% confidence interval (CI) is 1.6 to 5.4%) in those aged less than 70 years to 15.7% (95% CI is 12.8 to 19.0%) in those aged 90 years or older. The mean age at death in the cases was significantly higher than that in the controls (84.8 ± 8.4 years in the cases and 80.5 ± 8.8 years in the controls). No statistically significant differences were noted by case status and sex, *APOE* $\epsilon 4$ and microtubule-associated protein tau (*MAPT*) haplotype (H1 haplotype tagging rs8070723 A-allele and H2 tagging G-allele) frequencies (Table 2.2).

Single variant-based association

Table 2.3 shows the most associated variants on each of the four genes defined gene boundaries as ± 5 kb from 5' and 3' UTRs. The highest association signals came from SNPs on the *ABCC9* gene (rs7966849; $p = 7.1 \times 10^{-6}$ with an assumed recessive MOI and

$p = 4.4 \times 10^{-5}$ with an assumed additive MOI) and on the *KCNMB2* gene (rs73183328; $p = 8.2 \times 10^{-5}$ with an assumed additive MOI and $p = 1.6 \times 10^{-4}$ with an assumed dominant MOI). There was a series of small signals in high LD with the top SNP on the *TMEM106B* gene, and there was an associated region with small effects in low-to-moderate LD with the top SNP on the *ABCC9* gene.

Gene-based association

In the gene-based association analyses, 20, 222, 259 and 939 variants were mapped to the *GRN*, *TMEM106B*, *ABCC9* and *KCNMB2* genes, respectively. Table 2.4 shows the results of the gene-based association test in people aged 60 years. The *ABCC9* gene had a significant gene-based association with HS-Aging assuming a recessive MOI when applying the Bonferroni correction ($p = 2.4 \times 10^{-4}$). There were nominally significant gene-based associations for the *GRN* gene assuming a recessive MOI, the *TMEM106B* gene assuming a recessive and an additive MOI, the *ABCC9* gene assuming an additive MOI, and the *KCNMB2* gene assuming an additive and a dominant MOI. For sensitivity analysis in people aged 80 years or older ($n = 1,883$: 203 in HS-Aging cases and 1,680 in controls), we confirmed the same results that the *ABCC9* gene had a significant gene-based association with HS-Aging assuming a recessive MOI ($p = 0.0017$) (Table 2.5).

Haplotype-based association with HS-Aging

The single-variant-based association plots (Figure 2.4) imply that the significant gene-based association of the *ABCC9* gene is driven by the region in which the most significant variants were located on the position 21,982,262 - 22,015,114 (all

chromosomal positions we describe are referent to human assembly GRCh37/hg19). The top SNP (rs7966849) in this study is in high LD with rs704180 ($r^2 = 0.926$) which was identified as the predominant risk SNP of HS-Aging [51, 52]. Assuming a recessive MOI, there were 33 variants (30 SNPs and 3 indels) associated with HS-Aging pathology (each with $p < 1.0 \times 10^{-3}$) in this region, all of which are intronic. We selected four tag SNPs between exon 18 and 29 (Figure 2.4) of the *ABCC9* gene when assuming a recessive MOI. The most frequent haplotypes were “Hap1” T-A-G-T (from 5’ to 3’) estimated to be present in 40.1% of observed chromosomes (32.1% in cases and 40.8% in controls), and “Hap2” C-C-A-C (36.8%; 43.7% in cases and 36.2% in controls). Hap1 was significantly associated with a lower risk of HS-Aging (score statistic = -2.747 and $p = 0.0061$) and Hap2 with a higher risk of HS-Aging (score statistic = 4.277 and $p = 3.3 \times 10^{-5}$).

Haplotype-based expression Quantitative Trait Locus (eQTL) association with ABCC9 gene expression

In haplotype-based association tests assuming an additive MOI, Hap1 was significantly associated with *ABCC9* gene expression in both datasets ($p = 0.0026$ in the NABEC and $p = 0.024$ in the UKBEC). Compared with the association with rs704180 only, Hap1 had a stronger association with *ABCC9* gene expression in the NABEC (Table 2.6).

Discussion

In the large autopsy dataset derived from multiple research centers, we evaluated the genetic associations of four candidate genes (*GRN*, *TMEM106B*, *ABCC9*, and *KCNMB2*)

for HS-Aging pathology. We found significant gene- and haplotype-based associations of the *ABCC9* gene with HS-Aging, and these approaches provide new insights into the other candidate genes and variants that are associated with HS-Aging. The haplotype made up of the risk alleles at the region (Hap2: C-C-A-C) was significantly overrepresented in HS-Aging cases, and thus could be a risk haplotype, while the opposite haplotype (Hap1: T-A-G-T) was significantly overrepresented in controls, and thus could be a protective risk factor. We further revealed that the protective haplotype (i.e., Hap1) was associated with down-regulation of *ABCC9* gene expression, and the results were consistent in two independent datasets.

Unlike the *TMEM106B* and *GRN* genes, the association between the *ABCC9* gene and FTLT-TDP has never been reported. That is, the *ABCC9* gene could potentially be a key gene on the distinction between FTLT-TDP and HS-Aging pathogenesis. The *ABCC9* gene encodes a transmembrane protein, a part of an ATP-sensitive potassium (K_{ATP}) channel complex. K_{ATP} channel consists of two distinct subunits: an inwardly rectifying K^+ channel (Kir6.x) and a regulatory sulfonylurea receptor (SURx) [98]. When the ATP levels drop due to hypoxia/ischemia or other stressor, vascular smooth muscle cell K_{ATP} channels open to increase K^+ efflux, voltage-activated calcium channels close to block Ca^{2+} entry, and in turn, vasodilatation is induced [99, 100]. Given the critical roles in regulation of vascular tone, K_{ATP} channel dysfunction may be involved in cardio- and cerebrovascular diseases. In mouse experiments, knock-out Kir6.1 (encoded immediately downstream from *ABCC9* on chromosome 12) and *Abcc9* led to hypertension, coronary artery vasospasm, and sudden cardiac death [101, 102]. In addition, Leverenz and

colleagues found in their community-based study that HS-Aging cases were more likely to have history of stroke, small vessel disease, and hypertension than AD cases [47]. Our group also reported that brains with HS-Aging pathology tended to have arteriolosclerosis in multiple cortical and subcortical regions [103]. We note that known mutations in the human *ABCC9* gene lead to a toxic gain of function (“Cantu syndrome”) also are associated with human cerebrovascular pathology - a phenotype of “tortuous cerebral vessels” detected on neuroimaging [104]. These prior studies imply that cerebrovascular factors might be involved in developing HS-Aging via the K_{ATP} channel-dependent activity [105]. In addition, we recently reported that human brain gene expressions that are triiodothyronine (T3) responsive were correlated with the *ABCC9* gene expression, and total T3 levels in cerebrospinal fluid (CSF) were significantly higher in HS-Aging cases than in controls [106]. Prior studies showed links between thyroid hormone (TH) levels and dementia [107-109], as well as TH levels and vascular diseases [110-112]. Therefore it is possible that the *ABCC9* gene variants may help mediate links between TH dysregulation, cerebrovascular disease, and HS-Aging pathology.

The *TMEM106B* gene did not have a significant gene-based association with HS-Aging when applying the Bonferroni correction, but nominal significance was found assuming a recessive and an additive MOI. Van Deerlin and colleagues identified rs1990622 T-allele as a risk factor for FTLD with TDP inclusions (FTLD-TDP) [113]. Here we report that rs3823612, which is in strong LD with rs1990622 ($r^2 = 0.975$), is the variant on the *TMEM106B* gene that is most strongly associated with risk for HS-Aging pathology

assuming a recessive and an additive MOI. However, there are 108 gene variants (96 SNPs and 12 indels) in near perfect LD with the top SNP rs3823612 over the gene (the range of r^2 was from 0.930 to 0.996). Of the 108 variants, rs3173615 is a missense variant on exon 6, rs6460901 is a splice region variant, rs2302634 and rs2302633 are non-coding transcript exon variants, 19 variants are 5' or 3' UTR variants, 10 variants are upstream or downstream gene variants, and the remaining variants are intronic. Yu and colleagues reported that rs1990622 A-allele was associated with more advanced TDP-43 pathology which is the dominant feature of HS-Aging [114]. TDP-43 is also a major disease protein of other neurodegenerative diseases including FTLD and amyotrophic lateral sclerosis (ALS) [115]. Nicholson and colleagues showed that rs3173615 (missense variant on exon 6), dictating the amino acid at codon 185 of threonine (ACC: T185) or serine (AGC: S185), was associated with higher TMEM106B protein levels in *GRN* mutation carriers [116]. Aberrant TDP-43 immunoreactivity is seen in both HS-Aging and FTLD-TDP, and rs1990622 A-allele is reported to be a risk allele of both HS-Aging and FTLD-TDP. However, these two diseases differ in clinical symptoms and pathological characteristics [25, 117].

The SNP on the *KCNMB2* gene that was identified as a possible risk factor is rs9637454 [48], while in the current study we found that rs73183328 was the most strongly associated variant assuming an additive and a dominant MOI. Nominally significant gene-based association of the *KCNMB2* gene with HS-Aging were found assuming an additive and a dominant MOI, although the gene-based associations were not significant when applying the Bonferroni correction. The *KCNMB2* protein is the transmembrane β 2

subunit of the large-conductance Ca^{2+} - and voltage-activated K^+ (BK) channel. The channel is formed by pore-forming α -subunit encoded on the *KCNMA1* gene (chromosome 10) and four β -subunits ($\beta 1$ to $\beta 4$) [118]. The $\beta 2$ subunit induces the BK channel inactivation with the coexpressed α -subunit leading to neuronal excitability by inhibiting K^+ currents [119]. Since inactivating BK channels are found in CA1 hippocampal neurons [120], HS-Aging may be related to the *KCNMB2* gene via a process involving BK channel activation. It seems remarkable that both GWAS-identified putative HS-Aging risk genes (*ABCC9* and *KCNMB2*) encode proteins that modify potassium channels.

There are limitations in this study. Since NACC data are derived from ADCs, the study design is not population-based. Also, HS pathologic diagnoses vary across calendar time and ADCs. Thus, there was probably some misclassification of HS-Aging diagnosis. However, neuropathologic evaluation is the gold standard for HS diagnosis, and thus the problem of misclassification, while ever-present, was minimized as much as possible. We did not obtain dense genetic information on the *GRN* gene. The previously identified SNP rs5848 as a HS-Aging risk SNP was removed in the process of the QC due to high missing rate. Therefore, we could not evaluate the *GRN* gene well in this study.

In summary, we confirmed that the *ABCC9* gene had the significant gene-based association with HS-Aging when assuming a recessive MOI. The significant gene-based association of the *ABCC9* gene is driven by the region in which a significant haplotype-based association was found. Although we did not find statistically significant gene-based

associations of the other three genes (i.e., *GRN*, *TMEM106B*, and *KCMNB2*) with HS-Aging in this study, it does not mean that these genes are not associated with HS-Aging. Single variants may independently affect HS-Aging pathology rather than the entire gene, or there may be interactions between these genes conferring HS-Aging risk via other mechanisms, such as TDP-43 proteinopathies or ion channel dysfunction. In the future, we plan to examine what role the intronic region of the *ABCC9* gene plays in developing HS-Aging pathology, and whether there are single variant-based and gene-based gene-gene interactions among these four genes to HS-Aging.

Funding

This work was supported by the National Cell Repository for Alzheimer's Disease (U24 AG21886), and National Institute on Aging (K25 AG043546, UL1TR000117, and the UK-ADC P30 AG028383).

Table 2.1. Quality control filters for single nucleotide polymorphism

Criteria	# of excluded variants	# of passed variants
MAF < 1%	29,429,731	8,613,031
Call rate per variant < 95%	1,928,184	6,684,847
HWE test in controls < 10^{-5}	19,386	6,665,461

MAF = minor allele frequency; HWE = Hardy-Weinberg equilibrium

Table 2.2. Comparison of selected characteristics between hippocampal sclerosis of aging cases and controls who died at age 60 years or older (n = 3,251)

Variable	Cases n = 271	Controls n = 2,980	p-value
Age at death, mean (SD)	84.8 (8.4)	80.5 (8.8)	<0.001
Sex, n (%)			
Male	124 (45.8)	1,458 (48.9)	0.349
Female	147 (54.2)	1,522 (51.1)	
<i>APOE</i> , n (%) ^a			
-/-	114 (46.0)	1,207 (44.1)	0.749
-/ ϵ 4	109 (43.9)	1,216 (44.4)	
ϵ 4/ ϵ 4	25 (10.1)	314 (11.5)	
<i>MAPT</i> (rs8070723), n (%) ^b			
H1/H1	176 (66.2)	1,773 (60.1)	0.146
H1/H2	77 (28.9)	1,022 (34.7)	
H2/H2	13 (4.9)	154 (5.2)	

^a *APOE* genotype information was available for n = 2,985.

^b *MAPT* genotype information was available for n = 3,215.

SD = standard deviation; *APOE* = apolipoprotein E; *MAPT* = microtubule-associated protein tau.

Table 2.3. Most associated variant with hippocampal sclerosis of aging in four genes using a logistic regression model assuming a recessive/additive/dominant mode of inheritance in people who died at age 60 years or older (n = 3,251)

Gene	MOI	Variant	Risk/protective alleles	RAF in cases	RAF in controls	OR (95% CI) ^a	p-value
<i>GRN</i>	REC	rs72824731	C/G	9.5	8.4	3.88 (1.64 – 9.22)	0.0021
	ADD] rs2879096	T/C	28.6	24.4	1.25 (1.02 – 1.53)	0.032
	DOM					1.38 (1.07 – 1.78)	0.014
<i>TMEM106B</i>	REC] rs3823612	G/C	64.6	56.5	1.53 (1.19 – 1.98)	0.0011
	ADD					1.40 (1.16 – 1.68)	3.6×10^{-4}
	DOM	rs13229988	A/G	64.0	56.4	1.67 (1.16 – 2.40)	0.0062
<i>ABCC9</i>	REC] rs7966849	A/G	60.3	51.2	1.84 (1.41 – 2.40)	7.1×10^{-6}
	ADD					1.46 (1.22 – 1.76)	4.4×10^{-5}
	DOM	rs829080	C/T	59.1	40.9	1.76 (1.18 – 2.62)	0.0057
<i>KCNMB2</i>	REC	rs13091964	T/C	96.1	92.9	1.84 (1.15 – 2.96)	0.011
	ADD] rs73183328	A/G	5.0	2.2	2.42 (1.56 – 3.76)	8.2×10^{-5}
	DOM					2.40 (1.52 – 3.78)	1.6×10^{-4}

^a Adjusted for age at death, sex and the top three principal components

MOI = mode of inheritance; RAF = risk allele frequency; OR = odds ratio; CI = confidence interval; REC = recessive; ADD = additive; DOM = dominant

Table 2.4. Gene-based associations of the target four genes with hippocampal sclerosis of aging assuming a recessive/additive/dominant mode of inheritance in people who died at age 60 years or older (n = 3,251)

Gene	# of variants	Start position	End position	Gene-based p-value		
				REC	ADD	DOM
<i>GRN</i>	20	42,417,491	42,435,470	0.012	0.16	0.090
<i>TMEM106B</i>	222	12,245,848	12,281,890	0.028	0.0089	0.068
<i>ABCC9</i>	259	21,945,324	22,094,628	2.4×10^{-4}	0.0014	0.26
<i>KCNMB2</i>	939	178,249,224	178,567,217	0.57	0.0079	0.016

REC = recessive; ADD = additive; DOM = dominant

Table 2.5. Gene-based associations of the target four genes with hippocampal sclerosis of aging assuming a recessive/additive/dominant mode of inheritance in people who died at age 80 years or older (n = 1,883)

Gene	Gene-based p-value		
	REC	ADD	DOM
<i>GRN</i>	0.24	0.035	0.026
<i>TMEM106B</i>	0.027	0.0032	0.0070
<i>ABCC9</i>	0.0017	0.0099	0.18
<i>KCNMB2</i>	0.23	0.014	0.015

REC = recessive; ADD = additive; DOM = dominant

Table 2.6. Haplotype association with *ABCC9* gene expression in human brain assuming an additive mode of inheritance

	NABEC		UKBEC	
	(Frontal cortex; n = 130 brains)		(10 brain regions; n = 134 brains)	
	Score statistic ^a	p-value	Score statistic ^a	p-value
Hap1	-2.968	0.0026	-2.250	0.024
Hap2	1.450	0.15	1.740	0.081
Hap3	1.686	0.091	-0.048	0.96
Hap4	0.214	0.83	-0.822	0.41
Hap5	0.878	0.38	1.952	0.051
Global	10.255	0.034	8.455	0.074
rs704180 only		0.010		0.011

^a A positive sign indicates up-regulation of *ABCC9* gene expression and vice versa.

NABEC = North American Brain Expression Consortium (GEO accession: GSE36192);

UKBEC = United Kingdom Brain Expression Consortium (<http://www.braineac.org/>).

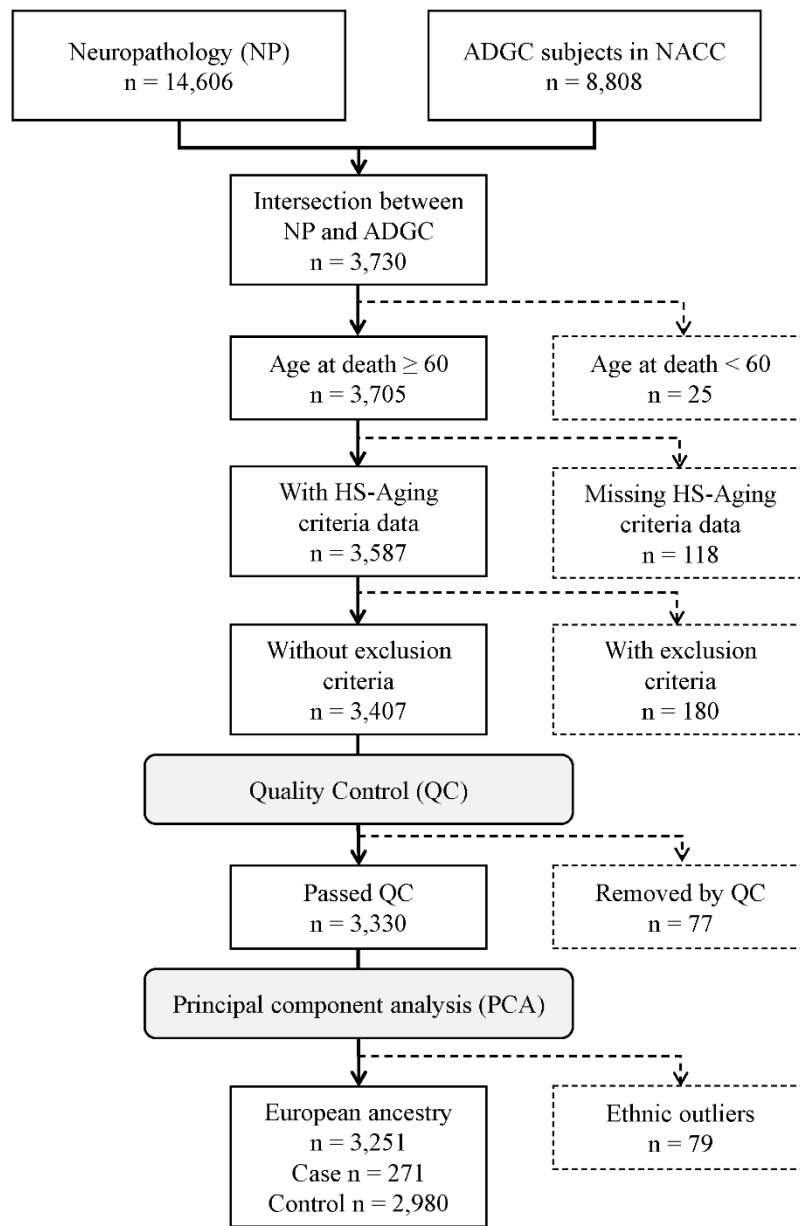


Figure 2.1. Flow diagram of the subjects included in the analyses. Genetic data were obtained from subjects in ADGC who had the NACC individual IDs. Phenotype data were available from the neuropathological dataset in NACC. The inclusion/exclusion criteria, quality control and removal of ethnic outliers were applied in order. ADGC = Alzheimer’s Disease Genetics Consortium; NACC = National Alzheimer’s Coordinating Center; NP = neuropathological dataset; HS-Aging = hippocampal sclerosis of aging

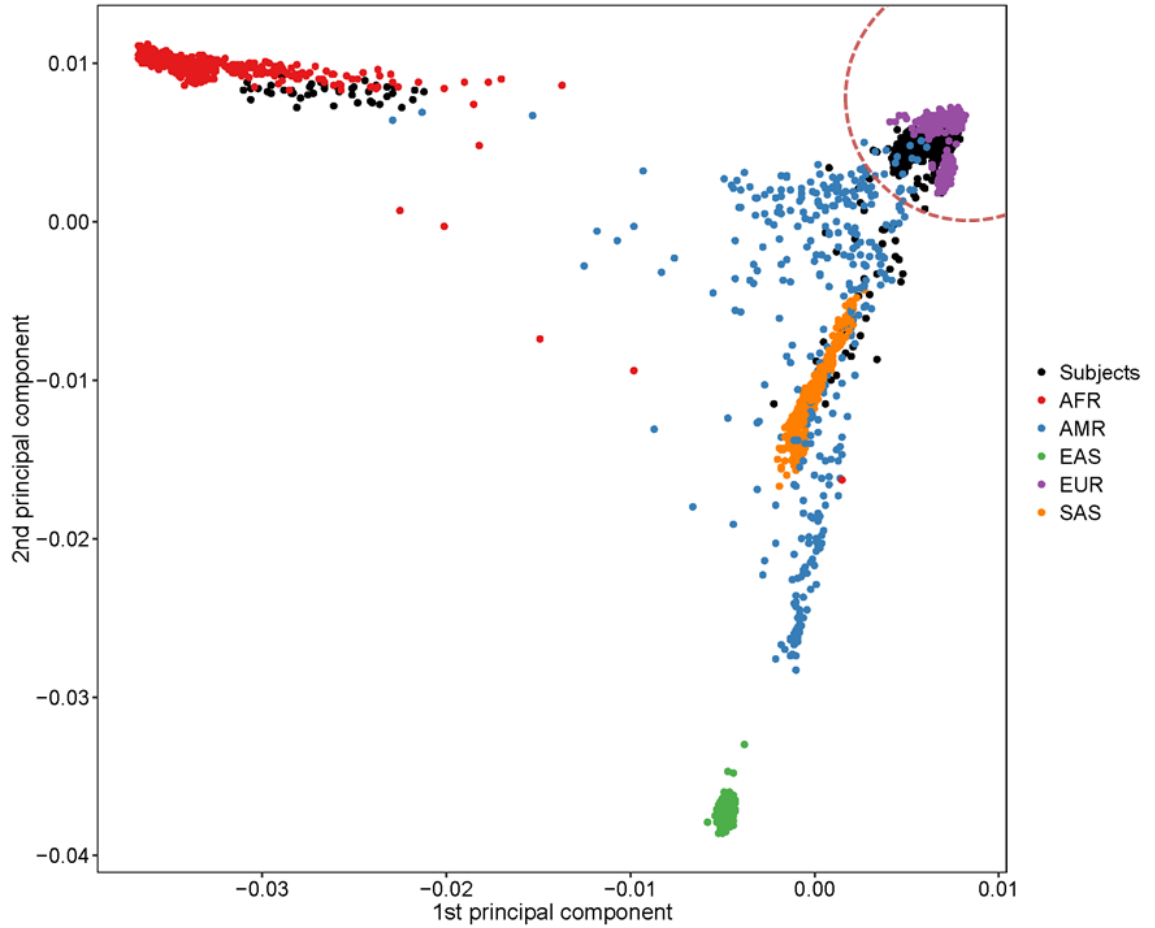


Figure 2.2. The first and second principal components plots along with 1000 genome reference samples. Block dot indicates individuals in this study. We chose individuals within the red dotted circle based on Euclidean distance from an individual with maximum first and second principal components.

AFR = African; AMR = Admixed American; EAS = East Asian; EUR = European; SAS = South Asian

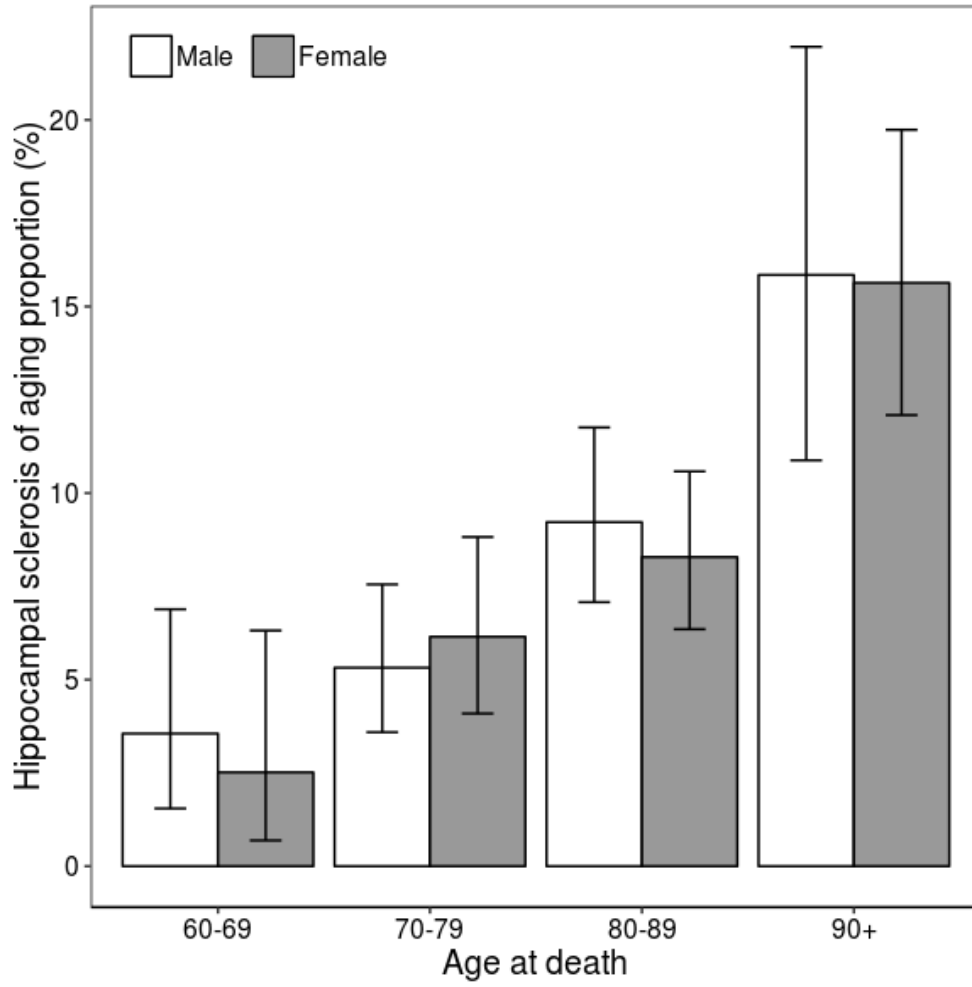
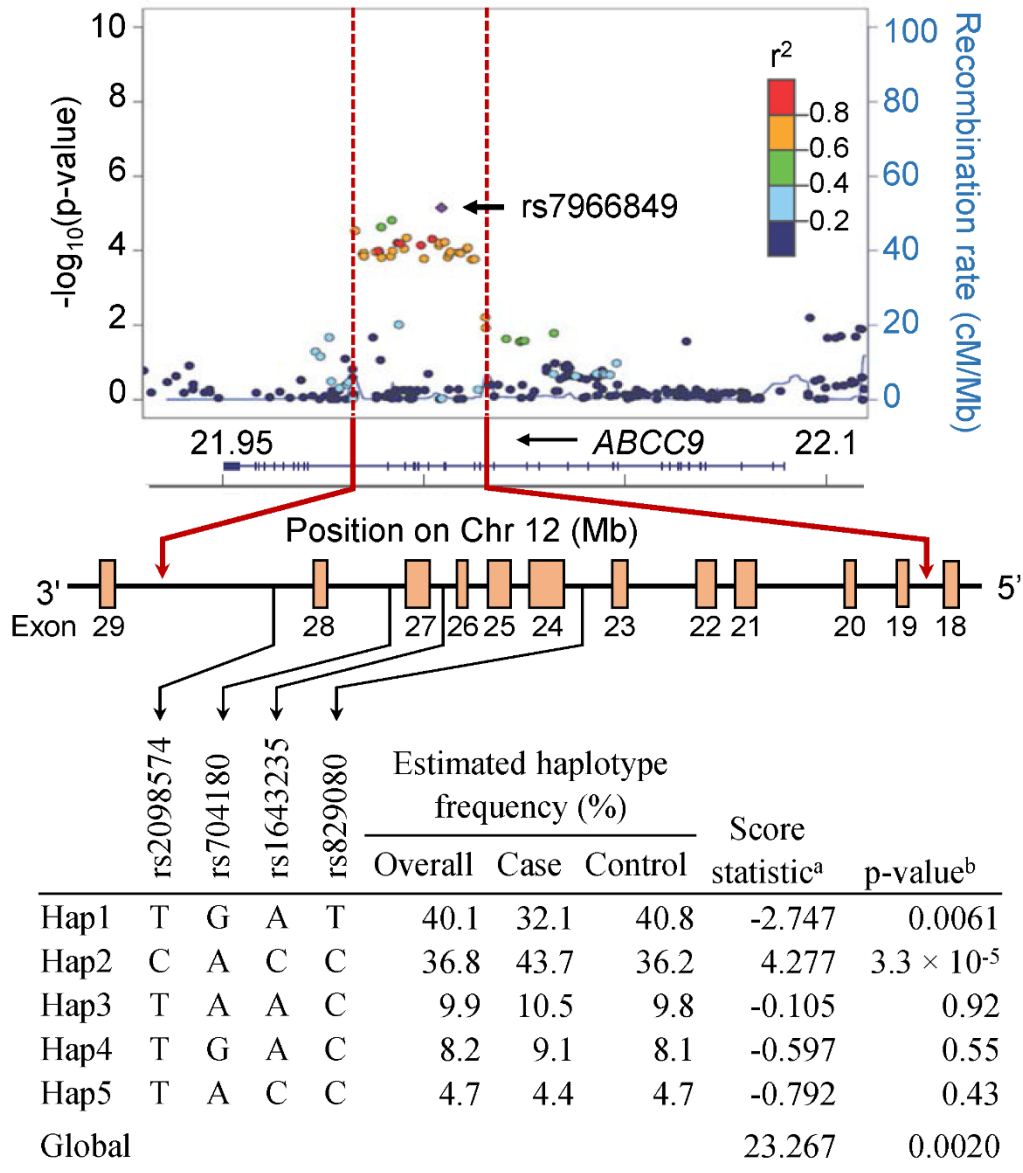


Figure 2.3. Proportion and 95% confidence interval of hippocampal sclerosis of aging cases



^a Adjusted for age at death, sex and the top three principal components assuming recessive mode of inheritance. A positive sign indicates a risk haplotype for development of HS-Aging and vice versa.

^b Monte-Carlo p-value from 10^7 replications

Figure 2.4. Estimation of haplotype frequencies and association using four tag single nucleotide polymorphisms on the *ABCC9* gene when assuming a recessive mode of inheritance. Box indicates an exon.

CHAPTER THREE

Translating Alzheimer's disease risk polymorphisms into functional candidates

: a survey of IGAP genes

Abstract

Alzheimer's disease (AD) is the most common form of dementia. The report with the largest numbers of cases and controls was the International Genomics of Alzheimer's Project (IGAP), a consortium to discover the genetic landscape of AD that included 74,046 individuals to show significant AD-associations with 19 SNPs. However, we have relatively little understanding of the functional impact of these loci in regards to AD pathogenesis. In this study, we elucidated the role of the non-coding SNPs identified in IGAP hypothesizing that each IGAP SNP is a proxy of a coding variant and/or a regulatory variant. Our genetic data were obtained from the Alzheimer's Disease Sequencing Project (ADSP). For the first hypothesis, rs2296160 in *CRI*, rs9270303, rs1049092, and rs1049086 in *HLA-DRB5*, rs2405442 and rs1859788 in *ZCWPW1*, rs7982 in *CLU*, rs12453 and rs7232 in *MS4A6A*, and rs3752246 in *ABCA7* may be proxies of coding SNPs. For the second hypothesis, rs6656401 in *CRI*, rs10838725 in *CELF1*, and rs8093731 in *DSG2* may be regulatory SNPs affecting AD-associated gene expression. Our approach for identifying proxies and examining eQTL lessens the impact of the crude gene assignment, although this still remains an open question in the field.

Introduction

Dementia is a clinical state caused by neurodegeneration and characterized by a loss of function in cognitive domains and behavior. Alzheimer's disease (AD) is the most common form of dementia, accounting for over 50% of dementia cases [1]. Although it has been more than 100 years since Alois Alzheimer published "About a Peculiar Disease of the Cerebral Cortex" in 1907 [3], the exact cause of AD is yet to be identified.

Amyloid β ($A\beta$) protein and hyperphosphorylated tau aggregates in the brain are considered to be the key pathological hallmarks in AD patients [4, 5]. A predominant mechanistic hypothesis for AD pathogenesis is the "amyloid cascade hypothesis" that suggests that AD is caused by lack of $A\beta$ clearance, which triggers downstream neuronal injury such as synaptic and neuronal loss, enhanced neuroinflammation, tau hyperphosphorylation, and eventually the clinical symptoms of AD [6].

Familial AD, which often occurs early in life, is linked mainly to mutations in three genes: amyloid precursor protein (*APP*) and the presenilin proteins (*PSEN1* and *PSEN2*) [8], which generally cause a shift in $A\beta$ production from $A\beta_{40}$ to less soluble and more neurotoxic $A\beta_{42}$ (e.g., Volga German mutation in *PSEN2* and Iberian mutation in *APP*) [26-29], an increased total $A\beta$ levels (Swedish mutation in *APP*) [30], and an increased protofibril formation of $A\beta$ (Arctic mutation in *APP*) [31]. On the other hand, late-onset AD (LOAD), which accounts for 95% of all AD cases [32], has a more complex genetic architecture. The $\epsilon 4$ allele of apolipoprotein E (*APOE*) gene is the most well-established susceptibility gene for LOAD. There are three apoE isoforms, apoE2 (cys112, cys158), apoE3 (cys112, arg158), and apoE4 (arg112, arg158), determined by two single

nucleotide polymorphisms (SNPs) rs429358 (T/C) and rs7412 (C/T) located on chromosome 19q13. These isoforms have different effects on A β metabolism, influencing age of onset of A β deposition. It is suggested that the binding ability of the apoE isoforms to A β follows the increasing order of apoE2, apoE3 and apoE4, and thus apoE2 and apoE3 inhibit the aggregation and enhance the clearance of A β compared to apoE4 [33]. The *APOE* alleles are also reported to be associated with tau levels in CSF [34, 35]. This association, however, has not been established as thoroughly as the association between *APOE* alleles and A β deposition [36].

A series of genome-wide association studies (GWAS) have identified AD-associated SNPs in addition to the *APOE* alleles. The report with the largest numbers of cases and controls was the International Genomics of Alzheimer's Project (IGAP), a consortium to discover the genetic landscape of AD that included 74,046 individuals to show significant AD-associations with 19 SNPs by meta-analyzing GWAS from four component consortia [41]. The SNPs are in or close to genes that include *CRI*, *BIN1*, *INPP5D*, *MEF2C*, *CD2AP*, *NME8*, *EPHA1*, *PTK2B*, *PICALM*, *SORL1*, *FERMT2*, *SLC24A4-RIN3*, *DSG2*, *CASS4*, *HLA-DRB5-DBR1*, *CLU*, *MS4A6A*, *ABCA7*, *CD33*, *ZCWPW1*, and *CELF1* (Table 1.2). Although GWAS have succeeded in revealing numerous susceptibility variants for AD, it is difficult to determine whether the genes and SNPs at these loci are functional and to understand how they contribute to AD pathogenesis.

Genetic variants located in coding regions are much less frequent than those in non-coding regions (about only 1% of variants are within a protein-coding sequence) [54].

However, it is estimated that about 85% of the mutations with large effects on diseases are located in protein-coding functional regions [55]. To understand disease development mechanisms that underlie AD-associated genetic variants, identifying functional genes and/or variants is an important challenge. Functional variants may be located in a coding region, an alternative splicing region, or a regulatory region such as promoter, operator, insulator, enhancer or silencer. Nonsynonymous variants may have effects on the protein structure and function. Many Mendelian diseases are due to nonsynonymous mutations causing deleterious amino acid substitutions. Synonymous mutations occur in the coding region but do not change the amino acid sequence. These variants were referred to as “silent mutations” until recently [56]. Several synonymous mutations have been reported to affect mRNA splicing and stability, gene expression, and protein folding and function [56]. Other disease-associated genetic variants are located in the intronic and intergenic regions (i.e., non-coding regions) which may contain regulatory or splice sites. Intronic and intergenic variants may have an important role in regulating expression level of disease-associated genes and modulating translation efficiency and stability [57].

In this study, we elucidated the role of the non-coding SNPs identified in the IGAP consortium (hereinafter referred to as “IGAP SNPs”) within a systems genetics framework. Systems genetics is a global approach to understand how genetic information flows from DNA to transcripts, proteins, metabolites, and ultimately diseases [121]. Figure 3.1A shows three possible pathways linking a SNP, a transcript (mRNA), and a phenotype [121-123]. The first model is the causal relationship model in which the SNP affects phenotype by acting through mRNA. Second, the reactive model proposes that the

SNP requires the phenotype to affect the mRNA. The third model is the independent model in which the SNP affects mRNA and phenotype independently. Focusing on these models in this study, we hypothesized that each IGAP SNP is: (1) a proxy of a coding SNP (Figure 3.1B and 3.1C) or (2) a regulatory SNP (Figure 3.1D). One frequently used approach to test the first hypothesis is to identify coding SNPs in strong linkage disequilibrium (LD) with the SNP identified by GWAS. LD is generally measured using the squared correlation coefficient (r^2) between two SNPs, and the most widely used threshold is $r^2 \geq 0.8$. For the second hypothesis, expression quantitative trait locus (eQTL) analysis can be used. eQTL is a genetic locus that contributes to variation in gene expression. By mapping eQTL, we investigate how the SNPs regulate gene expression.

Material and methods

Genetic datasets

Our genetic data were obtained from the Alzheimer's Disease Sequencing Project (ADSP) with whole exome sequence (WES) data to limit the possibility of imputation errors. ADSP is comprised of 18 cohorts from the Alzheimer's Disease Genetic Consortium (ADGC) and 6 from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium (Table 3.1). There were 10,913 unrelated subjects with WES data in ADSP. For our study, we limited the subjects to those who had AD diagnosis information and who were 65 years or older at the last visit or at death, yielding a total of 10,468 ADSP subjects with WES data (Figure 3.2).

Gene expression datasets

Quality-controlled microarray gene expression from blood samples and whole genome sequence (WGS) datasets were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (available through <http://adni.loni.usc.edu>). We included 661 subjects aged 65 years or older who had both gene expression and WGS data available. We considered AD diagnosis (normal, mild cognitive impairment (MCI), or AD) closest to the year when the blood sample was drawn.

We retrieved human brain gene expression and genotype dataset from the North American Brain Expression Consortium (NABEC) [93] and United Kingdom Brain Expression Consortium (UKBEC) [94]. Details are described in our previous report [124]. Briefly, the NABEC gene expression data in two brain regions (cerebellum and frontal cortex) were available at Gene Expression Omnibus (GEO) public repository and the genotype data were obtained from the database of Genotypes and Phenotypes (dbGaP: <http://www.ncbi.nlm.nih.gov/gap>). After performing standard QC procedures, we imputed the genotype data using the Michigan Imputation Server (<https://imputationserver.sph.umich.edu/start.html>) [95, 96] with the following parameters: 1000 Genome Phase 3 v5 reference panel, Eagle v2.3 phasing [96], and EUR population. Of the 228 neurologically normal donors, 119 subjects who died at age 65 years or older and passed QC were included in the analysis (all were US Caucasians). The UKBEC gene expression in ten brain regions (cerebellar cortex (CRBL), frontal cortex (FCTX), hippocampus (HIPPI), medulla (MEDU), occipital cortex (OCTX), putamen (PUTM), substantia nigra (SNIG), thalamus (THAL), temporal cortex (TCTX),

and white matter (WHMT)) and the genotype data in 134 neuropathologically normal individuals were obtained from the BRAINEAC website (<http://www.braineac.org/>). The dosage genotype data were converted into PLINK file format using Genome-wide Complex Trait Analysis (GCTA) software version 1.24.4 [97].

Since the NABEC and UKBEC datasets do not have AD diagnosis information, we obtained two datasets to examine whether the levels of gene expressions were different between AD statuses. The first dataset was derived from AD cases and controls available at Gene Expression Omnibus (GEO) public repository (<http://www.ncbi.nlm.nih.gov/geo/>) under the GEO accession GSE5281 [125]. The gene expression data consisted of 9 AD cases and 13 controls in entorhinal cortex (EC), 10 AD cases and 12 controls in hippocampus (HIPPO), 16 AD cases and 12 controls in medial temporal gyrus (MTG), 8 AD cases and 13 controls in posterior cingulate (PC), 23 AD cases and 11 controls in superior frontal gyrus (SFG), and 19 AD cases and 12 controls in primary visual cortex (VCX). The range of age at death was 63 to 102. The second was Allen Institute data downloaded at <http://aging.brain-map.org/> derived from 38 AD cases and 47 controls (AD diagnosis was based on NINCDS-ADRDA Alzheimer's Criteria) in white matter (FWM), hippocampus (HIPPO), parietal cortex (PCx), and temporal cortex (TCx) brain regions. The range of age at death was 77 to 100+.

We excluded probes in Affymetrix that targeted transcripts from different genes (i.e., probes with “_x” suffix) if a more reliable probe was available. We also excluded monoallelically expressed genes including genes on chromosomes X and Y, and HLA-

genes (i.e., *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DMA*, *HLA-DMB*, *HLA-DOA*, *HLA-DOB*, *HLA-DPA1*, *HLA-DPBI*, *HLA-DQA1*, *HLA-DQA2*, *HLA-DQB1*, *HLA-DQB2*, *HLA-DRA*, *HLA-DRB1*, *HLA-DRB5*, *HLA-E*, *HLA-F*, *HLA-G*, *HLA-J*, *HLA-P*, and *HLA-T*).

Expression data were normalized and log₂-transformed.

Statistical analysis

Hypothesis 1: identified IGAP SNPs are proxies of coding SNPs

For each of the 21 IGAP SNPs (including *CD33* and *DSG2*, although the SNPs rs3865444 and rs8093731 were reported not to reach statistical significance on meta-analysis in IGAP [41]), we first identified SNPs in the nearby coding-regions showing strong LD ($r^2 \geq 0.8$) and moderate LD ($0.4 \leq r^2 < 0.8$) by using 1000 Genomes Project Phase 3 in individuals of European ancestry (1000 Genomes EUR) [84]. We performed association tests under an additive mode of inheritance (MOI) assumption for the coding SNPs, using logistic regression adjusted for age at the last visit or death, sex, and the top 5 principal components (computed in PLINK v1.90a [82]). The pathogenic nature of nonsynonymous SNPs associated with AD was predicted by SIFT (<http://sift.jcvi.org/>) [126] and PolyPhen-2 (<http://genetics.bwh.harvard.edu/pph2/>) [127] which are *in silico* algorithm tools to predict the effect of amino acid substitution on a protein function.

Hypothesis 2: identified IGAP SNPs are regulatory SNPs

We evaluated whether the IGAP SNPs were *cis*- or *trans*-eQTL on the same chromosome as genes. We defined a locus within 1Mb of the 5' or 3' ends of the gene as *cis*-eQTL, and a locus more than 1Mb away from the transcription site as a *trans*-eQTL (Figure 3.1E).

We first tested the association between each of the IGAP SNPs and each of all the gene expression profiles, assuming an additive MOI as implemented in PLINK v1.90a [82]. We then examined whether the levels of gene expression modified by the IGAP SNPs were different between AD statuses. An analysis of covariance (ANCOVA) with sex and age as covariates was applied to each comparison for all gene expressions identified in the eQTL analysis.

For all analyses, we converted the nominal p-values into false discovery rate adjusted p-values (FDR adjusted p-value) using the Benjamini-Hochberg procedure [128] and defined associations with FDR adjusted p-value < 0.05 as significant.

Results

We considered individuals with either prevalent or incident AD at baseline (year 0) as AD cases in ADSP. Descriptive characteristics of the individuals are shown in Table 3.2. 5,374 (51.3%) were AD cases.

Hypothesis 1: identified IGAP SNPs are proxies of coding SNPs

We identified 10 exonic SNPs which were in strong LD ($r^2 \geq 0.8$) and 16 SNPs in moderate LD ($0.4 \leq r^2 < 0.8$) with the IGAP SNPs based on 1000 Genomes EUR (Table 3.3). We confirmed that the several exonic SNPs demonstrated statistically significant associations with AD after FDR adjustment, including rs2296160 in *CRI*, rs9270303, rs1049092, and rs1049086 in *HLA-DRB5*, rs2405442 and rs1859788 in *ZCWPW1*, rs7982 in *CLU*, rs12453 and rs7232 in *MS4A6A*, and rs3752246 in *ABCA7* (Table 3.4). The

association between rs4844600 in *CR1* and AD was not confirmed because of the lack of WES data in ADSP. Of these 10 coding SNPs, 5 SNPs (rs2296160 in *CR1*, rs9270303 in *HLA-DRB1*, rs1859788 in *PILRA*, rs3752246 in *ABCA7*, and rs7232 in *MS4A6A*) are nonsynonymous (Table 3.5). When we analyzed the nonsynonymous SNPs with SIFT and Polyphen-2, rs2296160, rs9270303, rs1859788, and rs3752246 were predicted to have minimal impact on their respective proteins (Table 3.5). In contrast, the minor allele of rs7232 was predicted to be deleterious and possibly damaging to *MS4A6A* protein.

Hypothesis 2: identified IGAP SNPs are regulatory SNPs

Table 3.6 shows gene expressions in the blood that were significantly associated with the IGAP SNPs, reaching FDR adjusted significance level. We found that the protective allele of rs1476679 in *ZCWPWI* was strongly associated with decreased expression of multiple *PILRB* probe sets and *TRIM4* expression, and associated with increased expressions of *ZKSCAN11*, *GATS*, and *PVRIG*. The protective allele of rs11771145 in *EPHA1* was associated with increased *LOC154761* expression. The risk allele of rs28834970 in *PTK2B* exhibited *cis*-eQTL for multiple probe sets for its own expression and for the contiguous gene *TRIM35*. This allele also had trans-association with *NSAP11*. The risk allele of rs10838725 was associated with increased *MYBPC3* expression. The protective allele of rs983392 was associated with decreased expressions of *MS4A6A* and another family members, *MS4A4A*. The protective alleles of rs8093731 in *DSG2* and rs7274581 in *CASS4* appeared to be *trans*-eQTLs. Of these significant gene expressions in the blood, only *MYBPC3* (11725151_at) potentially regulated by the *CELF1* SNP was significantly associated with AD status.

The FDR adjusted significant associations between the IGAP SNPs and brain gene expressions in NABEC and UKBEC are shown in Table 3.7. In NABEC, the protective allele of rs10792832 in *PICALM* was significantly associated with increased expression of *MRGPRD* in CRBL. The protective allele of rs8093731 in *DSG2* exhibited *trans*-eQTL for the expression of three genes, *DLGAP1*, *NETO1*, and *KCNG2*. The minor allele of rs7274581, which is also protective, was significantly associated with the *PCK1* expression in CRBL. Of these patterns, the *DLGAP1* and *NETO1* expressions in FCTX were highly correlated ($r^2 = 0.78$) (Figure 3.3). In UKBEC, rs6656401 in *CRI* acted as an eQTL for several genes, *COL9A2*, *CERS2* (also known as *KASS2*), *ARHGEF2*, *CNTN2*, and *CDK18* (also known as *PCTK3* and *PCTAIRE3*) in MEDU, and *CRI* itself in the average of all ten regions (AveALL) and WHMT. The genes potentially regulated by the *CRI* SNPs were highly correlated with each other except *CRI* expression ($r^2 = 0.68$ to 0.89) (Figure 3.4).

In comparison with AD controls using GSE5281, AD cases had significantly lower expressions of *DLGAP1* in MTG and SFG, *NETO1* in EC, MTG, PC, and SFG, *KCNG2* in MTG, *CRI* in MTG, and *ARHGEF2* in EC. On the other hand, AD cases had significantly higher expressions of *DLGAP1* in HIP, *COL9A2*, *CERS2*, and *CRI* in EC, and *ARHGEF2* in PC (Table 3.8 and Table 3.9). For the Allen Institute dataset, significantly higher expressions of *KCNG2*, *CNTN2*, *ARHGEF2*, and *CDK18* in AD cases have seen only in TCx (Table 3.10).

Discussion

Although recent studies have identified novel GWAS loci that affect AD risk, we have relatively little understanding of the functional impact of these loci in regards to AD pathogenesis. In this study, we examined the possible functional effects of the IGAP SNPs on AD under two hypotheses: “the IGAP SNP is a proxy of a coding SNP” and “the IGAP SNP is a regulatory SNP”. For the first hypothesis, rs2296160 in *CRI*, rs9270303, rs1049092, and rs1049086 in *HLA-DRB5*, rs2405442 and rs1859788 in *ZCWPW1*, rs7982 in *CLU*, rs12453 and rs7232 in *MS4A6A*, and rs3752246 in *ABCA7* are proxies of coding SNPs. For the second hypothesis, rs6656401 in *CRI*, rs10838725 in *CELF1*, and rs8093731 in *DSG2* are associated with gene expression, although whether these SNPs are proxies for the functional regulatory SNP or functional themselves requires further studies.

Hypothesis 1: identified IGAP SNPs are proxies of coding SNPs

CRI SNPs

The IGAP SNP rs6656401 in *CRI* was the most striking SNP in this study for several reasons. First, there were 2 coding SNPs (one is synonymous and the other is nonsynonymous) in strong LD with the IGAP SNP. Second, the IGAP SNP acted as *cis*- or *trans*-eQTL for the several gene expressions in the brain. Last, these expressions were associated with AD. *CRI*, located on chromosome 1q32.2, encodes complement receptor 1 in a cluster of complement-related proteins which plays an important role in the immune system [129]. The CR1 protein is a receptor for complement fragments binding to A β , and thus the change in CR1 protein structure and expression levels may be related

to A β clearance [130]. We found the two coding SNPs, rs4844600 and rs2296160, which were in strong LD with the IGAP SNP rs6656401. The coding SNP rs2296160 is nonsynonymous, causing alanine-to-threonine amino acid substitution at codon position 2419 (A2419T), while the SNP rs4844600 is synonymous (E60E). The IGAP SNP as well as the coding SNPs acted as *cis*-eQTL for *CR1* expression itself (Figure 3.5). The most associated SNP with the *CR1* expression was the nonsynonymous SNP rs2296160. This may imply that the change in CR1 protein structure may change the expression level of the gene itself through an autoregulatory mechanism.

Furthermore, the higher *CR1* expression detected by the probe set 244313_at in EC increased the risk of AD in GSE5281 dataset. Consistent with our findings, the SNP rs1408077 in *CR1*, which is in strong LD with the IGAP SNP rs6656401, was reported to be associated with loss of EC thickness [69], and the IGAP SNP rs6656401 A carriers had smaller local gray matter volume in EC of young health adults which may lead to an increased risk of LOAD [131]. These results may indicate that there is a causal relationship between *CR1* SNPs, *CR1* expression, and AD development, although we cannot mention which model this relationship is on: causal model or reactive model in Figure 3.1.

Other gene expressions than *CR1* potentially regulated by *CR1* SNPs were highly correlated with each other in MEDU in UKBEC. These genes are highly expressed in oligodendrocytes (http://web.stanford.edu/group/barres_lab/brainseqMariko/brainseq2.html) [132]. The

COL9A2 and *CERS2* genes were over-expressed in EC of AD cases. The *ARHGEF2* gene was over-expressed in PC and TCx of AD cases and under-expressed in EC of AD cases. The genes *CNTN2* and *CDK18* were significantly over-expressed in TCx of AD cases. In future work, gene co-expression network analysis will be required to understand how these gene expressions affect in each of the brain regions of AD cases.

ZCWPWI SNPs

ZCWPWI (zinc finger, CW type with PWWP domain 1) is located on chromosome 7q22.1. We found two coding SNPs, rs2405442 and rs1859788, which were in strong LD with the IGAP SNP rs1476679. The coding SNP rs1859788 is nonsynonymous, causing glycine-to-arginine amino acid substitution at codon positions 78 (G213R). The IGAP SNP as well as the coding SNPs acted as *cis*-eQTLs for three gene expressions (*GATS*, *TRIM4*, and *PILRB*) in the blood, although we did not find significant associations between these blood gene expressions and AD.

MS4A6A SNPs

MS4A6A, located on chromosome 11q12.2, encodes a member of membrane-spanning 4A gene family (membrane-spanning 4A domains, subfamily A, member 6A). We also found two coding SNPs, rs12453 and rs7232, which were in strong and moderate LD with the protective IGAP SNP rs983392. The coding SNP rs7232 is nonsynonymous, causing threonine-to-serine amino acid substitution at codon positions 213 (T213S), while the SNP rs12453 is synonymous (L137L). The IGAP SNP as well as the coding SNPs affected *MS4A6A* expression itself and its gene family *MS4A4A* in the blood.

However, we did not find FDR adjusted significant associations between these blood gene expressions and AD. *MS4A* genes are highly expressed in hematopoietic cells, and involved in the regulation of calcium signaling [133]. Although the function of *MS4A6A* protein are still unknown, it is possible that the *MS4A6A* SNPs is linked to AD via deregulation of calcium signaling implicated in neurodegenerative diseases [134, 135].

CLU SNPs

We confirmed synonymous SNP rs7982 was in strong LD with the IGAP SNP rs9331896 and was protectively associated with AD. However, we found no significant gene expressions regulated by the *CLU* IGAP SNP and the synonymous SNP. *CLU*, also known as apolipoprotein J, is located in chromosome 8p21.1, and encodes clusterin. Clusterin directly influences A β , regulating the conversion of A β into insoluble forms [136, 137]. *CLU* has mainly two isoforms, nuclear *CLU* (*nCLU*, isoform 1) and secretory *CLU* (*sCLU*, isoform 2) with different functions. The *sCLU* form is pro-survival, while *nCLU* is pro-apoptotic [138]. Since the coding SNP rs7982 is synonymous, it would affect alternative splicing as Ling et al. showed that the protective SNP rs11136000 (which is in almost perfect LD with rs7982 in 1000 genomes EUR) was associated with increased *nCLU* expression level [139]. We would need to examine how the isoforms affects A β clearance in future.

HLA-DRB5-DRB1 SNPs

The IGAP SNP rs9271192 is located in intergenic region (chromosome 6p21.32), contiguous to HLA class II genes (*HLA-DR*, *-DQ* and *-DP*). There were two coding

SNPs which were in strong or moderate LD with the IGAP SNP: the nonsynonymous rs9270303 is in *HLA-DRB1* and the synonymous rs1049092 is in *HLA-DQB1*. Several HLA-DR and HLA-DQ genes are monoallelically expressed. There are three classes of monoallelically expressed genes [140, 141]. One is the autosomal imprinted genes regulated in a parent-of-origin specific manner. The second one is X-inactivated. The last class is for randomly monoallelically expressed genes in autosome, in which several immune system genes are included [140, 141]. Given epigenetic association between DNA methylation in *HLA-DRB5* and AD pathology [142], allele specific expression may impact on biological function related to AD.

ABCA7 SNPs

We confirmed that nonsynonymous SNP rs3752246 was in strong LD with the IGAP SNP rs4147929 and was associated with AD risk. However, we found no significant gene expressions regulated by the *ABCA7* IGAP SNP or the nonsynonymous SNP. *ABCA7* is located in chromosome 19p13.3, and encodes a member of the superfamily of ATP-binding cassette transporters. *ABCA7* is expressed in microglia and oligodendrocytes [143] and potentially regulates lipid efflux and A β accumulation [144, 145]. Several SNPs in or close to *ABCA7* were identified as AD risk alleles [41, 42, 68]. However, the impact of these SNPs is not yet well understood.

Hypothesis 2: identified IGAP SNPs are regulatory SNPs

In *CELF1*, we did not have sufficient evidence that the IGAP SNP rs10838725 and the coding SNP rs2293576 in strong LD with the IGAP SNP were associated with AD.

However, rs10838725 acted as *cis*-eQTL for *MYBPC3* expression in the blood which was associated with AD. *MYBPC3* is located on chromosome 11p11.2, and encodes cardiac myosin binding protein C expressed exclusively in heart muscle [146]. Huang et al. reported that the *MYBPC3* and *SPII* were associated with the allele of rs1057233 in *CELF1* ($r^2 = 0.17$ and $D' = 0.97$ with the IGAP SNP rs10838725 as shown in Table 1.2), and suggested that the association of *MYBPC3* expression came from leaky transcription driven by the adjacent *SPII* expression [72].

In addition to *CRI*, several IGAP SNPs were associated with gene expression in the brain. *DLGAP1* and *NETO1* expressions were regulated by the *DSG2* IGAP SNP and were highly correlated with each other in FCTX. Interestingly, these genes were significantly under-expressed in MTG and SFG brain region of AD cases. The *DSG2* IGAP SNP also regulated *KCNG2* expression which was under-expressed in MTG of AD cases as well, although it was not correlated with the *DLGAP1* and *NETO1* expressions. *DLGAP1* is located in chromosome 18p11.31 more than 25Mb away from *DSG2*, and encodes disks large-associated protein 1 (also known as guanylate kinase-associated protein (GKAP)). *NETO1* is located in chromosome 18q22.3 more than 40Mb away from *DSG2* and encodes neuropilin and tolloid like 1. Both *DLGAP1* and *NETO1* are mainly expressed in neuros (http://web.stanford.edu/group/barres_lab/brainseqMariko/brainseq2.html) [132], and may be involved in the N-methyl-D-aspartate receptor-dependent synaptic plasticity [147, 148]. *KCNG2* located in chromosome 18q22.3, encodes a voltage-gated potassium channel subfamily G member 2, which is a potassium channel subunit. Potassium

channels are important regulatory proteins also associated with synaptic plasticity [149]. Synaptic plasticity is a fundamental property of the nervous system [150, 151]. Elevated A β levels induce synaptic dysfunction, and thus loss of synaptic proteins may contribute to AD progression. Although the role of *DSG2* gene is unknown, this may imply that *DSG2* is involved in brain functions including memory and learning.

There are limitations to this study. We aggregated data from many rich resources that aid in establishing a confluence of related information; however, these datasets are heterogeneous and can exhibit biases from the respective study designs, analytic protocols, and participant pools. As per common but inexact convention, we identified genes as those closest to the identified IGAP SNP. Although our approach for identifying proxies and examining eQTL lessens the impact of this crude gene assignment, this still remains an open question in the field.

In summary, investigating the functional role of the suspected and replicated SNPs associated with AD is an important next step to understanding the genetic contributions and the functional pathways linking AD developmental mechanisms. AD is a complex disease with a strong genetic component. However, much of the genetic contribution to AD remains unexplained. In future studies, we will need to investigate how RNA and protein levels as well as their interactions are affected the known genes.

Funding

This work was supported by the National Cell Repository for Alzheimer's Disease (U24 AG21886), and National Institute on Aging (K25 AG043546, UL1TR000117, the UK-ADC P30 AG028383, and R01 AG057187).

Table 3.1. ADGC and CHARGE studies in ADSP

Consortium	Study
ADGC	
	ACT
	ADC
	CHAP
	EFIGA
	GDF
	MAP
	MAYO
	MAYO PD
	MIA
	MIRAGE
	NCRAD
	NIA-LOAD
	RAS
	ROS
	TARCC
	TOR
	VAN
	WHICAP
CHARGE	
	ARIC
	ASPS
	CHS
	ERF
	FHS
	RS

ADGC = Alzheimer's Disease Genetic Consortium; CHARGE = Cohorts for Heart and Aging Research in Genomic Epidemiology; ADSP = Alzheimer's Disease Sequencing Project; ACT = Adult Changes in Thought; ADC = NIA Alzheimer Disease Centers; CHAP = Chicago Health and Aging Project; EFIGA = Estudio Familiar de la Influenza Genetica en Alzheimer; GDF = Genetic Differences; NIA-LOAD = National Institute on Aging (NIA) Late Onset Alzheimer's Disease Family Study; MAP = Memory and Aging Project; MAYO = Mayo Clinic; MAYO PD = Mayo PD; MIA = University of Miami; MIRAGE = Multi-Institutional Research in Alzheimer's Genetic Epidemiology; NCRAD = National Cell Repository for Alzheimer's Disease; RAS = University of Washington Families; ROS = Religious Orders Study; TARCC = Texas Alzheimer's Research and Care Consortium; TOR = University of Toronto; VAN = Vanderbilt University; WHICAP = Washington Heights-Inwood Columbia Aging Project; ARIC = Atherosclerosis Risk in Communities Study; ASPS = Austrian Stroke Prevention Study; CHS = Cardiovascular Health Study; ERF = Erasmus Rucphen Family; FHS = Framingham Heart Study; RS = Rotterdam Study

Table 3.2. Characteristics of the individual in ADSP

Variable	Overall n (%)	AD cases n (%)	AD controls n (%)
ADSP (n = 10,468)		5,374 (51.3)	5,094 (48.7)
Sex			
Male	4,380 (41.8)	2,313 (43.0)	2,067 (40.6)
Female	6,088 (58.2)	3,061 (57.0)	3,027 (59.4)
Age at the last visit or at death			
65-69	570 (5.5)	537 (10.0)	33 (0.6)
70-74	1,127 (10.8)	1,062 (19.8)	65 (1.3)
75-79	1,355 (12.9)	1,153 (21.5)	202 (4.0)
80-84	2,650 (25.3)	1,092 (20.3)	1,558 (30.6)
84-90	3,139 (30.0)	896 (16.7)	2,243 (44.0)
90+	1,627 (15.5)	634 (11.8)	993 (19.5)
<i>APOE</i>			
-/-	7,476 (71.4)	3,140 (58.4)	4,336 (85.1)
-/ ϵ 4	2,900 (27.7)	2,159 (40.2)	741 (14.5)
ϵ 4/ ϵ 4	91 (0.9)	75 (1.4)	17 (0.3)

ADSP = Alzheimer's Disease Sequencing Project; *APOE* = apolipoprotein E

Table 3.3. Exonic single nucleotide polymorphism correlated with the IGAP SNP

IGAP SNP	Closest Gene	Exonic SNP					1000 Genomes EUR		
		SNP ID	Position	Variant ^a	Alleles ^b	Gene	MAF	r ²	D'
Strong LD (r ² ≥ 0.8)									
rs6656401	<i>CRI</i>	rs4844600	207,679,307	E60E	G/A	<i>CRI</i>	0.19	0.88	0.99
		rs2296160	207,795,320	A2419T	G/A	<i>CRI</i>	0.18	0.84	0.93
rs9271192	<i>HLA-DRB5</i>	rs9270303	32,557,483	A13T	C/T	<i>HLA-DRB1</i>	0.25	0.92	0.99
rs1476679	<i>ZCWPW1</i>	rs2405442	99,971,313	L12L	C/T	<i>PILRA</i>	0.32	0.85	0.97
		rs1859788	99,971,834	G78R	G/A	<i>PILRA</i>	0.32	0.85	0.97
rs9331896	<i>CLU</i>	rs7982	27,462,481	H263H	G/A	<i>CLU</i>	0.39	0.90	0.97
rs10838725	<i>CELF1</i>	rs2293576	47,434,986	A191A	G/A	<i>SLC39A13</i>	0.31	0.84	0.99
rs983392	<i>MS4A6A</i>	rs12453	59,945,745	L137L	T/C	<i>MS4A6A</i>	0.40	0.80	0.91
rs4147929	<i>ABCA7</i>	rs3752246	1,056,492	A1527G	C/G	<i>ABCA7</i>	0.19	0.97	1
rs3865444	<i>CD33</i>	rs12459419	51,728,477	A14V	C/T	<i>CD33</i>	0.31	1	1

^a The first amino acid is linked to major allele and the second one to minor allele. ^b Major/minor alleles

IGAP = International Genomics of Alzheimer's Project; SNP = single nucleotide polymorphism; MAF = minor allele frequency; LD= linkage disequilibrium; 1000 Genomes = 1000 Genomes Project Phase 3 in individuals of European ancestry

Table 3.3. (Continued)

IGAP SNP	Closest Gene	Exonic SNP					1000 Genomes EUR		
		SNP ID	Position	Variant ^a	Alleles ^b	Gene	MAF	r ²	D'
Moderate LD ($0.4 \leq r^2 < 0.8$)									
rs9271192	<i>HLA-DRB5</i>	rs2308759	32,549,596	V130V	C/T	<i>HLA-DRB1</i>	0.14	0.45	1
		rs1049092	32,629,802	D201D	G/A	<i>HLA-DQB1</i>	0.40	0.51	0.97
		rs1049086	32,629,904	D167D	G/A	<i>HLA-DQB1</i>	0.40	0.50	0.97
rs2718058	<i>NME8</i>	rs2722372	37,890,267	R43K	G/A	<i>NME8</i>	0.25	0.49	0.93
		rs2598044	37,890,316	D59D	C/T	<i>NME8</i>	0.25	0.49	0.93
rs1476679	<i>ZCWPW1</i>	rs909152	100,175,473	G337G	C/T	<i>LRCH4</i>	0.31	0.53	0.75
rs10838725	<i>CELF1</i>	rs12286721	47,701,528	I671M	A/C	<i>AGBL2</i>	0.28	0.48	0.97
rs983392	<i>MS4A6A</i>	rs7232	59,940,599	T213S	T/A	<i>MS4A6A</i>	0.36	0.68	0.92
rs8093731	<i>DSG2</i>	rs16961975	29,046,606	V509M	G/A	<i>DSG3</i>	0.012	0.51	0.75
		rs61730311	29,049,138	R575W	C/T	<i>DSG3</i>	0.010	0.67	0.90
rs4147929	<i>ABCA7</i>	rs4147930	1,064,193	L1995L	A/G	<i>ABCA7</i>	0.28	0.56	1
		rs4147934	1,065,018	S2045A	T/G	<i>ABCA7</i>	0.28	0.55	0.98
		rs2074442	1,074,000	D275E	T/A	<i>HMHA1</i>	0.27	0.56	0.97
		rs2074454	1,080,311	P603P	C/G	<i>HMHA1</i>	0.26	0.49	0.88
		rs10404947	1,081,617	Q769Q	G/A	<i>HMHA1</i>	0.22	0.44	0.74
rs3865444	<i>CD33</i>	rs35112940	51,738,917	G304R	G/A	<i>CD33</i>	0.21	0.57	0.98

^a The first amino acid is linked to major allele and the second one to minor allele.

IGAP = International Genomics of Alzheimer's Project; SNP = single nucleotide polymorphism; MAF = minor allele frequency; LD = linkage disequilibrium; 1000 Genomes = 1000 Genomes Project Phase 3 in individuals of European ancestry

Table 3.4. Association results for exonic SNPs correlated with the IGAP SNPs

IGAP SNP	Closest gene	Exonic SNP			
		SNP ID	Gene	OR	P-value
Strong LD ($r^2 \geq 0.8$)					
rs6656401	<i>CR1</i>	rs4844600	<i>CR1</i>	-	-
		rs2296160	<i>CR1</i>	1.11	7.47×10^{-3}
rs9271192	<i>HLA-DRB5</i>	rs9270303	<i>HLA-DRB1</i>	1.16	1.68×10^{-4}
rs1476679	<i>ZCWPW1</i>	rs2405442	<i>PILRA</i>	0.89	1.22×10^{-3}
		rs1859788	<i>PILRA</i>	0.89	1.22×10^{-3}
rs9331896	<i>CLU</i>	rs7982	<i>CLU</i>	0.90	1.30×10^{-3}
rs10838725	<i>CELF1</i>	rs2293576	<i>SLC39A13</i>	1.08	0.030
rs983392	<i>MS4A6A</i>	rs12453	<i>MS4A6A</i>	0.89	3.56×10^{-4}
rs4147929	<i>ABCA7</i>	rs3752246	<i>ABCA7</i>	1.18	1.30×10^{-4}
rs3865444	<i>CD33</i>	rs12459419	<i>CD33</i>	0.95	0.11
Moderate LD ($0.4 \leq r^2 < 0.8$)					
rs9271192	<i>HLA-DRB5</i>	rs2308759	<i>HLA-DRB1</i>	1.09	0.062
		rs1049092	<i>HLA-DQB1</i>	1.12	7.29×10^{-4}
		rs1049086	<i>HLA-DQB1</i>	1.11	1.37×10^{-3}
rs2718058	<i>NME8</i>	rs2722372	<i>NME8</i>	0.90	3.19×10^{-3}
		rs2598044	<i>NME8</i>	0.90	3.61×10^{-3}
rs1476679	<i>ZCWPW1</i>	rs909152	<i>LRCH4</i>	0.97	0.33
rs10838725	<i>CELF1</i>	rs12286721	<i>AGBL2</i>	1.05	0.13
rs983392	<i>MS4A6A</i>	rs7232	<i>MS4A6A</i>	0.87	3.53×10^{-5}
rs8093731	<i>DSG2</i>	rs16961975	<i>DSG3</i>	1.21	0.23
		rs61730311	<i>DSG3</i>	1.15	0.45
rs4147929	<i>ABCA7</i>	rs4147930	<i>ABCA7</i>	1.10	9.92×10^{-3}
		rs4147934	<i>ABCA7</i>	1.10	6.04×10^{-3}
		rs2074442	<i>HMHA1</i>	1.08	0.045
		rs2074454	<i>HMHA1</i>	1.09	0.019
		rs10404947	<i>HMHA1</i>	1.07	0.086
rs3865444	<i>CD33</i>	rs35112940	<i>CD33</i>	0.98	0.67

A bold p-value represents the statistical significance after FDR adjustment.

IGAP = International Genomics of Alzheimer's Project; SNP = single nucleotide polymorphism; ADSP = Alzheimer's Disease Sequencing Project; OR = odds ratio; LD = linkage disequilibrium

Table 3.5. Pathogenic nature of nonsynonymous single nucleotide polymorphism associated with AD

Exonic SNP	Gene	Major/minor allele	Variant ^a	SIFT	PolyPhen-2
Nonsynonymous					
rs2296160	<i>CRI</i>	G/A	A2419T	Tolerated (0.39)	Benign (0.0)
rs9270303	<i>HLA-DRB1</i>	C/T	A13T	Tolerated (1.0)	Benign (0.0)
rs1859788	<i>PILRA</i>	G/A	G78R	Tolerated (1.0)	Benign (0.0)
rs3752246	<i>ABCA7</i>	C/G	A1527G	Tolerated (0.88)	Benign (0.0)
rs7232	<i>MS4A6A</i>	T/A	T213S	Deleterious (0.03)	Possibly damaging (0.827)

^aThe first amino acid is linked to major allele and the second one to minor allele, and the codon number is for the biggest isoform.
 SNP = single nucleotide polymorphism

Table 3.6. Significant *cis*- and *trans*-association of the IGAP SNPs with blood gene expressions in ADNI

IGAP SNP	Closest gene	Probe set ID ^a	Gene	<i>Cis</i> or <i>trans</i>	eQTL association		AD association
					β	P-value ^b	P-value ^c
rs1476679	<i>ZCWPW1</i>	11736388_a_at	<i>TRIM4</i>	<i>Cis</i>	-0.126	1.16×10^{-8}	0.78
		11760665_at	<i>ZKSCAN1</i>	<i>Cis</i>	0.150	1.96×10^{-7}	0.18
		11722909_a_at	<i>GATS</i>	<i>Cis</i>	0.177	4.14×10^{-17}	0.86
		11730247_a_at	<i>PVRIG</i>	<i>Cis</i>	0.088	8.39×10^{-5}	0.36
		11743311_a_at	<i>PILRB</i>	<i>Cis</i>	-0.113	1.20×10^{-8}	0.28
		11730023_s_at	<i>PILRB</i>	<i>Cis</i>	-0.106	3.44×10^{-8}	0.53
		11730022_a_at	<i>PILRB</i>	<i>Cis</i>	-0.129	5.41×10^{-8}	0.68
rs11771145	<i>EPHA1</i>	11755327_s_at	<i>LOC154761</i>	<i>Cis</i>	0.110	5.11×10^{-6}	0.59
rs28834970	<i>PTK2B</i>	11761824_at	<i>NSAP11</i>	<i>Trans</i>	0.028	8.78×10^{-5}	1.00
		11723344_at	<i>TRIM35</i>	<i>Cis</i>	-0.067	1.31×10^{-5}	0.29
		11720981_a_at	<i>PTK2B</i>	<i>Cis</i>	0.109	3.03×10^{-15}	0.13
		11720982_s_at	<i>PTK2B</i>	<i>Cis</i>	0.079	1.72×10^{-14}	0.077
		11720980_a_at	<i>PTK2B</i>	<i>Cis</i>	0.089	3.04×10^{-10}	0.12
rs10838725	<i>CELF1</i>	11725151_at	<i>MYBPC3</i>	<i>Cis</i>	0.145	1.82×10^{-7}	3.80×10^{-4}
rs983392	<i>MS4A6A</i>	11716846_a_at	<i>MS4A6A</i>	<i>Cis</i>	-0.087	1.27×10^{-12}	0.045
		11732865_a_at	<i>MS4A4A</i>	<i>Cis</i>	-0.178	2.69×10^{-6}	0.70
		11751570_a_at	<i>MS4A4A</i>	<i>Cis</i>	-0.148	3.02×10^{-6}	0.97
rs8093731	<i>DSG2</i>	11735070_a_at	<i>GNAL</i>	<i>Trans</i>	-0.317	9.65×10^{-6}	0.35
rs7274581	<i>CASS4</i>	11720252_s_at	<i>C20orf194</i>	<i>Trans</i>	-0.210	7.33×10^{-5}	0.74
		11723408_a_at	<i>MKKS</i>	<i>Trans</i>	-0.130	5.03×10^{-5}	0.85

A bold p-value represents the statistical significance after FDR adjustment.

^a Probe set IDs on Affymetrix Human Genome U219 Array

^b P-values less than false discovery rate (FDR) adjusted significance level

^c P-values calculated by analysis of covariance with the outcome of gene expression and the predictor of AD status (normal/MCI/AD)

IGAP = International Genomics of Alzheimer's Project; SNP = single nucleotide polymorphism; eQTL = expression quantitative trait locus; ADNI = Alzheimer's Disease Neuroimaging Initiative

Table 3.7. Significant *cis*- and *trans*-association of the IGAP SNPs with brain gene expressions in NABEC and UKBEC

SNP ID	Closest gene	Probe set ID ^a	Gene expression	<i>Cis</i> or <i>trans</i>	Brain region	β	P-value ^b
NABEC							
rs10792832	<i>PICALM</i>	ILMN_1714980	<i>MRGPRD</i>	<i>Trans</i>	CRBLM	0.066	1.62×10^{-5}
rs8093731	<i>DSG2</i>	ILMN_2380779	<i>DLGAP1</i>	<i>Trans</i>	FCTX	0.848	4.87×10^{-10}
		ILMN_1783168	<i>NETO1</i>	<i>Trans</i>	FCTX	0.757	1.66×10^{-6}
		ILMN_1780373	<i>KCNG2</i>	<i>Trans</i>	CRBLM	0.408	6.76×10^{-5}
rs7274581	<i>CASS4</i>	ILMN_1731948	<i>PCK1</i>	<i>Trans</i>	CRBLM	0.241	7.29×10^{-6}
UKBEC							
rs6656401	<i>CRI</i>	t2408244	<i>COL9A2</i>	<i>Trans</i>	MEDU	0.152	3.50×10^{-6}
		t2434716	<i>CERS2</i>	<i>Trans</i>	MEDU	0.284	1.87×10^{-6}
		t2437801	<i>ARHGEF2</i>	<i>Trans</i>	MEDU	0.183	1.15×10^{-5}
		t2376299	<i>CNTN2</i>	<i>Trans</i>	MEDU	0.297	5.59×10^{-6}
		t2376457	<i>CDK18</i>	<i>Trans</i>	MEDU	0.323	9.83×10^{-6}
		t2377332	<i>CRI, CRIL</i>	<i>Cis</i>	AveALL	0.078	4.22×10^{-6}
					WHMT	0.172	6.55×10^{-7}

^a Probe set IDs on HumanHT-12_v3 Expression BeadChips in NABEC, and on Affymetrix Exon 1.0 ST Arrays in UKBEC

^b P-values less than false discovery rate (FDR) adjusted significance level

IGAP = International Genomics of Alzheimer's Project; SNP = single nucleotide polymorphism; NABEC = North American Brain Expression Consortium; UKBEC = United Kingdom Brain Expression Consortium; FCTX = frontal cortex; CRBLM = cerebellum; MEDU = medulla; WHMT = white matter; AveALL = average of all 10 regions

Table 3.8. Associations between gene expressions identified in NABEC and UKBEC and AD status in GSE5281 dataset, entorhinal cortex (EC), hippocampus (HIP), and medial temporal gyrus (MTG)

Probe set ID ^a	EC		HIP		MTG	
	β	P-value	β	P-value	β	P-value
Identified in NABEC ^b						
<i>DLGAP1</i>						
206489_s_at	-0.979	0.058	-1.511	0.010	-1.383	0.03
206490_at	-0.320	0.31	-0.350	0.084	-1.266	1.04 × 10⁻³
210750_s_at	0.876	0.24	0.915	0.15	-0.348	0.56
235527_at	-0.761	0.15	0.875	4.59 × 10⁻⁵	1.363	0.033
<i>NETO1</i>						
1552736_a_at	-0.787	0.10	0.009	0.97	-1.140	0.058
1552904_at	-1.728	8.03 × 10⁻⁴	0.023	0.95	-1.400	0.027
1562713_a_at	-1.742	0.024	-0.614	0.41	-1.993	6.86 × 10⁻⁴
236440_at	-0.580	0.22	-0.744	0.064	-1.988	7.41 × 10⁻⁵
<i>KCNG2</i>						
208550_x_at	-0.413	0.26	0.431	0.21	-1.933	6.97 × 10⁻⁴
<i>PCK1</i>						
208383_s_at	0.948	0.20	0.787	0.26	1.306	0.030

A bold p-value represents the statistical significance after FDR adjustment.

^aProbe set IDs on Affymetrix U133 Plus 2.0 array

^bThere is no *MRGPRD* expression data available.

NABEC = North American Brain Expression Consortium; UKBEC = United Kingdom Brain Expression Consortium; EC = entorhinal cortex; HIP = hippocampus (HIP); MTG = medial temporal gyrus

Table 3.8. (Continued)

Probe set ID ^a	EC		HIP		MTG	
	β	P-value	β	P-value	β	P-value
Identified in UKBEC						
<i>COL9A2</i>						
213622_at	-0.384	0.32	0.502	0.37	1.031	0.021
232542_at	1.696	7.83×10^{-3}	0.148	0.76	-0.577	0.32
<i>CERS2</i>						
222212_s_at	0.744	3.63×10^{-3}	-0.526	0.11	0.480	0.021
<i>ARHGEF2</i>						
1554783_s_at	-1.296	0.092	0.137	0.84	-0.942	0.17
207629_s_at	-0.319	0.52	-0.390	0.33	-0.373	0.33
209435_s_at	-0.920	1.79×10^{-6}	0.246	0.16	0.058	0.77
235595_at	1.033	0.041	0.963	0.058	0.590	0.18
<i>CNTN2</i>						
206970_at	-0.993	0.11	-1.725	0.020	-0.460	0.46
230045_at	0.751	0.023	-0.550	0.20	-0.146	0.58
<i>CDK18</i>						
214797_s_at	-0.352	0.62	0.032	0.94	-0.152	0.78
<i>CRI</i>						
206244_at	-0.783	0.24	-1.259	0.013	-2.028	2.29×10^{-3}
208488_s_at	0.732	0.33	0.151	0.80	-0.231	0.69
217484_at	0.840	0.069	0.457	0.35	0.717	0.15
244313_at	1.501	3.81×10^{-3}	-0.099	0.86	-0.098	0.87

A bold p-value represents the statistical significance after FDR adjustment.

^aProbe set IDs on Affymetrix U133 Plus 2.0 array

NABEC = North American Brain Expression Consortium; UKBEC = United Kingdom Brain Expression Consortium; EC = entorhinal cortex; HIP = hippocampus (HIP); MTG = medial temporal gyrus

Table 3.9. Associations between gene expressions identified in NABEC and UKBEC and AD status in GSE5281 dataset, posterior cingulate (PC), superior frontal gyrus (SFG), and primary visual cortex (VCX)

Probe set ID ^a	PC		SFG		VCX	
	β	P-value	β	P-value	β	P-value
Identified in NABEC ^b						
<i>DLGAP1</i>						
206489_s_at	-1.449	0.012	-1.801	3.57×10^{-4}	-0.647	0.20
206490_at	-0.183	0.51	-1.229	3.20×10^{-3}	-0.995	0.011
210750_s_at	1.027	0.094	-0.352	0.60	0.755	0.20
235527_at	-0.722	0.040	-0.671	0.16	-0.937	0.024
<i>NETO1</i>						
1552736_a_at	-0.098	0.74	-1.591	6.53×10^{-4}	-0.556	0.29
1552904_at	-1.210	0.068	-0.740	0.061	-0.158	0.81
1562713_a_at	-1.180	0.13	-2.066	1.10×10^{-3}	-0.557	0.16
236440_at	-1.029	5.19×10^{-3}	-1.079	0.049	-0.404	0.29
<i>KCNG2</i>						
208550_x_at	0.729	0.066	-1.089	0.024	0.965	0.017
<i>PCK1</i>						
208383_s_at	0.906	0.28	-0.016	0.98	-0.470	0.48

A bold p-value represents the statistical significance after FDR adjustment.

^aProbe set IDs on Affymetrix U133 Plus 2.0 array

^bThere is no *MRGPRD* expression data available.

NABEC = North American Brain Expression Consortium; UKBEC = United Kingdom Brain Expression Consortium; PC = posterior cingulate; SFG = superior frontal gyrus; VCX = primary visual cortex

Table 3.9. (Continued)

Probe set ID ^a	PC		SFG		VCX	
	β	P-value	β	P-value	β	P-value
Identified in UKBEC						
<i>COL9A2</i>						
213622_at	-0.029	0.97	-0.142	0.73	0.734	0.034
232542_at	0.948	0.16	0.121	0.84	-0.537	0.22
<i>CERS2</i>						
222212_s_at	-0.316	0.32	-0.275	0.33	0.567	0.031
<i>ARHGEF2</i>						
1554783_s_at	0.910	0.21	-0.759	0.15	0.924	0.15
207629_s_at	0.185	0.69	-0.761	0.041	0.280	0.33
209435_s_at	0.461	3.20 × 10⁻⁴	-0.671	0.019	0.253	0.13
235595_at	2.245	1.02 × 10⁻³	0.172	0.73	0.400	0.33
<i>CNTN2</i>						
206970_at	-1.270	0.070	-1.402	0.011	0.284	0.56
230045_at	-0.104	0.72	-0.135	0.74	0.006	0.98
<i>CDK18</i>						
214797_s_at	0.714	0.18	-0.899	0.12	0.796	0.18
<i>CRI</i>						
206244_at	1.336	0.082	-1.075	0.10	1.282	0.049
208488_s_at	0.332	0.52	-1.239	0.015	0.655	0.20
217484_at	0.688	0.24	-0.344	0.45	1.139	4.21 × 10 ⁻³
244313_at	0.780	0.14	0.299	0.61	1.248	0.01

A bold p-value represents the statistical significance after FDR adjustment.

^aProbe set IDs on Affymetrix U133 Plus 2.0 array

NABEC = North American Brain Expression Consortium; UKBEC = United Kingdom Brain Expression Consortium; PC = posterior cingulate; SFG = superior frontal gyrus; VCX = primary visual cortex

Table 3.10. Associations between gene expressions identified in NABEC and UKBEC and AD status in Allen Institute dataset

Gene expression ^a	FWM		HIP		PCx		TCx	
	β	P-value	β	P-value	β	P-value	β	P-value
Identified in NABEC ^b								
<i>DLGAP1</i>	0.749	0.028	-0.109	0.056	-0.034	0.51	-0.036	0.38
<i>NETO1</i>	0.495	0.095	0.010	0.91	-0.033	0.55	-0.041	0.36
<i>KCNG2</i>	0.085	0.43	0.004	0.97	0.035	0.73	0.203	0.020
<i>PCK1</i>	0.380	0.25	0.248	0.36	0.533	0.04	0.368	0.096
Identified in UKBEC								
<i>COL9A2</i>	-0.262	0.12	0.016	0.88	0.062	0.62	0.232	0.041
<i>CERS2</i>	-0.087	0.55	0.094	0.11	-0.052	0.55	0.083	0.19
<i>ARHGEF2</i>	-0.088	0.27	-0.028	0.51	-0.021	0.59	0.112	1.98 × 10⁻³
<i>CNTN2</i>	-0.143	0.41	0.034	0.70	-0.086	0.42	0.219	0.011
<i>CDK18</i>	-0.218	0.22	0.063	0.53	0.059	0.59	0.263	2.41 × 10⁻³
<i>CRI</i>	0.151	0.46	0.128	0.47	0.005	0.97	0.045	0.78

A bold p-value represents the statistical significance after FDR adjustment.

^a RNA sequencing on Illumina HighSeq 2500

^b *MRGPRD* expression was removed because of little expressions.

NABEC = North American Brain Expression Consortium; UKBEC = United Kingdom Brain Expression Consortium; FWM = white matter; HIP = hippocampus; PCx = parietal cortex; TCx = temporal cortex

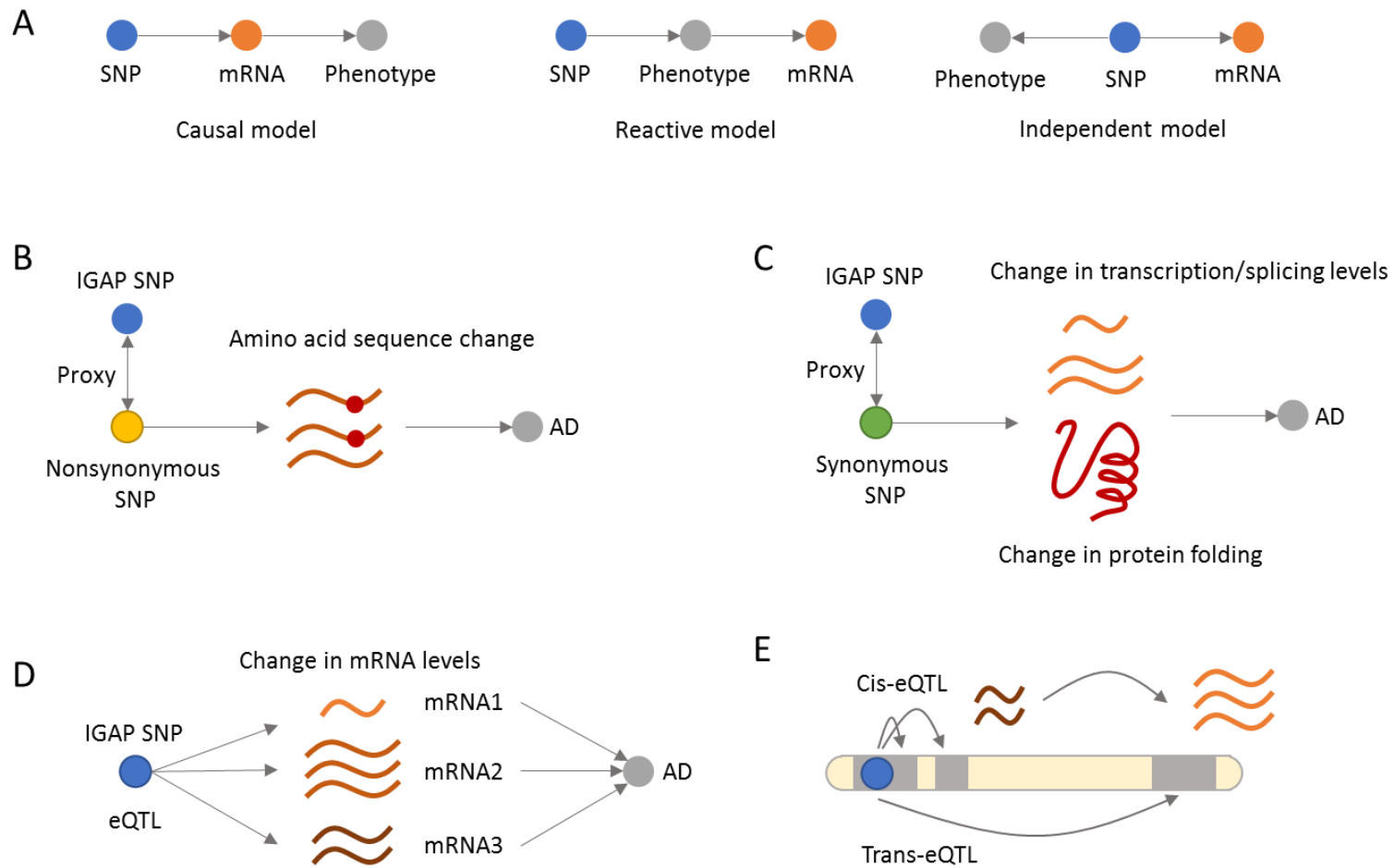


Figure 3.1. Possible causal relationships between single nucleotide polymorphisms (SNPs), mRNA, and phenotype. IGAP = International Genomics of Alzheimer's Project; eQTL = expression quantitative trait locus

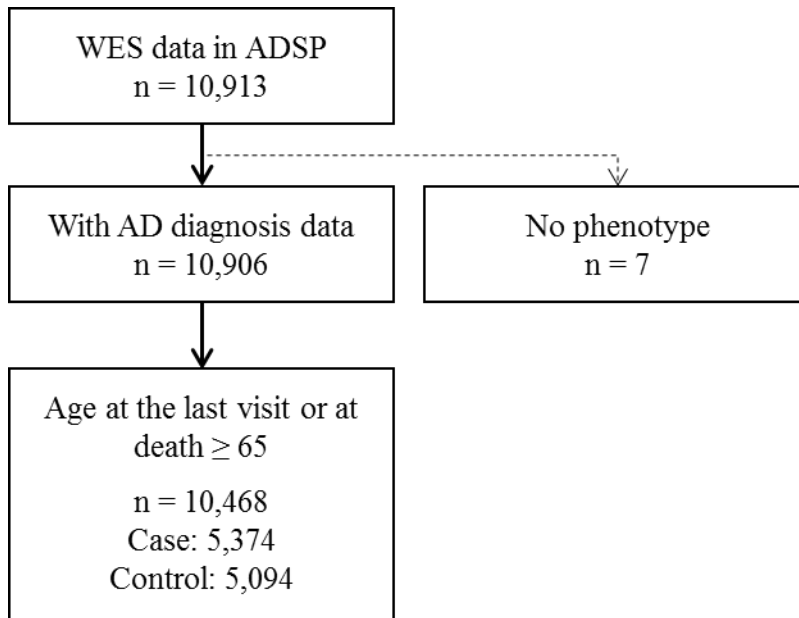


Figure 3.2. Flow diagram of the subjects included in the analyses.
ADSP = Alzheimer's Disease Sequencing Project

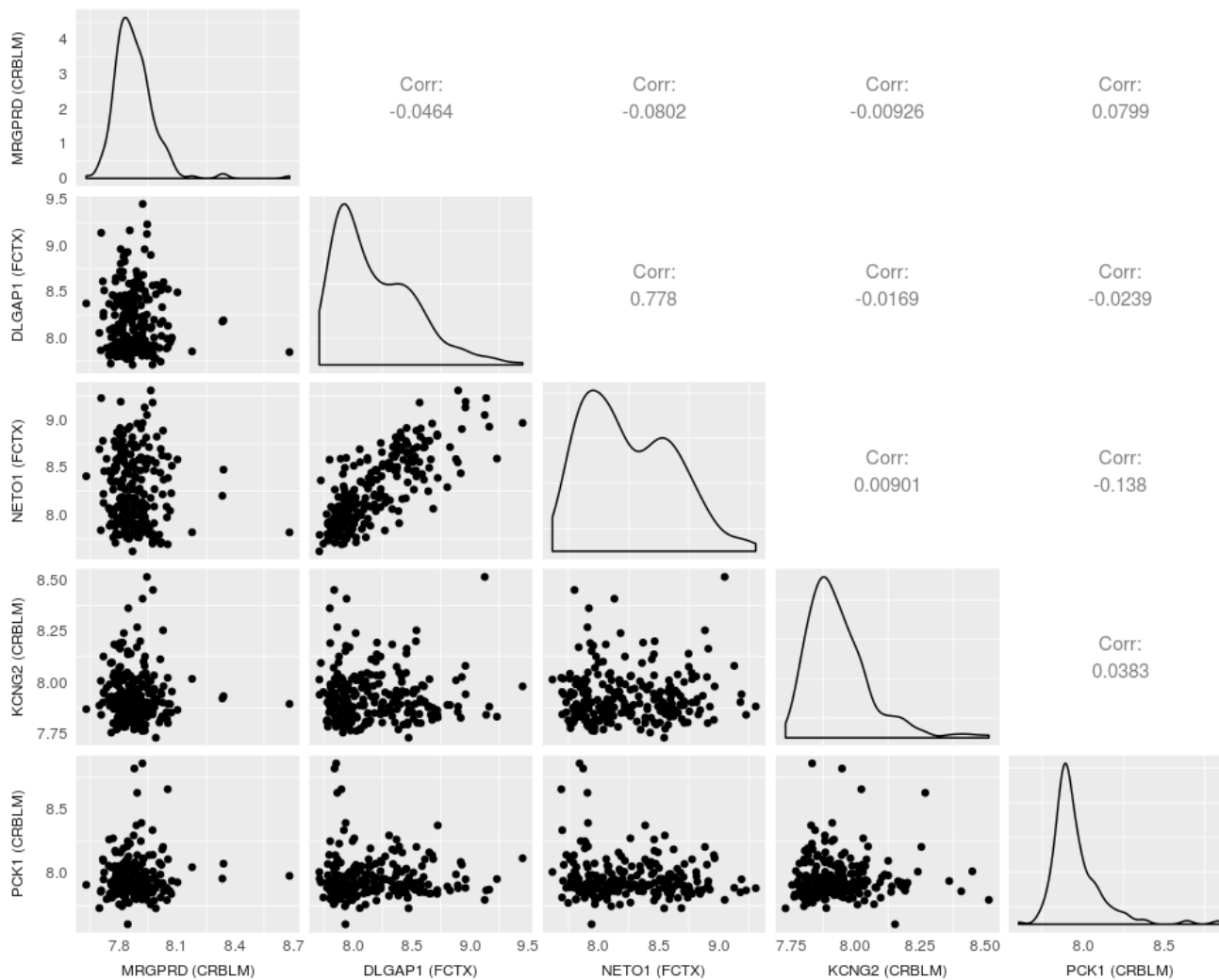


Figure 3.3. Correlation between gene expressions potentially regulated by IGAP SNPs in NABEC

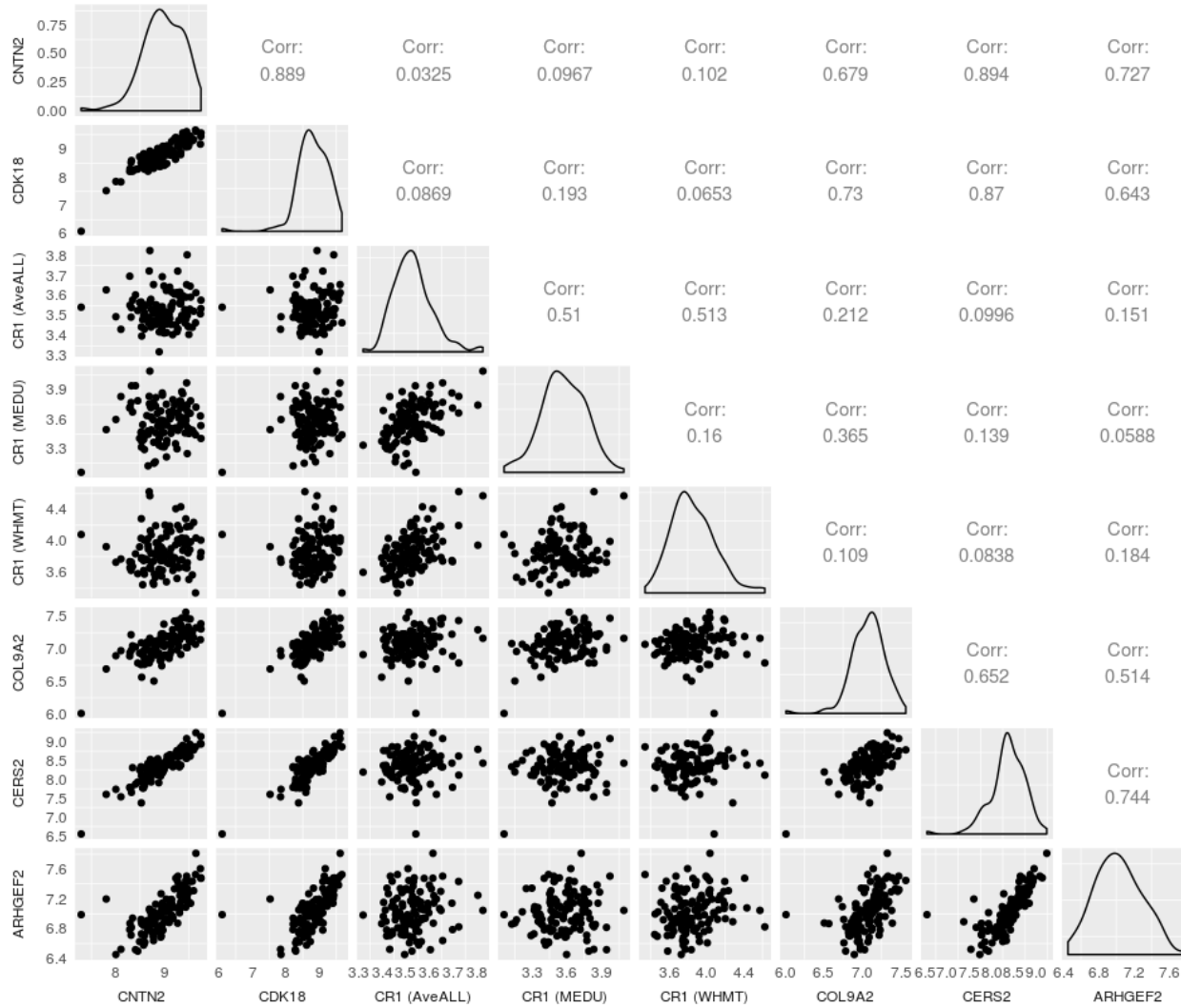


Figure 3.4. Correlation between gene expressions potentially regulated by *CR1* SNPs in UKBEC

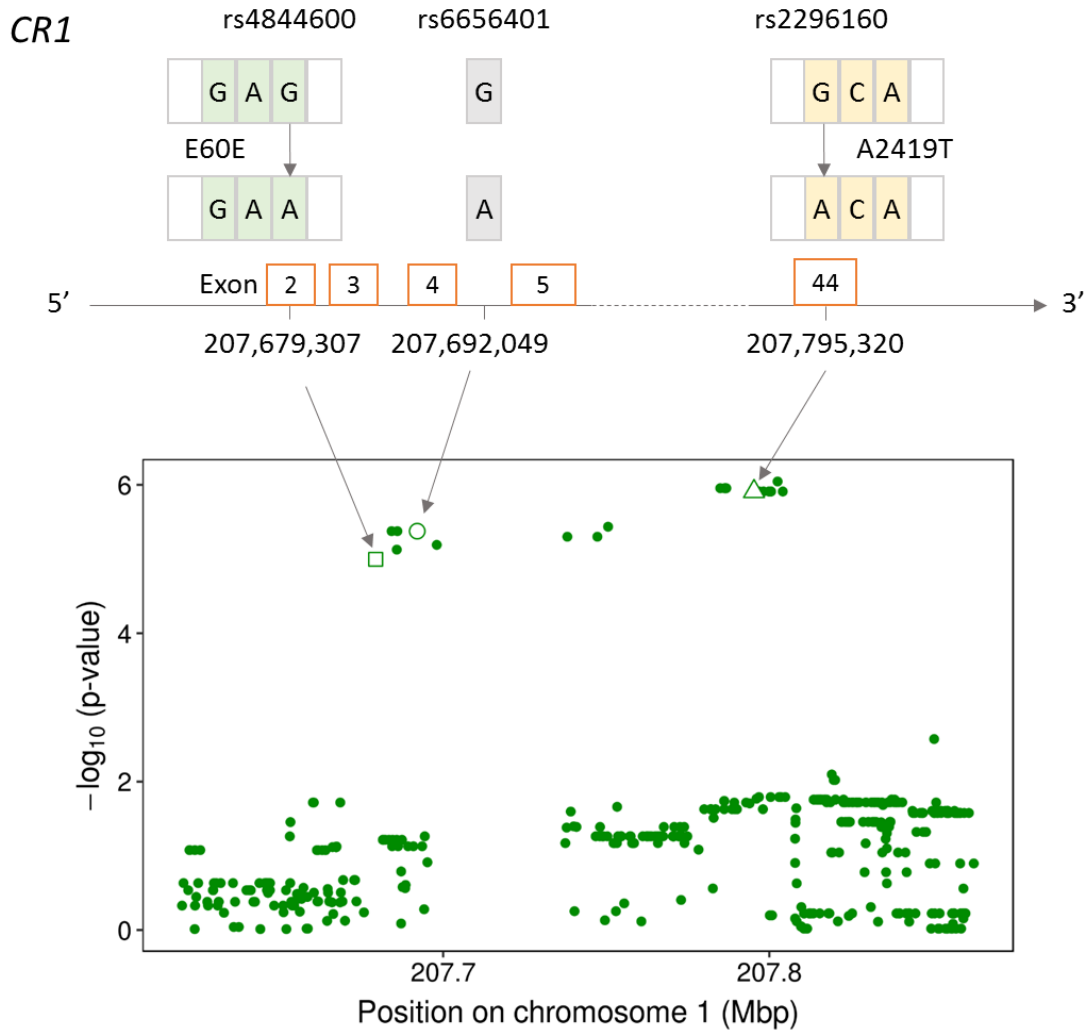


Figure 3.5. Plot for the associations of the *CR1* SNPs with the *CR1* expression in the average of 10 brain regions of UKBEC. The outlined square, circle, and triangle indicate the synonymous SNP rs4844600, the IGAP SNP rs6656401, and the nonsynonymous SNP rs2296160, respectively.

SNP = single nucleotide polymorphism; ADSP = Alzheimer's Disease Sequencing Project; UKBEC = United Kingdom Brain Expression Consortium

CHAPTER FOUR

Identifying regions harboring Alzheimer's disease related rare variants using scan-statistic-based analysis of exome sequencing data

Abstract

Recent advances in sequencing technologies have allowed the move toward comprehensive genome-wide approaches, enabling more accurate genotyping of rare variants. To date, several studies have succeeded in identifying rare variants associated with Alzheimer's disease (AD), yielding protective or risk effects. We applied a scan-statistic-based approach and developed an approach to construct optimized windows within a gene to find meaningful clusters harboring risk or protective rare variants for AD. Data in this study came from the Alzheimer's Disease Sequencing Project (ADSP) comprising 18 studies from the Alzheimer's Disease Genetic Consortium (ADGC), and 6 studies from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium. Of the 10,031 subjects included in this study, 5,142 (51.3%) were diagnosed as AD. We evaluated the scan statistics with different settings in *TREM2* and *TOMM40* as highly-replicated positive controls. We obtained very similar scan statistic values when we specified the whole genome and chromosome as a large genetic region. Our optimized window approach captured almost the entire gene in *TREM2* and the single variant in *TOMM40* as a meaningful cluster. Applying the optimized window approach in all genes, we detected clusters harboring risk or protective variants for AD including *MUC6*, *NXNLI*, and *BCAM*. As more NGS data become available, it is of great

interest to see whether these findings are replicated in other cohorts of European ancestry as well as different populations.

Introduction

Over the past decade, genome-wide association studies (GWAS) have identified common risk loci for Alzheimer's disease (AD). It is well known that the $\epsilon 4$ allele of apolipoprotein E (*APOE*) is the major genetic risk factor for late onset of AD. In addition to *APOE*, GWAS have detected and replicated several single nucleotide polymorphisms (SNPs) that are associated with susceptibility of AD, including in or close to *CRI*, *BIN1*, *INPP5D*, *MEF2C*, *CD2AP*, *NME8*, *EPHA1*, *PTK2B*, *PICALM*, *SORL1*, *FERMT2*, *SLC24A4-RIN3*, *DSG2*, *CASS4*, *HLA-DRB5-DBR1*, *CLU*, *MS4A6A*, *ABCA7*, *CD33*, *ZCWPW1*, and *CELF1* [41-45].

Rare variants have become a focus in the recent past. Although GWAS have been successful in interrogating genetic variants for association with disease, GWAS are performed under the "common disease – common variant" hypothesis positing that common traits are caused by the combination of common variants with a small to moderate effect [58]. GWAS rely on genotyping preselected SNPs and imputing ungenotyped variants based on local linkage disequilibrium (LD) of a set of some haplotypes from reference population. Imputation approaches have continually improved and are quite accurate for common variants [59, 60] but are not as reliable for rare variants [61]. Therefore, imputed rare variants are typically removed from GWAS analysis. Although GWAS for common variants have revealed numerous susceptibility

variants for common diseases, much of the genetic contribution to common diseases remains unexplained – the so called ‘missing heritability’ problem [152, 153]. One possible explanation of this missing heritability is that rare variants may account for additional disease variability [153, 154].

Recent advances in sequencing technologies have allowed to move toward comprehensive genome-wide approaches, enabling to accurately genotype rare variants generally defined as a variant with minor allele frequency (MAF) < 1-5%. These next-generation sequencing (NGS) technologies have the potential to improve our understanding the role of both common and rare variants in the underlying biological mechanisms of developing a disease. Whole-exome sequencing (WES) and whole-genome sequencing (WGS) are ideal approaches to identify novel variants and genes associated with complex traits.

To date, several studies have reported rare variants associated with AD, yielding protective or risk effects. Jonsson et al. found that a rare missense variant in *TREM2* -- rs75932628 producing an amino acid substitution (arginine to histidine at codon 47) -- was associated with the risk of AD [155]. Concurrently, Guerreiro et al. confirmed that rs75932628 was the most associated variant with AD in *TREM2* [156]. This variant is considered the first NGS-based finding of a novel rare variant associated with AD risk. Since then, several rare variants have been reported as risk or protective factors for AD, including in or near *BIN1*, *UNC5C*, *TREM2*, *CD2AP*, *AKAP9*, *EPHA1*, *SORL1*, *TM2D3*, *PLCG2*, *ABI3*, *PLD3*, and *ABCA7* (Table 4.1).

Most coding variants, however, are very rare, and thus an extremely large sample size is required to identify a single variant associated with a disease. There are significant computational and statistical challenges for these sequencing studies. Traditional single-variant-based association tests are underpowered to detect rare variant associations unless sample size and/or effect size is very large [62]. Instead of testing single variant individually, more powerful and computationally efficient approaches for aggregating the effects of rare variants have become a standard approach for association testing. Many such approaches for testing association between rare variants within a pre-specified region and a disease have been proposed. A recently-proposed scan-statistic-based test can be used to detect the location of rare variant clusters influencing a disease. The scan-statistic-based test was introduced into human genetics by Hoh et al [65] to locate susceptibility genes. Ionita-Laza et al. adapted this test to identify clusters of rare disease risk variants based on a likelihood ratio under a Bernoulli model as proposed by Kulldorff [66, 67].

Variants within a functional protein-coding domain may be located in close proximity and may play a similar role in genetic mechanisms of a disease. Unlike association tests or other cluster detection analyses, the scan-statistic-based test [67] adopted in this work can both detect the location of clusters and examine the association against the null hypothesis that variants within a certain scan window are equally likely to confer AD risk compared to those outside the window. This approach is powerful when there are clusters of disease-related variants with the same direction of association within a selected

region/window [67]. In this study, we applied the scan-statistic-based approach and optimized windows within a gene to find meaningful clusters harboring risk or protective rare variants for AD.

Material and methods

Study subjects

Data in this study came from the Alzheimer's Disease Sequencing Project (ADSP) comprising 18 studies from the Alzheimer's Disease Genetic Consortium (ADGC), and 6 studies from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium. The ADSP consists of a smaller family-based study and a case-control study of unrelated subjects. In this study, we used the case-control study dataset containing whole exome sequencing (WES) data on 10,913 subjects (Table 4.2). Of the 10,913 subjects in WES data, 10,468 subjects who had the AD diagnosis available and who were 65 years or older at the last visit or at death were included in a principal component analysis (PCA) to identify ethnic outliers. PCA was performed in PLINK v1.90a [82] using a linkage disequilibrium (LD) pruned subset of markers (pairwise $r^2 < 0.2$) from these data and 1000 Genomes Project Phase 3 (1000 Genomes) [84] data after removing symmetric SNPs and flipping SNPs discordant for DNA strands between ADSP WES and 1000 Genomes data. We then plotted the first and second principal components (PCs) for each individual ($n = 10,468$ from ADSP WES and $n = 2,504$ from 1000 Genomes) using the ggplot2 R package (version 2.2.0) [85] in R (version 3.4.1; <http://www.r-project.org>). Based on the PC plot, 437 subjects were removed as ethnic

outliers (Figure 4.1 and Figure 4.2). We considered individuals who were marked as prevalent or incident AD at year 0 AD cases.

ADSP WES data

We obtained the public access WES data containing biallelic single nucleotide variants and insertion/deletion (indels) genotypes from ADSP in the unrelated sample set. The WES data was generated by applying a quality control protocol designed by the ADSP Quality Control Working Group, removing low-quality variants and samples (see the details in Supplementary Method 1).

Statistical analysis

The likelihood ratio (LR) statistic for the Bernoulli model from the scan-statistic-based test is defined as:

$$LR_W = \begin{cases} \left(\frac{\hat{p}_W}{\hat{r}_G}\right)^{y_W} \left(\frac{1-\hat{p}_W}{1-\hat{r}_G}\right)^{n_W-y_W} \left(\frac{\hat{q}_W}{\hat{r}_G}\right)^{y_G-y_W} \left(\frac{1-\hat{q}_W}{1-\hat{r}_G}\right)^{n_G-n_W-(y_G-y_W)} & \text{if } \hat{p}_W > \hat{q}_W \\ 1 & \text{otherwise} \end{cases}$$

where

$$y_W = \sum_{i \in W} y_i, \quad y_G = \sum_{i \in G} y_i, \quad n_W = \sum_{i \in W} n_i, \quad n_G = \sum_{i \in G} n_i$$

$$\hat{p}_W = \frac{y_W}{n_W}, \quad \hat{q}_W = \frac{y_G - y_W}{n_G - n_W}, \quad \hat{r}_G = \frac{y_G}{n_G}$$

W is a window with fixed size w defined by the number of base pairs, G is a large genetic region of interest, and y_i and n_i are the numbers of AD cases and total subjects carrying the i th rare variant, respectively [67]. The window W with the highest LR_W is the most likely region to harbor a cluster of disease-associated variants.

The definitions of the large genetic region G and the window W are flexible. The power of the test primarily depends on the total number of risk variants in the entire genetic region G , the expected number of risk variants in the cluster, and the ratio of the probabilities of being risk variants inside and outside of the cluster [157]. Other factor influencing the power is how well the scanning window size matches the true cluster size. The scan statistics are sensitive for the chosen scanning window. A cluster within a 10 kb sized region, for example, may not be detected by two consecutive scanning windows of 5 kb (i.e., the cluster is split by two windows). A sliding window approach (i.e., partially overlapping windows) may be one of the solutions for this problem. Another issue is that the power may decrease when the true cluster window size is larger relative to the entire region and when scanning window size is too large compared to the true cluster size [67]. Ideally, one would hope to define a scanning window so that it is matched to the true cluster harboring variants related to disease.

We first compared the likelihood ratio (LR) statistic between different settings for a large genetic region G , a fixed window size W , and a sliding window size using a sliding window approach. We used *TREM2* (located on chromosome 6p21.1) and *TOMM40* (located on chromosome 19q13.32) as highly-replicated positive controls to evaluate the scan statistics for risk and protective effects. Using Fisher's exact test as implemented in PLINK v1.90a [82], we preliminarily confirmed that there was a significant variant in each of *TREM2* and *TOMM40* based on the false discovery rate (FDR)-adjusted p-value (i.e., q-value) < 0.05 defined as a significance level (Table 4.3). The variant rs75932628

(R47H, position: 41,129,252) in *TREM2* was a significant risk rare variant for AD, showing a MAF of 0.94% in AD cases and 0.21% in controls. The odds ratio (OR) was 4.41 and p-value was 4.30×10^{-12} . On the other hand, the variant rs1160983 (S183S, position: 45,397,229) in *TOMM40* had a protective effect with a MAF of 1.80% in AD cases and 4.34% in controls (OR = 0.39 and p-value = 2.12×10^{-21}). *TOMM40* is close to the *APOE* locus and the variant was in LD with *APOE* $\epsilon 2$ known as a protective factor ($r^2 = 0.35$ and $D' = 0.95$).

We tried to find optimized windows in *TREM2* and *TOMM40* by maximizing LR_W under the difference genetic region G settings. In addition to these positive control genes, we applied the optimized window approach in all genes to identify regions that presumably target AD associated loci (see the details in Supplementary Method 2). We then evaluated the optimized windows using the burden test and sequence-kernel association test (SKAT) for the optimized windows (Supplementary Method 3) [63, 64]. The SKAT is powerful when both risk and protective variants are mixed and when a small proportion of variants are causal, whereas the burden test is more powerful than SKAT when most of the variants are causal and have the same direction of effect [63]. We used these properties to confirm whether the optimized windows successfully captured a cluster harboring risk or protective rare variants for AD.

In all analyses above, we assumed a dominant mode of inheritance (MOI), and defined a rare variant as a variant of $MAF < 0.05$. We removed variants with a minor allele count

(MAC) < 5. Variants were assigned to genes based on their physical positions at the UCSC Genome Browser GRCh37/hg19 human assembly (<https://genome.ucsc.edu/>) [87].

Results

Of the 10,031 subjects included in this study, 5,142 (51.3%) were diagnosed as AD.

There were more females than males in both AD cases and controls. The majority of controls were aged 80 years or older and had no *APOE* ϵ 4 alleles (Table 4.4).

Scan-statistic-based analysis by a sliding window approach in TREM2 and TOMM40

Table 4.5 shows the $\log LR_W$ s for each window determined by a sliding window approach with several different settings in *TREM2* (risk effect) and *TOMM40* (protective effect). In both *TREM2* and *TOMM40*, the $\log LR_W$ s were very similar when the large genetic region G was the whole genome or entire chromosome (6 and 19). When the large genetic region G was the band level (6p21.1 and 19q13.32), the $\log LR_W$ s were a little bit smaller than those for the whole genome and entire chromosome specified as G . The largest $\log LR_W$ ($\log LR_W \approx 31$) in *TREM2* was obtained with the 5kb fixed and 2.5 kb sliding window sizes (41,125,156 - 41,130,156) and the 10kb fixed and 5kb sliding window sizes (41,120,156 - 41,130,156 and 41,125,156 - 41,135,156). From these results, we expected that a meaningful cluster would exist in 41,125,156 - 41,130,156. On the other hand, the setting with the 10kb fixed and 5kb sliding window sizes produced less $\log LR_W$ than other two window settings in *TOMM40*, implying that an interesting cluster would be in 45,395,468 - 45,397,468.

Scan-statistic-based analysis by an optimized window approach in TREM2 and TOMM40

We optimized windows with each large genetic region G in *TREM2* and *TOMM40* (Table 4.6). As expected above, we detected the optimized window 41,126,655 - 41,129,252 in *TREM2* existing within 41,125,156 - 41,130,156 and containing rs75932628 (R47H, position: 41,129,252), whichever large genetic region G was specified. In *TOMM40*, the optimized window contained the associated single variant rs1160983 (S183S, position: 45,397,229) that was located within 45,395,468 - 45,397,468 detected by the sliding window approach.

Genome-widely scan-statistic-based analysis by an optimized window approach

We applied the optimized window approach for all genes to identify clusters harboring risk or protective variants associated with AD (Figure 4.3). Top 10 genes with large $\log LR_W$ for each risk and protective effects were displayed in Table 4.7. Besides *TREM2* and *TOMM40*, the optimized windows in *MUC6*, *NXNLI*, and *BCAM* had larger $\log LR_W$ than that in *TREM2*, and the optimized windows in *MUC6*, *CADPS2*, *TYRO3*, *ADM*, *BCAM*, *CBLC*, and *LNPI* captured the variants significantly associated with AD in the single-variant-based analysis (Table 4.3). For *MUC6*, we detected two optimized windows, one with a cluster of risk variants and the other with protective variants, both of which had larger $\log LR_W$ than that in *TREM2*. The risk region 1,018,274 - 1,018,379 contained 4 risk variants including two significant variants rs202193006 (P1485S, position: 1,018,348) and rs765785447 (A1474A, position: 1,018,379). Similarly, the protective region 1,016,800 - 1,017,015 contained 9 protective variants including the significant variant rs373231068 (P1971L, position: 1,016,889) (Table 4.8). For *NXNLI*,

the optimized window 17,565,477 - 17,566,489 were detected, harboring 4 risk rare variants with relatively high effect sizes, although these variants did not reach the FDR-adjusted significance level (Table 4.8). The optimized protective window 45,312,432 - 45,316,606 in *BCAM* was observed including two significant protective variants; rs28399653 (R77H, position: 45,315,445) and rs28399654 (V196I, position: 45,316,588) (Table 4.8).

Discussion

Based on the assumption that variants located in close proximity within a functional protein-coding domain play a similar role in increased or decreased disease susceptibility, we identified clusters of rare variants using the scan-statistic-based analysis with an optimized window approach. First, we evaluated the scan statistics of *TREM2* and *TOMM40* with different settings for a large genetic region G , a fixed window size W , and a sliding window size by a sliding window approach. We obtained very similar $\log LR_W$ values when we specified the whole genome and chromosome as a large genetic region G . This may indicate that the scan statistics are comparable between windows in different large genetic regions if the G is large enough. Second, we optimized windows in the scan-statistic-based analysis. Ideally, it would be the best that an optimized window covers an entire gene if the gene itself affects a disease and captures one single variant if only the variant within the gene is significant. Our optimized window approach successfully identified the gene and the single variant in *TREM2* and *TOMM40*, respectively. Of the 9 variants existing in the *TREM2* gene region in the analysis, 8 variants were contained in the optimized window. In addition, the p-value from the

burden test was smaller than that from SKAT as shown in Table 4.7. These may indicate that this optimized window successfully captured variants with the same direction of effect, and the entire *TREM2* gene is associated with AD susceptibility as consistent to the previous report [158]. On the other hand, only one protective variant in *TOMM40* was identified by the optimized window approach. There existed 3 variants in the *TOMM40* gene region. One was insignificant (OR = 0.71 and p-value = 0.71) and the other was rather risk (OR = 8.1 and p-value = 7.70×10^{-4}).

Applying the optimized window approach in all genes, we detected two regions in *MUC6*; 1,018,274 - 1,018,379 with a risk effect and 1,016,800 - 1,017,015 with a protective effect, both of which were larger $\log LR_W$ than *TREM2*. That is, risk and protective clusters coexist in *MUC6*. If a region has heterogeneity effect (i.e., including both risk and protective variant), the sliding window approach may not be able to detect regions unless proper window sizes are specified. The optimized window approach demonstrated that there exists two interesting regions in *MUC6*; one with a cluster of risk variants including two significant variants and the other with protective variants including one significant variant. The optimized window approach also detected the cluster in *NXNLI* that captured 4 adjacent risk variants which are in strong LD with each other ($r^2 = 0.87$ to 0.98 and $D' = 0.99$ to 1). Although these variants were not shown to be significantly associated with AD in the single-variant-based test, the $\log LR_W$ of the optimized region was larger than that in *TREM2*, indicating that there is a cluster of disease risk variants within this region. The optimized window in *BCAM* was relatively large (the length was about 4kb) that captured 9 variants including two significant

protective variants. *BCAM* is located close to the *APOE* locus. The pairwise LD values were $r^2 = 0.09$ and $D' = 0.46$ between two significant protective variants and *APOE* $\epsilon 2$. Thus, the protective effect of *BCAM* may come from the *APOE* $\epsilon 2$ protective impact on AD as well as *TOMM40*.

In recent years, the importance of *TREM2* has been highlighted due to increased risk for AD. Consistent with previous studies [155, 156, 159, 160], the missense variant rs75932628 (R47H) in *TREM2* was detected as a risk variant of AD in the single-variant-based analysis. *TREM2* located on chromosome 6p21.1 encodes a transmembrane receptor primarily expressed in microglia [161, 162], and has been shown to suppress inflammatory responses [163]. We optimized a window in which there were 8 risk variants with a relatively large effect size, all of which were missense variants except rs144250872 (L133L). The two variants rs2234253 (T96K) and rs2234256 (L211P) were in perfect LD with each other. The variant rs2234255 (H157Y) had the largest estimated OR (OR = 10.48, p-value = 6.50×10^{-3}), showing a MAF of 0.11% in AD cases and 0.01% in controls.

The association of rs75932628 (R47H) was replicated in many studies for individuals of European ancestry [158, 164-167], while other studies failed to replicate this association in East Asian and African American populations [168-171]. The association of the variant rs2234255 (H157Y) was found in the study of the Han Chinese [172] but not of European populations so far. Also, the variant rs2234256 (L211P) was identified in African American [171]. These disparities may be raised because of high variabilities of

MAF in the rare variants among different populations. According to Exome Aggregation Consortium (ExAC) data, for example, the MAFs are 0.26% in European, 0.088% in African American, and 0% in East Asian for rs75932628 (R47H), 0.0030% in European, 0.059% in African American, and 0.20% in East Asian for rs2234255 (H157Y), and 0.096% in European, 12.8% in African American, and 0.15% in East Asian for rs2234256 (L211P) [173]. Since we extracted individuals of European ancestry based on PCs in this study, there may not be enough power to detect rs2234255 (H157Y) in the single-variant-based analysis, although the effect size was large. We did identify the genetic region (i.e., 41,126,655 - 41,129,252) using the scan-statistic-based analysis with an optimized window approach in which potential risk variants previously identified across various populations were harbored.

Using the large WES dataset derived from multiple research centers, we demonstrated the practicability of the optimized window approach in the scan-statistic-based analysis to detect a single variant and a cluster harboring risk or protective variants for AD. We identified several candidate genes exhibiting risk for AD and others playing a protective role using the optimized window approach. Our results strongly suggest the need for more investigation for these novel genes. As more NGS data become available, it is of great interest to see whether these findings are replicated in other cohorts of European ancestry as well as different populations. We also plan in future studies to examine how these mutations change the protein structure and function, and/or whether transgenic mice with these mutations exhibit AD-related pathogenesis.

Funding

This work was supported by the National Cell Repository for Alzheimer's Disease (U24 AG21886), and National Institute on Aging (K25 AG043546, UL1TR000117, the UK-ADC P30 AG028383, and R01 AG057187).

Table 4.1. Summary of the major findings in rare variants associated with late-onset Alzheimer's disease

Gene	Chr	SNV	Position	Variant	Reference
<i>BINI</i>	2	rs138047593	127,808,046	K358R	[174]
<i>UNC5C</i>	4	rs137875858	96,091,431	T835M	[175]
<i>TREM2</i>	6	rs143332484	41,129,207	R62H	[158, 176]
		rs75932628	41,129,252	R47H	[155, 156]
<i>CD2AP</i>	6	rs116754410	47,591,941	K633R	[174]
<i>AKAP9</i>	7	rs144662445	91,709,085	I2546M	[177]
		rs149979685	91,732,110	S3767L	[177]
<i>EPHA1</i>	7	rs202178565	143,095,499	P460L	[174]
<i>SORL1</i>	11	rs117260922	121,367,627	E270K	[178]
		rs143571823	121,429,476	T947M	[178]
<i>TM2D3</i>	15	rs139709573	102,186,966	P155L	[179]
<i>PLCG2</i>	16	rs72824905	81,942,028	P522R	[176]
<i>ABI3</i>	17	rs616338	47,297,297	S209F	[176]
<i>PLD3</i>	19	rs145999145	40,877,595	V232M	[180]
<i>ABCA7</i>	19	rs72973581	1,043,103	G215S	[74]
		rs770510230	1,058,154	E1679X	[174]

Chr = chromosome; SNV = single nucleotide variant

Table 4.2. Number of cases/controls and age range in each case-control study of ADSP

Consortium	Study	Cases		Controls	
		n	Age range	n	Age range
ADGC		4,966	40-99+	3,209	42-99+
	ACT	273	69-89	996	68-89
	ADC	2,417	60-90+	839	64-90+
	CHAP	27	68-90+	204	78-90+
	EFIGA	160	59-90+	171	42-90+
	GDF	111	59-90+	96	77-90+
	NIA-LOAD	364	37-90+	111	78-90+
	MAP	132	71-90+	283	72-90+
	MAYO	250	60-87	99	78-90+
	MAYO PD	181	59-89	14	79-90+
	MIA	316	56-88	15	78-89
	MIRAGE	0	-	20	74-90+
	NCRAD	160	58-90+	0	-
	RAS	46	56-88	0	-
	ROS	144	63-90+	207	67-90+
	TARCC	132	60-90+	12	80-89
TOR	9	40-84	0	-	
VAN	210	60-90+	26	79-90+	
WHICAP	34	73-90+	116	78-90+	
CHARGE		805	60-99+	1,927	61-99+
	ARIC	39	67-89	18	77-85
	ASPS	121	60-89	5	78-86
	CHS	251	68-90+	583	76-90+
	ERF	45	60-88	0	-
	FHS	126	65-90+	455	61-90+
	RS	223	61-90+	866	76-90+

ADSP = Alzheimer's Disease Sequencing Project; ACT =Adult Changes in Thought; ADC = NIA Alzheimer Disease Centers; CHAP = Chicago Health and Aging Project; EFIGA = Estudio Familiar de la Influencia Genetica en Alzheimer; GDF = Genetic Differences; NIA-LOAD = National Institute on Aging (NIA) Late Onset Alzheimer's Disease Family Study; MAP = Memory and Aging Project; MAYO = Mayo Clinic; MAYO PD = Mayo PD; MIA = University of Miami; MIRAGE = Multi-Institutional Research in Alzheimer's Genetic Epidemiology; NCRAD = National Cell Repository for Alzheimer's Disease; RAS = University of Washington Families; ROS = Religious Orders Study; TARCC = Texas Alzheimer's Research and Care Consortium; TOR = University of Toronto; VAN = Vanderbilt University; WHICAP = Washington Heights-Inwood Columbia Aging Project; ARIC = Atherosclerosis Risk in Communities Study; ASPS = Austrian Stroke Prevention Study; CHS = Cardiovascular Health Study; ERF = Erasmus Rucphen Family; FHS = Framingham Heart Study; RS = Rotterdam Study

Table 4.3. Significantly associated rare coding variants with Alzheimer's disease assuming a dominant mode of inheritance

Chr	SNV ID	Position	Gene	Alleles ^a	MAF (%)		No. of subjects with minor allele		OR	P-value ^b	Variant
					Cases	Controls	Cases	Controls			
3	rs9844083	100,170,628	<i>LNP1</i>	A/G	0.33	0.97	28	76	0.33	1.59×10^{-7}	Q74Q
6	rs75932628	41,129,252	<i>TREM2</i>	C/T	0.94	0.21	96	21	4.41	4.30×10^{-12}	R47H
7	rs746999306	122,303,575	<i>CADPS2</i>	A/C	0.35	0.03	35	3	11.31	7.49×10^{-8}	C168G
11	rs373231068	1,016,889	<i>MUC6</i>	G/A	2.26	4.73	194	380	0.45	4.97×10^{-19}	P1971L
11	rs202193006	1,018,348	<i>MUC6</i>	G/A	4.38	2.29	434	219	2.00	1.16×10^{-16}	P1485S
11	rs765785447	1,018,379	<i>MUC6</i>	G/C	3.59	1.90	359	182	1.96	1.41×10^{-13}	A1474A
11	rs764691516	10,327,875	<i>ADM</i>	G/C	0.97	0.32	87	29	3.08	3.10×10^{-8}	R14P
15	rs149022093	41,862,356	<i>TYRO3</i>	T/C	2.25	1.12	188	92	2.05	1.18×10^{-8}	1382+2T>C
19	rs3208856	45,296,806	<i>CBLC</i>	C/T	2.68	4.15	272	392	0.64	3.35×10^{-8}	H405Y
19	rs28399653	45,315,445	<i>BCAM</i>	G/A	2.47	3.90	249	369	0.62	1.45×10^{-8}	R77H
19	rs28399654	45,316,588	<i>BCAM</i>	G/A	2.45	4.02	248	383	0.60	5.75×10^{-10}	V196I
19	rs1160983	45,397,229	<i>TOMM40</i>	G/A	1.80	4.34	140	340	0.39	2.12×10^{-21}	S183S

^a Major/minor alleles

^b P-values were calculated by Fisher's exact test and displayed if significant based on false discovery rate (FDR) adjusted p-value < 0.05.

Chr = chromosome; SNV= single nucleotide variant; MAF = minor allele frequency; OR = odds ratio

Table 4.4. Characteristics of ADSP study subjects (n = 10,031)

Variable	AD cases (n = 5,142)		AD controls (n = 4,889)	
	n	%	n	%
Sex				
Male	2,226	43.3	1,995	40.8
Female	2,916	56.7	2,894	59.2
Age ^a				
65 - 69	506	9.8	12	0.2
70 - 74	1,008	19.6	29	0.6
75 - 79	1,105	21.5	169	3.5
80 - 84	1,041	20.2	1,510	30.9
85 - 89	866	16.8	2,197	55.9
90 +	616	12.0	972	19.9
<i>APOE</i>				
- / -	3,001	58.4	4,189	85.7
ε4 / -	2,068	40.2	683	14.0
ε4 / ε4	73	1.4	17	0.3

^a Age at the last visit or at death

ADSP = Alzheimer's Disease Sequencing Project

Table 4.5. Scan-statistic (logLRw) for windows determined by a sliding window approach in several different settings, *TREM2* and *TOMM40*

Gene	Window size		Window position		G	logLRw	
	Fixed	Sliding	Start	End			
Risk effect							
<i>TREM2</i>	2kb	1kb	41,128,156	41,130,156	Genome	25.42	
					Chromosome 6	25.43	
					Band 6p21.1	24.19	
				41,129,156	41,131,156	Genome	20.27
						Chromosome 6	20.27
						Band 6p21.1	19.24
	5kb	2.5kb	41,125,156	41,130,156	Genome	31.02	
					Chromosome 6	31.03	
					Band 6p21.1	29.57	
				41,127,656	41,132,656	Genome	25.42
						Chromosome 6	25.43
						Band 6p21.1	24.19
	10kb	5kb	41,120,156	41,130,156	Genome	31.01	
					Chromosome 6	31.02	
					Band 6p21.1	29.55	
				41,125,156	41,135,156	Genome	31.02
						Chromosome 6	31.03
						Band 6p21.1	29.57

logLRw = log likelihood ratio for the window

Table 4.5. (Continued)

Gene	Window size		Window position		G	logLRw
	Fixed	Sliding	Start	End		
Protective effect						
<i>TOMM40</i>	2kb	1kb	45,395,468	45,397,468	Genome	48.43
					Chromosome 19	48.13
					Band 19q13.32	44.85
			45,396,468	45,398,468	Genome	48.43
					Chromosome 19	48.13
					Band 19q13.32	44.85
	5kb	2.5kb	45,393,968	45,398,968	Genome	48.43
					Chromosome 19	48.13
					Band 19q13.32	44.85
			45,396,468	45,401,468	Genome	48.43
					Chromosome 19	48.13
					Band 19q13.32	44.85
	10kb	5kb	45,391,468	45,401,468	Genome	38.37
					Chromosome 19	38.06
					Band 19q13.32	34.64
			45,396,468	45,406,468	Genome	40.10
					Chromosome 19	39.82
					Band 19q13.32	36.74

logLRw = log likelihood ratio for the window

Table 4.6. Scan-statistic (logLRw) for windows determined by an optimized window approach in each large genetic region, *TREM2* and *TOMM40*

Gene	Optimized window		G	logLRw
	Start	End		
Risk effect				
<i>TREM2</i>	41,126,655	41,129,252	Genome	31.19
	41,126,655	41,129,252	Chromosome 6	31.20
	41,126,655	41,129,252	Band 6p21.1	29.75
Protective effect				
<i>TOMM40</i>	45,397,229	45,397,229	Genome	48.43
	45,397,229	45,397,229	Chromosome 19	48.13
	45,397,229	45,397,229	Band 19q13.32	44.85

logLRw = log likelihood ratio for the window

Table 4.7. Optimized windows for top 10 genes with a large scan-statistic in risk and protective effects

Chr	Optimized window		Gene	logLRw	Optimized window p-value ^a		No. of subjects	No. of variants in the window
	Start	End			SKAT	Burden		
Risk effect								
11	1,018,274	1,018,379	<i>MUC6</i>	56.19	1.16×10^{-16}	9.05×10^{-17}	9,664	4
19	17,566,477	17,566,489	<i>NXNL1</i>	45.88	7.39×10^{-7}	7.20×10^{-7}	7,679	4
6	41,126,655	41,129,252	<i>TREM2</i>	31.19	2.55×10^{-11}	7.22×10^{-14}	9,801	8
7	149,482,580	149,520,553	<i>SSPO</i>	21.19	0.18	0.30	2,369	137
4	190,903,822	190,903,880	<i>TUBB4Q</i> ^b	18.59	0.045	0.072	9,175	9
17	18,907,029	18,923,180	<i>SLC5A10</i>	17.21	5.18×10^{-4}	4.51×10^{-4}	9,829	8
14	74,763,064	74,766,360	<i>ABCD4</i>	16.16	8.48×10^{-3}	8.48×10^{-3}	10,029	4
7	122,303,575	122,303,598	<i>CADPS2</i>	16.09	1.93×10^{-7}	1.11×10^{-7}	9,825	2
15	41,862,356	41,862,520	<i>TYRO3</i>	14.48	8.12×10^{-10}	1.81×10^{-8}	8,243	3
11	10,327,875	10,327,875	<i>ADM</i>	13.68	3.10×10^{-8c}	-	9,020	1
Protective effect								
11	1,016,800	1,017,015	<i>MUC6</i>	51.62	1.14×10^{-12}	2.00×10^{-10}	6,933	9
19	45,397,229	45,397,229	<i>TOMM40</i>	48.43	2.12×10^{-21c}	-	7,978	1
19	45,312,432	45,316,606	<i>BCAM</i>	34.95	1.89×10^{-8}	1.00×10^{-8}	8,909	9
19	45,296,767	45,297,479	<i>CBLC</i>	24.83	2.66×10^{-5}	4.47×10^{-6}	9,457	6
11	55,594,868	55,595,291	<i>OR5L2</i>	22.98	1.98×10^{-4}	9.39×10^{-5}	9,655	7
2	89,246,948	89,246,978	<i>IGKV1-5</i>	16.31	2.19×10^{-3}	1.02×10^{-3}	9,561	4
19	55,179,184	55,179,217	<i>LILRB4</i>	16.29	6.38×10^{-5}	2.65×10^{-5}	9,774	3
6	30,954,438	30,955,218	<i>MUC21</i>	15.71	0.11	0.17	6,702	55
3	100,170,589	100,170,628	<i>LNPI</i>	14.92	2.11×10^{-7}	1.02×10^{-6}	7,798	2
14	50,799,018	50,857,010	<i>CDKLI</i>	14.61	1.71×10^{-4}	9.19×10^{-3}	7,995	15

^a P-values obtained from the subjects with no missing variants; ^b Pseudogene; ^c P-values were calculated by Fisher's exact test
logLRw = log likelihood ratio for the window; SKAT = sequence-kernel association test; Burden = the burden test

Table 4.8. Single-variant-based association within the optimized windows

SNV ID	Position	Alleles ^a	MAF (%)		No. of subjects with minor allele		OR	P-value ^b	Variant
			Cases	Controls	Cases	Controls			
Risk effect									
11: 1,018,274 - 1,018,379 (<i>MUC6</i>)									
rs79073076	1,018,274	G/C	0.04	0.01	4	1	3.81	0.38	T1509T
rs200240449	1,018,334	C/T	0.27	0.19	27	19	1.35	0.38	T1489T
rs202193006	1,018,348	G/A	4.38	2.29	434	219	2.00	1.16×10⁻¹⁶	P1485S
rs765785447	1,018,379	G/C	3.59	1.90	359	182	1.96	1.41×10⁻¹³	A1474A
19: 17,566,477 - 17,566,489 (<i>NXNLI</i>)									
rs773959663	17,566,477	G/C	1.02	0.38	88	32	2.73	4.41×10 ⁻⁷	G206G
rs761407534	17,566,481	T/C	0.93	0.33	80	28	2.84	5.46×10 ⁻⁷	E205G
rs767189869	17,566,484	T/C	0.75	0.27	71	23	2.84	6.43×10 ⁻⁷	E204G
rs750720749	17,566,489	A/C	0.86	0.28	68	23	3.10	6.58×10 ⁻⁶	G203G
6: 41,126,655 - 41,129,252 (<i>TREM2</i>)									
rs2234256	41,126,655	A/G	0.21	0.09	22	9	2.33	0.031	L211P
rs2234255	41,127,543	G/A	0.11	0.01	11	1	10.48	6.50×10 ⁻³	H157Y
rs144250872	41,127,613	C/A	0.16	0.10	17	10	1.61	0.25	L133L
rs145080901	41,129,078	G/A	0.04	0.02	4	2	1.90	0.69	A105V
rs2234253	41,129,105	G/T	0.21	0.09	22	9	2.33	0.031	T96K
rs142232675	41,129,133	C/T	0.21	0.09	22	9	2.33	0.031	D87N
rs143332484	41,129,207	C/T	1.35	0.90	137	87	1.51	2.88×10 ⁻³	R62H
rs75932628	41,129,252	C/T	0.94	0.21	96	21	4.41	4.30×10⁻¹²	R47H

A bold SNV and p-value represent the significant variants based on the false discovery rate (FDR)-adjusted significance level.

^a Major/minor alleles, ^b P-values were calculated by Fisher's exact test.

SNV= single nucleotide variant; MAF = minor allele frequency; OR = odds ratio

Table 4.8. (Continued)

SNV ID	Position	Alleles ^a	MAF (%)		No. of subjects with minor allele		OR	P-value ^b	Variant
			Cases	Controls	Cases	Controls			
Protective effect									
11: 1,016,800 - 1,017,015 (<i>MUC6</i>)									
rs199626069	1,016,800	G/A	0.04	0.07	4	7	0.54	0.38	P2001S
rs761958882	1,016,801	G/C	0.03	0.06	3	6	0.48	0.33	H2000Q
rs756062369	1,016,809	G/T	0.04	0.07	4	7	0.54	0.38	P1998T
rs747778866	1,016,818	A/G	0.54	0.97	55	93	0.56	6.33×10 ⁻⁴	Y1995H
rs762086454	1,016,835	A/G	0.50	1.02	51	98	0.49	3.09×10 ⁻⁵	F1989S
rs76307106	1,016,870	G/A	0.10	0.11	10	10	0.94	1	P1977P
rs373231068	1,016,889	G/A	2.26	4.73	194	380	0.45	4.97×10⁻¹⁹	P1971L
rs200695483	1,016,957	T/C	0.03	0.03	3	3	0.94	1	R1948R
rs767697427	1,017,015	T/G	0.02	0.06	2	6	0.32	0.17	E1929A
19: 45,312,432 - 45,316,606 (<i>BCAM</i>)									
rs767090237	45,312,432	G/A	0.02	0.05	2	4	0.47	0.44	L17L
rs573141230	45,314,496	GTGCGCT/G	0.08	0.14	8	13	0.58	0.28	R34_L35del
rs28399653	45,315,445	G/A	2.47	3.90	249	369	0.62	1.45×10⁻⁸	R77H
rs3745159	45,315,539	G/A	0.22	0.30	23	29	0.75	0.33	G108G
rs144124876	45,315,573	G/A	0.03	0.02	3	2	1.43	1.00	E120K
rs200398713	45,315,656	T/A	0.10	0.06	10	6	1.59	0.46	433+8T>A
rs143018179	45,315,799	C/G	0.22	0.32	23	31	0.70	0.22	A166A
rs28399654	45,316,588	G/A	2.45	4.02	248	383	0.60	5.75×10⁻¹⁰	V196I
rs776849980	45,316,606	G/A	0.02	0.07	2	7	0.27	0.10	601+3G>A

A bold SNV and p-value represent the significant variants based on the false discovery rate (FDR)-adjusted significance level.

^a Major/minor alleles, ^b P-values were calculated by Fisher's exact test.

SNV= single nucleotide variant; MAF = minor allele frequency; OR = odds ratio

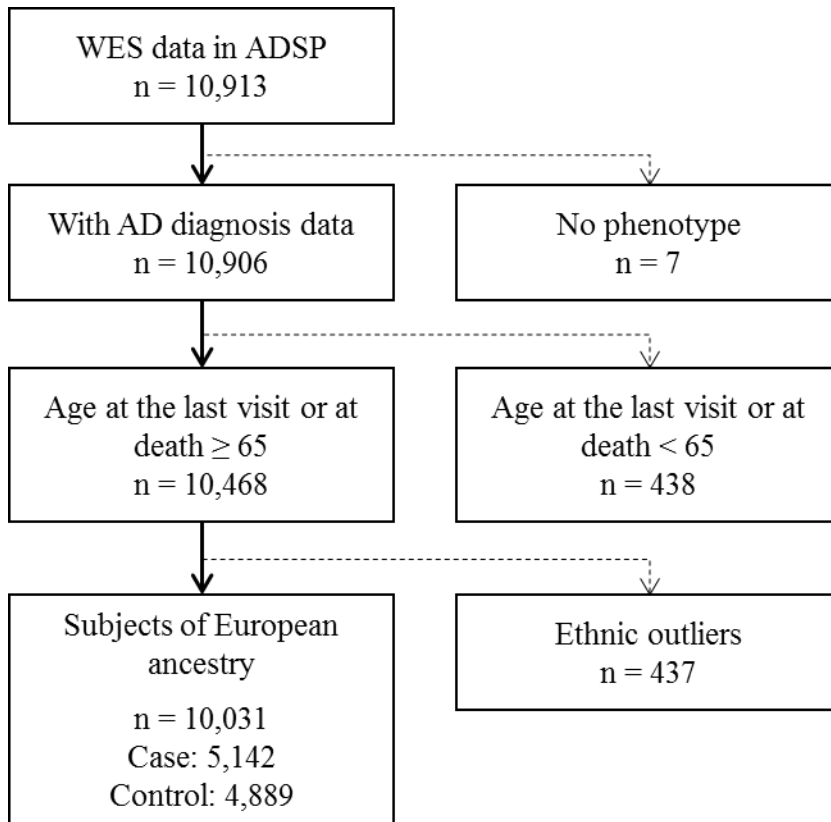


Figure 4.1. Flow diagram of the subjects included in the analyses.

WES = Whole-exome sequencing; ADSP = Alzheimer's Disease Sequencing Project

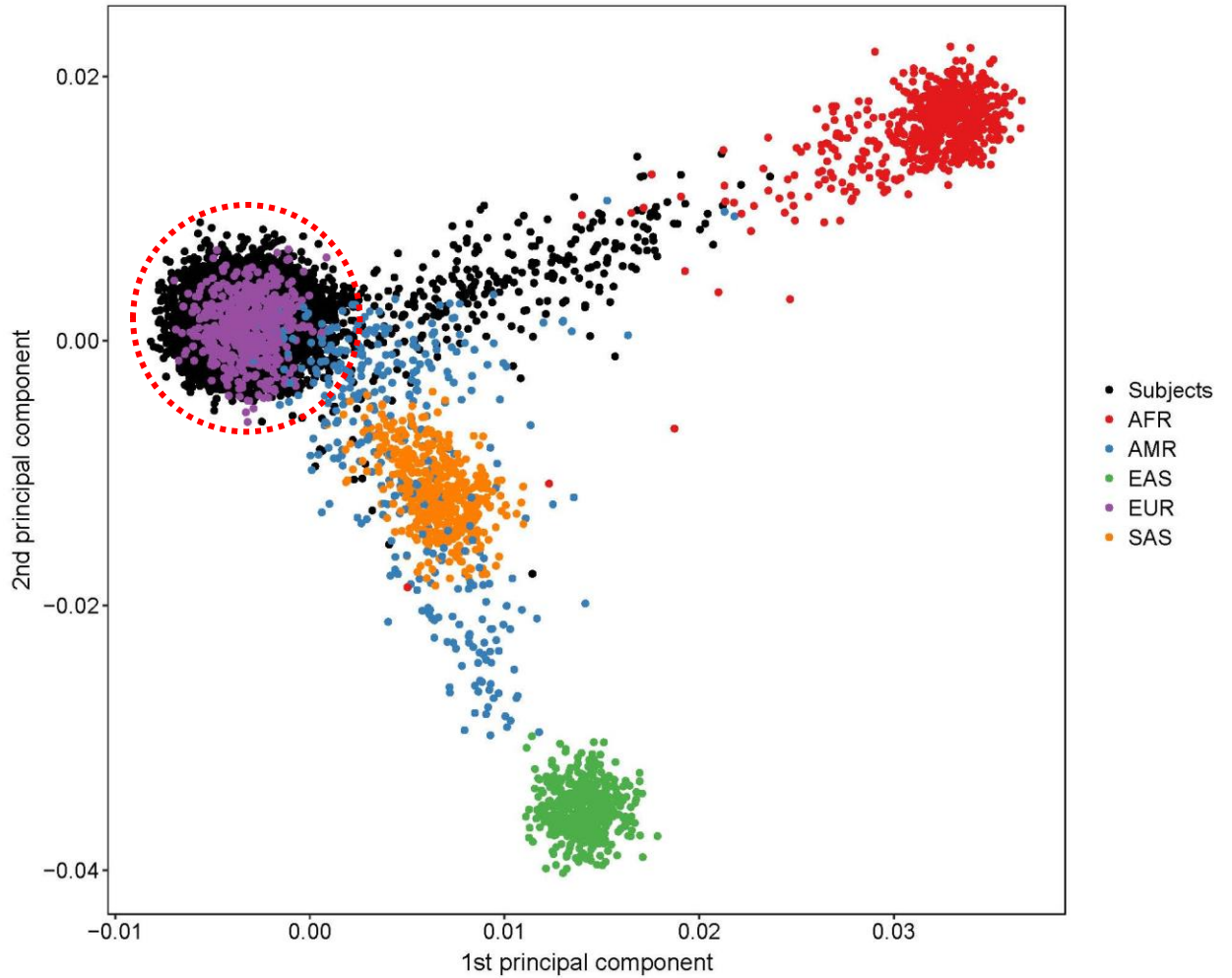


Figure 4.2. First and second principal components plots along with 1000 genome reference samples. Block dots indicate individuals in this study. We chose individuals within the red dotted circle based on Euclidean distance.

AFR = African; AMR = Admixed American; EAS = East Asian; EUR = European; SAS = South Asian

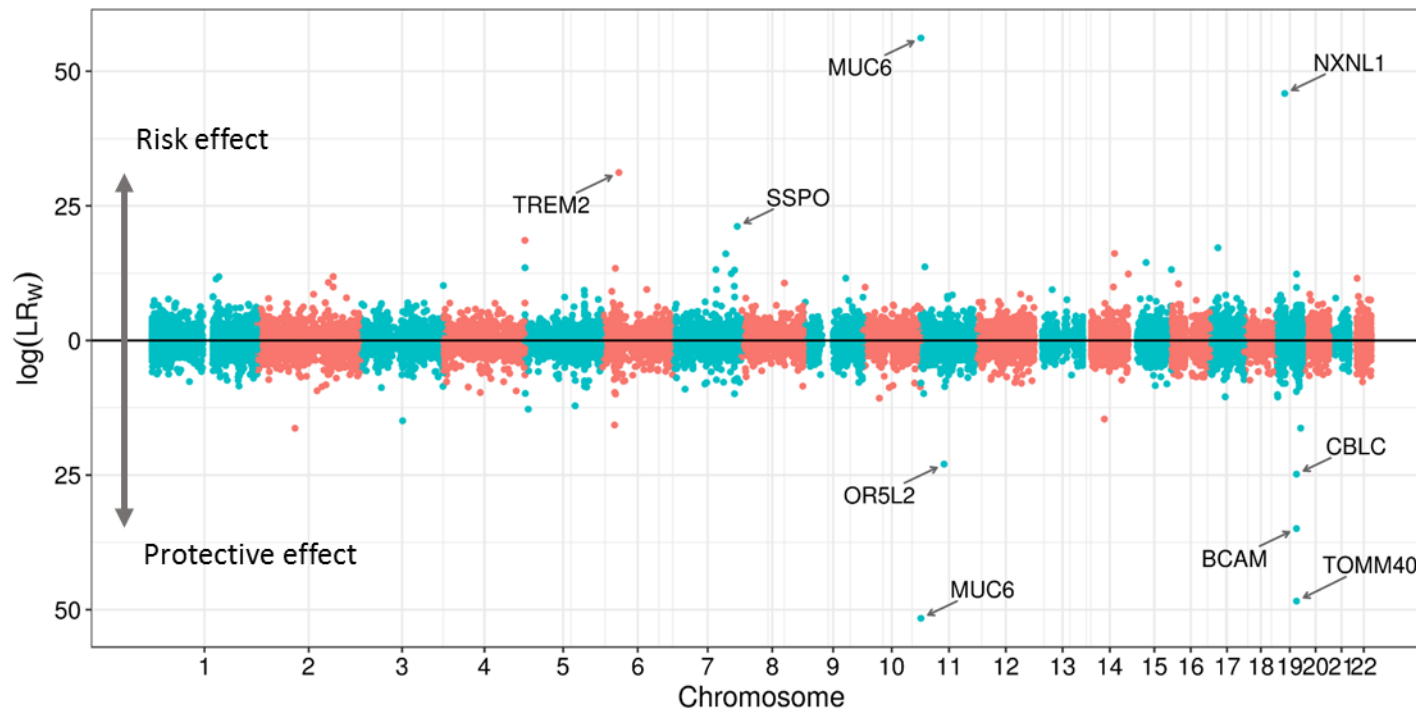


Figure 4.3. Manhattan plot of scan-statistic ($\log LR_w$) for the optimized windows in each gene.

Supplementary Method 1

Quality control (QC) was performed for each of the variant call formats (VCFs) received from two institutions; the Human Genome Sequencing Center at Baylor College of Medicine (hereafter, Baylor) in which genotype calling was conducted using Atlas V2 software [181] and the Broad Institute (hereafter, Broad) in which genotype calling was conducted using Genome Analysis Toolkit (GATK)-HaplotypeCaller [182]. Both of the VCFs were derived from the same set of BAM file.

Primary QC

Distinct primary QCs were applied to each of the sets of VCFs due to differences in the calling pipelines. For the Broad VCFs, variants that did not have a “PASS” in the FILTER field were deleted. For the Baylor VCFs, variants with a low mapping score and genotypes that did not have a “PASS”, that had a low read depth, or that had an out-of-range variant read to total read depth ratio were deleted. Then, monomorphic variants, variants with high missing rate ($\geq 20\%$), variants with high read depth (> 500 reads), or variants (only for $MAF > 0.001$) with Hardy-Weinberg equilibrium p-value $< 5 \times 10^{-6}$ were deleted.

Concordance check between the Baylor and Broad VCFs

Variants were compared between the QCed VCFs. The consensus protocol was defined as follows.

- (1) Variants in which a different alternative allele was called between the two VCFs were excluded.

- (2) All genotypes in remaining variants that were concordant between the two QCed sets of VCFs were included in the final consensus set.
- (3) All genotypes that were discordant between the two QCed sets of VCFs were set to missing.
- (4) All genotypes that were present only in the QCed Baylor VCF (but were missing in the QCed Broad VCF) were included in the final consensus set.
- (5) Genotypes that were present only in the QCed Broad VCF (but were missing in the QCed Baylor VCF) that met a GQ threshold were included in the final consensus set.
 - i. The genotype quality score (GQ) threshold was set to the 0.1 percentile (genotype-specific) based on genome-wide comparisons of WES and GWAS genotypes.
 - ii. Genotypes from the Broad VCF that were not present in the Baylor VCF were excluded if they had GQ values less than 21 for “0/0” genotypes, less than 85 for “0/1” genotypes or less than 36 for “1/1” genotypes.

Second round of variant level QC

After the consensus genotypes were determined, the same QC protocol as the primary QC was applied.

Supplementary Method 2

In order to optimize a window for a genetic region of interest G (e.g., chromosome), we modified the sliding window approach as follows.

1. We specified a large base genetic region G and a window of interest W to get it to be optimized.
2. A total of M variants in G were sorted in an ascending order based on the physical position.
3. We set the j th window as w_j containing m_j variants ($m_j < M$). We denoted the physical position of the i th variant as p_i with $i = 1, \dots, m_j$.
4. We then calculated forward $\log LR_{W_{ij}}^f$ for each window from p_1 to p_i with $i = 1, \dots, m_j$, and obtained $\arg \max_{p_i} \log LR_{W_{ij}}^f$.
5. We set the end position of window as p_{m_j} and calculated backward $\log LR_{W_{ij}}^b$ for each window from p_i to p_{m_j} with $i = 2, \dots, m_j$, and obtained $\arg \max_{p_i} \log LR_{W_{ij}}^b$.
6. We compared between $\max \log LR_{W_{ij}}^f$ and $\max \log LR_{W_{ij}}^b$.

If $\max \log LR_{W_{ij}}^f > \max \log LR_{W_{ij}}^b$, we set $\log LR_{W_j} = \max \log LR_{W_{ij}}^f$,
otherwise, $\log LR_{W_j} = \max \log LR_{W_{ij}}^b$.

7. We compared $\log LR_{W_{j-1}}$ and $\log LR_{W_j}$.

If $\log LR_{W_j} \leq \log LR_{W_{j-1}}$, we stopped and obtained $\log LR_{W_j}$ and the window as an optimized window so that the length W_j was minimized.

Otherwise, we repeated 3 to 7 until we obtained the maximum value of $\log LR_W$ and its window positions.

Supplementary Method 3 [63, 64]

The SKAT aggregates score test statistics assuming that regression coefficient β_j for the j th variant follows an arbitrary distribution with mean 0 and variance $w_j^2\psi$, where w_j is a weight that depends on MAF. The SKAT test statistic is expressed as

$$Q_S = \sum_{j=1}^m w_j^2 S_j^2 = \sum_{j=1}^m w_j^2 \left\{ \sum_{i=1}^n g_{ij} (y_i - \hat{p}_{i,0}) \right\}^2$$

where g_{ij} is a genotype of the j th variant for the i th subject ($g_{ij} = 0, 1, \text{ or } 2$), and $\hat{p}_{i,0}$ is a estimated probability of the phenotype under the null logistic regression model. The test hypothesis $H_0: \boldsymbol{\beta} = 0$ is equivalent to the hypothesis $H_0: \psi = 0$. Instead of summing up the square of weighted score test statistics, the burden test treats the square of the sum of weighed score test statistics defined as

$$Q_B = \left(\sum_{j=1}^m w_j S_j \right)^2 = \left\{ \sum_{j=1}^m w_j \sum_{i=1}^n g_{ij} (y_i - \hat{p}_{i,0}) \right\}^2$$

The SKAT is powerful when both risk and protective variants are mixed, and when a small proportion of variants are causal, whereas the burden test is more powerful than SKAT when most of the variants are causal and have the same direction of effect [63].

CHAPTER FIVE

Conclusion

Summary

Dementia is a complex condition caused by a number of diseases each with an interplay of genetic and environmental factors. With the advance of molecular genetic technologies, studies have been performed in search of genes influencing dementia susceptibility. GWAS have identified many genetic loci that contribute to dementia. However, most of the loci have moderate to small estimated effects, often making it difficult to reproduce in experimental work such as transgenic mouse models. Ideally, we would perform global analyses including information from DNA sequence to levels of proteins and metabolites, which is deemed “systems genetics” [121, 123]. Genetic data from GWAS can play a part of the role of systems genetics by demonstrating interactions among genes and between gene and gene expression by utilizing eQTL mapping. Recent sequencing technologies have allowed us not only to increase our resolution of genetic associations but also to integrate information for a multi-omics/systems genetics approach that includes genomics, epigenomics, transcriptomics, proteomics, metabolomics, and microbiomics [183, 184].

Disease associated genetic variants have the potential to be a powerful anchor point for unidirectional flow in disease causal networks, in which the genetic variants affect downstream layers (i.e., the levels of transcripts, proteins and metabolites). Integrating omics layer data with genomic data can help to identify causal SNPs and genes/regions and to examine causal pathways leading to disease. The purpose of this dissertation

research was to identify SNPs and genes/regions that are potentially causal for dementia, especially focusing on HS-Aging and AD using genomic and transcript data. Genomic and transcript data were used to conduct three studies: (1) Gene-based association study of genes linked to hippocampal sclerosis of aging neuropathology: *GRN*, *TMEM106B*, *ABCC9*, and *KCNMB2*, (2) Translating Alzheimer's disease risk polymorphisms into functional candidates: a survey of IGAP genes, and (3) Identifying regions harboring Alzheimer's disease related rare variants using scan- statistic-based analysis of exome sequencing data. The major findings from each of these studies are summarized below.

In the first study, we examined the genetic associations of four candidate genes (*GRN*, *TMEM106B*, *ABCC9*, and *KCNMB2*) for HS-Aging pathology using the large autopsy dataset derived from multiple research centers. The important findings are that the significant gene-based association between *ABCC9* and HS-Aging appeared to be driven by a region in which a significant haplotype-based association was found, and that the protective haplotype was associated with down-regulation of the *ABCC9* expression in two independent datasets. The association between *ABCC9* and other dementias has never been reported. That is, *ABCC9* could potentially be a key gene that distinguishes HS-Aging from other types of dementia.

The *ABCC9* gene encodes a transmembrane protein, a part of an ATP-sensitive potassium (K_{ATP}) channel complex. K_{ATP} channels are widely expressed in various brain regions including hippocampus [100, 185]. This channel consists of two distinct subunits: an inwardly rectifying K^+ channel (*Kir6.x*) and a regulatory sulfonylurea receptor (*SURx*)

[98]. Two alternatively spliced isoforms of SUR2 (i.e., SUR2A and SUR2B) are encoded by *ABCC9* differing only in the last 42 C-terminal amino acid residues [100, 186, 187]. K_{ATP} channels are formed by the combination of Kir6.x and SURx in different tissues; Kir6.2/SUR1 in pancreatic β -cells and brain, Kir6.2/SUR2A in heart and skeletal muscle, Kir6.2/SUR2B in smooth muscle and brain, and Kir6.1/SUR2B in vascular smooth muscle [99, 186]. K_{ATP} channels are important for neuroprotection against brain injury. When the ATP levels drop due to hypoxia/ischemia, vascular smooth muscle cell K_{ATP} channels open to increase K^+ efflux, voltage-activated calcium channels close to block Ca^{2+} entry, and in turn, vasodilatation is induced for limiting tissue damage [99, 100]. Given the critical roles in regulation of vascular tone, K_{ATP} channel dysfunction may be involved in cardiovascular-related diseases. Leverenz et al. showed that HS-Aging cases were more likely to have history of stroke, small vessel disease, and hypertension than AD cases [47]. Neltner et al. reported that brains with HS-Aging pathology tended to have arteriolosclerosis in multiple cortical and subcortical regions [103]. Therefore, cerebrovascular factors might be involved in developing HS-Aging via the K_{ATP} channel-dependent activity [105].

Chapter Three examined the SNPs which have been reported to be associated with AD. It is well known that the $\epsilon 4$ allele of *APOE* is the major genetic risk factor for late onset of AD. There are three apoE isoforms, apoE2 (cys112, cys158), apoE3 (cys112, arg158), and apoE4 (arg112, arg158), determined by two SNPs rs429358 (T/C) and rs7412 (C/T), both of which are missense variants. These isoforms have different effects on A β metabolism. The binding ability of the apoE isoforms to A β follows the increasing order

of apoE2, apoE3 and apoE4, and thus apoE2 and apoE3 inhibit the aggregation and enhance the clearance of A β compared to apoE4 [33]. A series of genome-wide association studies (GWAS) have identified AD-associated SNPs in addition to the *APOE* alleles. Although GWAS have succeeded in revealing numerous susceptibility variants for AD, we have relatively little understanding of the functional impact of these loci in regards to AD pathogenesis. To understand disease development mechanisms that genetic variants are associated with, identifying functional genes and/or variants is an important challenge.

In this study, the possible functional effects of the IGAP SNPs on AD were examined under two hypotheses: “the IGAP SNP is a proxy of a coding SNP” and “the IGAP SNP is a regulatory SNP”. For the first hypothesis, rs2296160 in *CRI*, rs9270303, rs1049092, and rs1049086 in *HLA-DRB5*, rs2405442 and rs1859788 in *ZCWPW1*, rs7982 in *CLU*, rs12453 and rs7232 in *MS4A6A*, and rs3752246 in *ABCA7* may be proxies of coding SNPs. For the second hypothesis, rs6656401 in *CRI*, rs10838725 in *CELF1*, and rs8093731 in *DSG2* may be regulatory SNPs affecting AD-associated gene expression. Investigating the functional role of the suspected and replicated SNPs associated with AD is an important next step to understanding the genetic contributions and the functional pathways linking AD developmental mechanisms. AD is a complex disease with a strong genetic component. However, much of the genetic contribution to AD remains unexplained. In future studies, more investigations are needed to investigate how RNA and protein levels as well as their interactions are affected by known AD-correlated genes.

Chapter Four presented rare variants associated with AD. Rare variants have become a focus in the recent past. Imputed rare variants are typically removed from GWAS analyses because of low accuracy of imputation. Recent advances in sequencing technologies enable to accurately genotype rare variants. Whole-exome sequencing (WES) and whole-genome sequencing (WGS) are ideal approaches to identify novel variants and genes associated with complex traits. However, traditional single-variant-based association tests are underpowered to detect rare variant associations unless sample size and/or effect size is very large [62]. Instead of testing single variant individually, more powerful and computationally efficient approaches for aggregating the effects of rare variants have become a standard approach for association testing. In this study, we applied a scan-statistic-based approach and developed an approach to construct optimized windows within a gene to find meaningful clusters harboring risk or protective rare variants for AD based on the assumption that variants located in close proximity within a functional protein-coding domain play a similar role in increased or decreased disease susceptibility. We used *TREM2* and *TOMM40* as highly-replicated positive controls to evaluate the scan-statistic-based analysis with a sliding window and an optimized window approach. In addition to these positive control genes, we applied the optimized window approach in all genes to identify regions that harbor AD associated loci.

We obtained very similar scan statistics when we specified the whole genome and chromosome as a large genetic region G . This may indicate that the scan statistics are comparable between windows in different large genetic regions if the G is large enough.

Ideally, it would be the best that an optimized window covers an entire gene if the gene itself affects a disease and captures one single variant if only the variant within the gene is significant. Our optimized window approach successfully identified the gene and the single variant in *TREM2* and *TOMM40*, respectively. Applying the optimized window approach across the genome, we identified several candidate genes exhibiting risk for AD and others playing a protective role using the optimized window approach.

Strengths and Limitations

A major strength of this dissertation is that we evaluated genetic associations of both common and rare variants with neurodegenerative diseases using several statistical methods. Single-variant-based association tests such as through conventional logistic and linear regression are powerful and useful tools to identify associated common variants having marginal significant effects on disease. Unlike Mendelian disease, however, complex disease is not explained by a single variant in a single gene. As a functional unit, a gene/region contains one or more significant SNPs that jointly affect a disease. Therefore, gene/region-based association analysis is an effective strategy to identify candidate genes/regions contributing to natural variation in a disease, as it combines signals from all SNPs within a putative gene/region. Haplotype analysis is also another informative approach to handle multiple SNPs. It accounts for not only heterogeneity but also possible statistical interactions among SNPs.

There are some limitations. First, all data used in three studies come from multiple research centers. Since a large number of observations are required in genetic studies to

boost statistical power and to obtain more accurate estimates, collaborative groups combine the raw individual data from each study and jointly analyze the pooled data (referred to as mega-analysis). In mega-analysis, a concern about “heterogeneity among studies” of the same trait has been raised. Sources of heterogeneity among studies contains different study designs, trait measurements, ethnic groups, genotyping chips, and so on. The statistical methods applied in this dissertation did not take into account heterogeneity among studies. Second, dementia diagnoses vary across calendar time and research centers. In addition, the accuracy of clinical diagnosis of AD may be hampered by difficulty in differentiating AD from other type of dementia or AD mixed with other neuropathological conditions [16]. Third, we aggregated data from some resources that aid in establishing a confluence within a systems genetics framework. These datasets are heterogeneous and can exhibit biases from the respective study designs, analytic protocols, and participant pools. Last, we used WES to identify novel rare variants associated with AD in the third study. Although WES is a powerful approach for discovering mutations in coding regions, it is difficult or impossible to detect important elements including variants in non-coding regions, large indels and repeat expansions, and exon- or gene-level copy number variations. Therefore, WGS is better than WES to capture comprehensive genomic associations.

Future Research

There are several future research ideas suggested by these dissertation studies. First, we incorporated only common SNPs into the gene-based and haplotype-based analyses in the first study. Although we showed that the associated haplotype with HS-Aging was also

associated with the *ABCC9* expression, it is still possible that the common SNP-association signals come from a synthetic association, i.e., that detectable common SNPs reflect the net effect of multiple functional rare variants on a disease. The synthetic association hypothesis would suggest that confirmation studies for known SNPs should encompass a larger region surrounding the detected common SNP in which possible rare variants creating the synthetic association are contained. Second, we need to evaluate impacts of allele specific expression on biological function related to AD. We excluded monoallelically expressed genes including genes on chromosomes X and Y, and HLA-genes in the second study. However, allele specific expression is widely spread across the genome, and have important implications in the genotype-phenotype associations [188]. Given epigenetic association between DNA methylation in *HLA-DRB5* and AD pathology [142], we will require more investigations for the association between allele specific expression of *HLA-DRB5* and AD pathogenesis. Third, the scan-statistic-based analysis we used in the third study is performed under the important assumption that each risk is independent because it is based on Bernoulli model. That is, the scan-statistic-based analysis does not take into account LD. If multiple associated variants are in LD, using all of them would artificially inflate the test statistic. Therefore, we need to modify it as considering LD structure. In addition, we could expand the scan-statistic-based analysis for a continuous outcome using a normal distribution model so that biomarkers related to AD development would be handled.

REFERENCES

1. Querfurth HW, LaFerla FM: **Alzheimer's disease**. *N Engl J Med* 2010, **362**: 329-344
2. Alzheimer's Association: **2014 Alzheimer's disease facts and figures**. *Alzheimers Dement* 2014, **10**: e47-92
3. Alzheimer A: **Über eine eigenartige Erkrankung der Hirnrinde**. *Allgemeine Z Psychiatrie Psychisch-Gerichtliche Med* 1907, **64**: 146-148
4. Reitz C, Brayne C, Mayeux R: **Epidemiology of Alzheimer disease**. *Nat Rev Neurol* 2011, **7**: 137-152
5. Selkoe DJ, American College of P, American Physiological S: **Alzheimer disease: mechanistic understanding predicts novel therapies**. *Ann Intern Med* 2004, **140**: 627-638
6. Hardy J, Selkoe DJ: **The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics**. *Science* 2002, **297**: 353-356
7. Donahue JE, Johanson CE: **Apolipoprotein E, amyloid-beta, and blood-brain barrier permeability in Alzheimer disease**. *J Neuropathol Exp Neurol* 2008, **67**: 261-270
8. van Es MA, van den Berg LH: **Alzheimer's disease beyond APOE**. *Nat Genet* 2009, **41**: 1047-1048
9. Shen J, Kelleher RJ, 3rd: **The presenilin hypothesis of Alzheimer's disease: evidence for a loss-of-function pathogenic mechanism**. *Proc Natl Acad Sci U S A* 2007, **104**: 403-409
10. Holtzman DM, Morris JC, Goate AM: **Alzheimer's disease: the challenge of the second century**. *Sci Transl Med* 2011, **3**: 77sr71
11. Guo T, Noble W, Hanger DP: **Roles of tau protein in health and disease**. *Acta Neuropathol* 2017, **133**: 665-704
12. Wang Y, Mandelkow E: **Tau in physiology and pathology**. *Nat Rev Neurosci* 2016, **17**: 5-21
13. Hanger DP, Noble W: **Functional implications of glycogen synthase kinase-3-mediated tau phosphorylation**. *Int J Alzheimers Dis* 2011, **2011**: 352805
14. Alonso AC, Zaidi T, Grundke-Iqbal I, Iqbal K: **Role of abnormally phosphorylated tau in the breakdown of microtubules in Alzheimer disease**. *Proc Natl Acad Sci U S A* 1994, **91**: 5562-5566
15. Lim A, Tsuang D, Kukull W, Nochlin D, Leverenz J, McCormick W, Bowen J, Teri L, Thompson J, Peskind ER *et al*: **Clinico-neuropathological correlation of Alzheimer's disease in a community-based case series**. *J Am Geriatr Soc* 1999, **47**: 564-569
16. Beach TG, Monsell SE, Phillips LE, Kukull W: **Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010**. *J Neuropathol Exp Neurol* 2012, **71**: 266-273
17. Gaugler JE, Ascher-Svanum H, Roth DL, Fafowora T, Siderowf A, Beach TG: **Characteristics of patients misdiagnosed with Alzheimer's disease and their medication use: an analysis of the NACC-UDS database**. *BMC Geriatr* 2013, **13**: 137

18. Ranginwala NA, Hynan LS, Weiner MF, White CL, 3rd: **Clinical criteria for the diagnosis of Alzheimer disease: still good after all these years.** *Am J Geriatr Psychiatry* 2008, **16**: 384-388
19. Brenowitz WD, Monsell SE, Schmitt FA, Kukull WA, Nelson PT: **Hippocampal sclerosis of aging is a key Alzheimer's disease mimic: clinical-pathologic correlations and comparisons with both Alzheimer's disease and non-tauopathic frontotemporal lobar degeneration.** *J Alzheimers Dis* 2014, **39**: 691-702
20. Zarow C, Sitzer TE, Chui HC: **Understanding hippocampal sclerosis in the elderly: epidemiology, characterization, and diagnostic issues.** *Curr Neurol Neurosci Rep* 2008, **8**: 363-370
21. Pao WC, Dickson DW, Crook JE, Finch NA, Rademakers R, Graff-Radford NR: **Hippocampal sclerosis in the elderly: genetic and pathologic findings, some mimicking Alzheimer disease clinically.** *Alzheimer Dis Assoc Disord* 2011, **25**: 364-368
22. Hyman BT, Phelps CH, Beach TG, Bigio EH, Cairns NJ, Carrillo MC, Dickson DW, Duyckaerts C, Frosch MP, Masliah E *et al*: **National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease.** *Alzheimers Dement* 2012, **8**: 1-13
23. Nelson PT, Smith CD, Abner EL, Wilfred BJ, Wang WX, Neltner JH, Baker M, Fardo DW, Kryscio RJ, Scheff SW *et al*: **Hippocampal sclerosis of aging, a prevalent and high-morbidity brain disease.** *Acta Neuropathol* 2013, **126**: 161-177
24. Amador-Ortiz C, Lin WL, Ahmed Z, Personett D, Davies P, Duara R, Graff-Radford NR, Hutton ML, Dickson DW: **TDP-43 immunoreactivity in hippocampal sclerosis and Alzheimer's disease.** *Ann Neurol* 2007, **61**: 435-445
25. Nelson PT, Schmitt FA, Lin Y, Abner EL, Jicha GA, Patel E, Thomason PC, Neltner JH, Smith CD, Santacruz KS *et al*: **Hippocampal sclerosis in advanced age: clinical and pathological features.** *Brain* 2011, **134**: 1506-1518
26. Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, Pettingell WH, Yu CE, Jondro PD, Schmidt SD, Wang K *et al*: **Candidate gene for the chromosome 1 familial Alzheimer's disease locus.** *Science* 1995, **269**: 973-977
27. Jayadev S, Leverenz JB, Steinbart E, Stahl J, Klunk W, Yu CE, Bird TD: **Alzheimer's disease phenotypes and genotypes associated with mutations in presenilin 2.** *Brain* 2010, **133**: 1143-1154
28. Walker ES, Martinez M, Brunkan AL, Goate A: **Presenilin 2 familial Alzheimer's disease mutations result in partial loss of function and dramatic changes in A β 42/40 ratios.** *J Neurochem* 2005, **92**: 294-301
29. Lichtenthaler SF, Wang R, Grimm H, Uljon SN, Masters CL, Beyreuther K: **Mechanism of the cleavage specificity of Alzheimer's disease gamma-secretase identified by phenylalanine-scanning mutagenesis of the transmembrane domain of the amyloid precursor protein.** *Proc Natl Acad Sci U S A* 1999, **96**: 3053-3058
30. Mullan M, Crawford F, Axelman K, Houlden H, Lilius L, Winblad B, Lannfelt L: **A pathogenic mutation for probable Alzheimer's disease in the APP gene at the N-terminus of beta-amyloid.** *Nat Genet* 1992, **1**: 345-347

31. Nilsberth C, Westlind-Danielsson A, Eckman CB, Condrón MM, Axelman K, Forsell C, Sten C, Luthman J, Teplow DB, Younkin SG *et al*: **The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Abeta protofibril formation.** *Nat Neurosci* 2001, **4**: 887-893
32. Mancuso M, Orsucci D, Siciliano G, Murri L: **Mitochondria, mitochondrial DNA and Alzheimer's disease. What comes first?** *Curr Alzheimer Res* 2008, **5**: 457-468
33. Tokuda T, Calero M, Matsubara E, Vidal R, Kumar A, Permanne B, Zlokovic B, Smith JD, Ladu MJ, Rostagno A *et al*: **Lipidation of apolipoprotein E influences its isoform-specific interaction with Alzheimer's amyloid beta peptides.** *Biochem J* 2000, **348 Pt 2**: 359-365
34. Cruchaga C, Kauwe JS, Harari O, Jin SC, Cai Y, Karch CM, Benitez BA, Jeng AT, Skorupa T, Carrell D *et al*: **GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease.** *Neuron* 2013, **78**: 256-268
35. Kauwe JS, Cruchaga C, Bertelsen S, Mayo K, Latu W, Nowotny P, Hinrichs AL, Fagan AM, Holtzman DM, Alzheimer's Disease Neuroimaging I *et al*: **Validating predicted biological effects of Alzheimer's disease associated SNPs using CSF biomarker levels.** *J Alzheimers Dis* 2010, **21**: 833-842
36. Morris JC, Roe CM, Xiong C, Fagan AM, Goate AM, Holtzman DM, Mintun MA: **APOE predicts amyloid-beta but not tau Alzheimer pathology in cognitively normal aging.** *Ann Neurol* 2010, **67**: 122-131
37. Buee L, Bussiere T, Buee-Scherrer V, Delacourte A, Hof PR: **Tau protein isoforms, phosphorylation and role in neurodegenerative disorders.** *Brain Res Brain Res Rev* 2000, **33**: 95-130
38. Allen M, Kachadoorian M, Quicksall Z, Zou F, Chai HS, Younkin C, Crook JE, Pankratz VS, Carrasquillo MM, Krishnan S *et al*: **Association of MAPT haplotypes with Alzheimer's disease risk and MAPT brain gene expression levels.** *Alzheimers Res Ther* 2014, **6**: 39
39. Pastor P, Moreno F, Clarimon J, Ruiz A, Combarros O, Calero M, Lopez de Munain A, Bullido MJ, de Pancorbo MM, Carro E *et al*: **MAPT H1 Haplotype is Associated with Late-Onset Alzheimer's Disease Risk in APOEepsilon4 Noncarriers: Results from the Dementia Genetics Spanish Consortium.** *J Alzheimers Dis* 2016, **49**: 343-352
40. Myers AJ, Kaleem M, Marlowe L, Pittman AM, Lees AJ, Fung HC, Duckworth J, Leung D, Gibson A, Morris CM *et al*: **The H1c haplotype at the MAPT locus is associated with Alzheimer's disease.** *Hum Mol Genet* 2005, **14**: 2399-2404
41. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, DeStafano AL, Bis JC, Beecham GW, Grenier-Boley B *et al*: **Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.** *Nat Genet* 2013, **45**: 1452-1458
42. Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo MM, Abraham R, Hamshere ML, Pahwa JS, Moskvina V *et al*: **Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease.** *Nat Genet* 2011, **43**: 429-435
43. Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, Combarros O, Zelenika D, Bullido MJ, Tavernier B *et al*: **Genome-wide association study**

- identifies variants at CLU and CR1 associated with Alzheimer's disease.** *Nat Genet* 2009, **41**: 1094-1099
44. Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buross J, Gallins PJ, Buxbaum JD, Jarvik GP, Crane PK *et al*: **Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease.** *Nat Genet* 2011, **43**: 436-441
45. Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, Pahwa JS, Moskvina V, Dowzell K, Williams A *et al*: **Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease.** *Nat Genet* 2009, **41**: 1088-1093
46. Troncoso JC, Kawas CH, Chang CK, Folstein MF, Hedreen JC: **Lack of association of the apoE4 allele with hippocampal sclerosis dementia.** *Neurosci Lett* 1996, **204**: 138-140
47. Leverenz JB, Agustin CM, Tsuang D, Peskind ER, Edland SD, Nochlin D, DiGiacomo L, Bowen JD, McCormick WC, Teri L *et al*: **Clinical and neuropathological characteristics of hippocampal sclerosis: a community-based study.** *Arch Neurol* 2002, **59**: 1099-1106
48. Beecham GW, Hamilton K, Naj AC, Martin ER, Huentelman M, Myers AJ, Corneveaux JJ, Hardy J, Vonsattel JP, Younkin SG *et al*: **Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias.** *PLoS Genet* 2014, **10**: e1004606
49. Dickson DW, Baker M, Rademakers R: **Common variant in GRN is a genetic risk factor for hippocampal sclerosis in the elderly.** *Neurodegener Dis* 2010, **7**: 170-174
50. Aoki N, Murray ME, Ogaki K, Fujioka S, Rutherford NJ, Rademakers R, Ross OA, Dickson DW: **Hippocampal sclerosis in Lewy body disease is a TDP-43 proteinopathy similar to FTLN-TDP Type A.** *Acta Neuropathol* 2015, **129**: 53-64
51. Nelson PT, Estus S, Abner EL, Parikh I, Malik M, Neltner JH, Ighodaro E, Wang WX, Wilfred BR, Wang LS *et al*: **ABCC9 gene polymorphism is associated with hippocampal sclerosis of aging pathology.** *Acta Neuropathol* 2014, **127**: 825-843
52. Nelson PT, Wang WX, Partch AB, Monsell SE, Valladares O, Ellingson SR, Wilfred BR, Naj AC, Wang LS, Kukull WA *et al*: **Reassessment of risk genotypes (GRN, TMEM106B, and ABCC9 variants) associated with hippocampal sclerosis of aging pathology.** *J Neuropathol Exp Neurol* 2015, **74**: 75-84
53. Murray ME, Cannon A, Graff-Radford NR, Liesinger AM, Rutherford NJ, Ross OA, Duara R, Carrasquillo MM, Rademakers R, Dickson DW: **Differential clinicopathologic and genetic features of late-onset amnesic dementias.** *Acta Neuropathol* 2014, **128**: 411-421
54. Rabbani B, Tekin M, Mahdiah N: **The promise of whole-exome sequencing in medical genetics.** *J Hum Genet* 2014, **59**: 5-15
55. Majewski J, Schwartzenuber J, Lalonde E, Montpetit A, Jabado N: **What can exome sequencing do for you?** *J Med Genet* 2011, **48**: 580-589

56. Sauna ZE, Kimchi-Sarfaty C: **Understanding the contribution of synonymous mutations to human disease.** *Nat Rev Genet* 2011, **12**: 683-691
57. Mockenhaupt S, Makeyev EV: **Non-coding functions of alternative pre-mRNA splicing in development.** *Semin Cell Dev Biol* 2015, **47-48**: 32-39
58. Gibson G: **Rare and common variants: twenty arguments.** *Nat Rev Genet* 2012, **13**: 135-145
59. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR: **Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.** *Nat Genet* 2012, **44**: 955-959
60. van Leeuwen EM, Kanterakis A, Deelen P, Kattenberg MV, Genome of the Netherlands C, Slagboom PE, de Bakker PI, Wijmenga C, Swertz MA, Boomsma DI *et al*: **Population-specific genotype imputations using minimac or IMPUTE2.** *Nat Protoc* 2015, **10**: 1285-1296
61. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**: 1061-1073
62. Asimit J, Zeggini E: **Rare variant association analysis methods for complex traits.** *Annu Rev Genet* 2010, **44**: 293-308
63. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Team NGESP-ELP, Christiani DC, Wurfel MM, Lin X: **Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.** *Am J Hum Genet* 2012, **91**: 224-237
64. Lee S, Wu MC, Lin X: **Optimal tests for rare variant effects in sequencing association studies.** *Biostatistics* 2012, **13**: 762-775
65. Hoh J, Ott J: **Scan statistics to scan markers for susceptibility genes.** *Proc Natl Acad Sci U S A* 2000, **97**: 9615-9617
66. Kulldorff M: **A spatial scan statistic.** *Commun Stat Theory Methods* 1997, **26**: 1484-1496
67. Ionita-Laza I, Makarov V, Consortium AAS, Buxbaum JD: **Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets.** *Am J Hum Genet* 2012, **90**: 1002-1013
68. Shulman JM, Chen K, Keenan BT, Chibnik LB, Fleisher A, Thiyyagura P, Roontiva A, McCabe C, Patsopoulos NA, Corneveaux JJ *et al*: **Genetic susceptibility for Alzheimer disease neuritic plaque pathology.** *JAMA Neurol* 2013, **70**: 1150-1157
69. Biffi A, Anderson CD, Desikan RS, Sabuncu M, Cortellini L, Schmansky N, Salat D, Rosand J, Alzheimer's Disease Neuroimaging I: **Genetic variation and neuroimaging measures in Alzheimer disease.** *Arch Neurol* 2010, **67**: 677-685
70. Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, Bis JC, Smith AV, Carassquillo MM, Lambert JC *et al*: **Genome-wide analysis of genetic loci associated with Alzheimer disease.** *JAMA* 2010, **303**: 1832-1840
71. Shuai P, Liu Y, Lu W, Liu Q, Li T, Gong B: **Genetic associations of CLU rs9331888 polymorphism with Alzheimer's disease: A meta-analysis.** *Neurosci Lett* 2015, **591**: 160-165

72. Huang KL, Marcora E, Pimenova AA, Di Narzo AF, Kapoor M, Jin SC, Harari O, Bertelsen S, Fairfax BP, Czajkowski J *et al*: **A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease.** *Nat Neurosci* 2017, **20**: 1052-1061
73. Wang Z, Lei H, Zheng M, Li Y, Cui Y, Hao F: **Meta-analysis of the Association between Alzheimer Disease and Variants in GAB2, PICALM, and SORL1.** *Mol Neurobiol* 2016, **53**: 6501-6510
74. Sassi C, Nalls MA, Ridge PG, Gibbs JR, Ding J, Lupton MK, Troakes C, Lunnon K, Al-Sarraj S, Brown KS *et al*: **ABCA7 p.G215S as potential protective factor for Alzheimer's disease.** *Neurobiol Aging* 2016, **46**: 235 e231-239
75. Corey-Bloom J, Sabbagh MN, Bondi MW, Hansen L, Alford MF, Masliah E, Thal LJ: **Hippocampal sclerosis contributes to dementia in the elderly.** *Neurology* 1997, **48**: 154-160
76. Zarow C, Weiner MW, Ellis WG, Chui HC: **Prevalence, laterality, and comorbidity of hippocampal sclerosis in an autopsy sample.** *Brain Behav* 2012, **2**: 435-442
77. Dickson DW, Davies P, Bevona C, Van Hoesven KH, Factor SM, Grober E, Aronson MK, Crystal HA: **Hippocampal sclerosis: a common pathological feature of dementia in very old (> or = 80 years of age) humans.** *Acta Neuropathol* 1994, **88**: 212-221
78. Nelson PT, Trojanowski JQ, Abner EL, Al-Janabi OM, Jicha GA, Schmitt FA, Smith CD, Fardo DW, Wang WX, Kryscio RJ *et al*: **"New Old Pathologies": AD, PART, and Cerebral Age-Related TDP-43 With Sclerosis (CARTS).** *J Neuropathol Exp Neurol* 2016, **75**: 482-498
79. Pickering-Brown SM, Rollinson S, Du Plessis D, Morrison KE, Varma A, Richardson AM, Neary D, Snowden JS, Mann DM: **Frequency and clinical characteristics of progranulin mutation carriers in the Manchester frontotemporal lobar degeneration cohort: comparison with patients with MAPT and no known mutations.** *Brain* 2008, **131**: 721-731
80. Rademakers R, Eriksen JL, Baker M, Robinson T, Ahmed Z, Lincoln SJ, Finch N, Rutherford NJ, Crook RJ, Josephs KA *et al*: **Common variation in the miR-659 binding-site of GRN is a major risk factor for TDP43-positive frontotemporal dementia.** *Hum Mol Genet* 2008, **17**: 3631-3642
81. Rutherford NJ, Carrasquillo MM, Li M, Bisceglia G, Menke J, Josephs KA, Parisi JE, Petersen RC, Graff-Radford NR, Younkin SG *et al*: **TMEM106B risk variant is implicated in the pathologic presentation of Alzheimer disease.** *Neurology* 2012, **79**: 717-718
82. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**: 559-575
83. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**: 904-909
84. 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**: 1061-1073

85. Wickham H: **ggplot2 : elegant graphics for data analysis**. New York ; London: Springer Science + Business Media; 2009.
86. Li MX, Gui HS, Kwan JS, Sham PC: **GATES: a rapid and powerful gene-based association test using extended Simes procedure**. *Am J Hum Genet* 2011, **88**: 283-293
87. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome Res* 2002, **12**: 996-1006
88. Nelson PT, Head E, Schmitt FA, Davis PR, Neltner JH, Jicha GA, Abner EL, Smith CD, Van Eldik LJ, Kryscio RJ *et al*: **Alzheimer's disease is not "brain aging": neuropathological, genetic, and epidemiological human studies**. *Acta Neuropathol* 2011, **121**: 571-587
89. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ: **LocusZoom: regional visualization of genome-wide association scan results**. *Bioinformatics* 2010, **26**: 2336-2337
90. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics* 2005, **21**: 263-265
91. Sinnwell JP, Schaid DJ: **haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous**. In., R package version 1.7.7 edn; 2016.
92. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: **Score tests for association between traits and haplotypes when linkage phase is ambiguous**. *Am J Hum Genet* 2002, **70**: 425-434
93. Hernandez DG, Nalls MA, Moore M, Chong S, Dillman A, Trabzuni D, Gibbs JR, Ryten M, Arepalli S, Weale ME *et al*: **Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain**. *Neurobiol Dis* 2012, **47**: 20-28
94. Trabzuni D, Ryten M, Walker R, Smith C, Imran S, Ramasamy A, Weale ME, Hardy J: **Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies**. *J Neurochem* 2011, **119**: 275-282
95. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M *et al*: **Next-generation genotype imputation service and methods**. *Nat Genet* 2016, **48**: 1284-1287
96. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, Schoenherr S, Forer L, McCarthy S, Abecasis GR *et al*: **Reference-based phasing using the Haplotype Reference Consortium panel**. *Nat Genet* 2016, **48**: 1443-1448
97. Yang J, Lee SH, Goddard ME, Visscher PM: **GCTA: a tool for genome-wide complex trait analysis**. *Am J Hum Genet* 2011, **88**: 76-82
98. Quast U, Stephan D, Bieger S, Russ U: **The impact of ATP-sensitive K⁺ channel subtype selectivity of insulin secretagogues for the coronary vasculature and the myocardium**. *Diabetes* 2004, **53 Suppl 3**: S156-164
99. Sun XL, Hu G: **ATP-sensitive potassium channels: a promising target for protecting neurovascular unit function in stroke**. *Clin Exp Pharmacol Physiol* 2010, **37**: 243-252
100. Sun HS, Feng ZP: **Neuroprotective role of ATP-sensitive potassium channels in cerebral ischemia**. *Acta Pharmacol Sin* 2013, **34**: 24-32

101. Miki T, Suzuki M, Shibasaki T, Uemura H, Sato T, Yamaguchi K, Koseki H, Iwanaga T, Nakaya H, Seino S: **Mouse model of Prinzmetal angina by disruption of the inward rectifier Kir6.1.** *Nat Med* 2002, **8**: 466-472
102. Chutkow WA, Pu J, Wheeler MT, Wada T, Makielski JC, Burant CF, McNally EM: **Episodic coronary artery vasospasm and hypertension develop in the absence of Sur2 K(ATP) channels.** *J Clin Invest* 2002, **110**: 203-208
103. Neltner JH, Abner EL, Baker S, Schmitt FA, Kryscio RJ, Jicha GA, Smith CD, Hammack E, Kukull WA, Brenowitz WD *et al*: **Arteriolosclerosis that affects multiple brain regions is linked to hippocampal sclerosis of ageing.** *Brain* 2014, **137**: 255-267
104. Leon Guerrero CR, Pathak S, Grange DK, Singh GK, Nichols CG, Lee JM, Vo KD: **Neurologic and neuroimaging manifestations of Cantu syndrome: A case series.** *Neurology* 2016, **87**: 270-276
105. Nelson PT, Jicha GA, Wang WX, Ighodaro E, Artiushin S, Nichols CG, Fardo DW: **ABCC9/SUR2 in the brain: Implications for hippocampal sclerosis of aging and a potential therapeutic target.** *Ageing Res Rev* 2015, **24**: 111-125
106. Nelson PT, Katsumata Y, Nho K, Artiushin SC, Jicha GA, Wang WX, Abner EL, Saykin AJ, Kukull WA, Alzheimer's Disease Neuroimaging I *et al*: **Genomics and CSF analyses implicate thyroid hormone in hippocampal sclerosis of aging.** *Acta Neuropathol* 2016, **132**: 841-858
107. Rieben C, Segna D, da Costa BR, Collet TH, Chaker L, Aubert CE, Baumgartner C, Almeida OP, Hogervorst E, Trompet S *et al*: **Subclinical Thyroid Dysfunction and the Risk of Cognitive Decline: a Meta-Analysis of Prospective Cohort Studies.** *J Clin Endocrinol Metab* 2016, **101**: 4945-4954
108. Annerbo S, Lökk J: **A clinical review of the association of thyroid stimulating hormone and cognitive impairment.** *ISRN Endocrinol* 2013, **2013**: 856017
109. Pasqualetti G, Pagano G, Rengo G, Ferrara N, Monzani F: **Subclinical Hypothyroidism and Cognitive Impairment: Systematic Review and Meta-Analysis.** *J Clin Endocrinol Metab* 2015, **100**: 4240-4248
110. Sara JD, Zhang M, Gharib H, Lerman LO, Lerman A: **Hypothyroidism Is Associated With Coronary Endothelial Dysfunction in Women.** *J Am Heart Assoc* 2015, **4**: e002225
111. Delitala AP, Orru M, Filigheddu F, Pilia MG, Delitala G, Ganau A, Saba PS, Decandia F, Scuteri A, Marongiu M *et al*: **Serum free thyroxine levels are positively associated with arterial stiffness in the SardiNIA study.** *Clin Endocrinol (Oxf)* 2015, **82**: 592-597
112. Gao CX, Yang B, Guo Q, Wei LH, Tian LM: **High thyroid-stimulating hormone level is associated with the risk of developing atherosclerosis in subclinical hypothyroidism.** *Horm Metab Res* 2015, **47**: 220-224
113. Van Deerlin VM, Sleiman PM, Martinez-Lage M, Chen-Plotkin A, Wang LS, Graff-Radford NR, Dickson DW, Rademakers R, Boeve BF, Grossman M *et al*: **Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions.** *Nat Genet* 2010, **42**: 234-239
114. Yu L, De Jager PL, Yang J, Trojanowski JQ, Bennett DA, Schneider JA: **The TMEM106B locus and TDP-43 pathology in older persons without FTL.** *Neurology* 2015, **84**: 927-934

115. Neumann M, Sampathu DM, Kwong LK, Truax AC, Micsenyi MC, Chou TT, Bruce J, Schuck T, Grossman M, Clark CM *et al*: **Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis.** *Science* 2006, **314**: 130-133
116. Nicholson AM, Finch NA, Wojtas A, Baker MC, Perkerson RB, 3rd, Castanedes-Casey M, Rousseau L, Benussi L, Binetti G, Ghidoni R *et al*: **TMEM106B p.T185S regulates TMEM106B protein levels: implications for frontotemporal dementia.** *J Neurochem* 2013, **126**: 781-791
117. Ighodaro ET, Jicha GA, Schmitt FA, Neltner JH, Abner EL, Kryscio RJ, Smith CD, Duplessis T, Anderson S, Patel E *et al*: **Hippocampal Sclerosis of Aging Can Be Segmental: Two Cases and Review of the Literature.** *J Neuropathol Exp Neurol* 2015, **74**: 642-652
118. Wu RS, Marx SO: **The BK potassium channel in the vascular smooth muscle and kidney: alpha- and beta-subunits.** *Kidney Int* 2010, **78**: 963-974
119. Wallner M, Meera P, Toro L: **Molecular basis of fast inactivation in voltage and Ca²⁺-activated K⁺ channels: a transmembrane beta-subunit homolog.** *Proc Natl Acad Sci U S A* 1999, **96**: 4137-4142
120. Hicks GA, Marrion NV: **Ca²⁺-dependent inactivation of large conductance Ca²⁺-activated K⁺ (BK) channels in rat hippocampal neurones produced by pore block from an associated particle.** *J Physiol* 1998, **508 (Pt 3)**: 721-734
121. Civelek M, Lusk AJ: **Systems genetics approaches to understand complex traits.** *Nat Rev Genet* 2014, **15**: 34-48
122. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C *et al*: **An integrative genomics approach to infer causal associations between gene expression and disease.** *Nat Genet* 2005, **37**: 710-717
123. MacLellan WR, Wang Y, Lusk AJ: **Systems-based approaches to cardiovascular disease.** *Nat Rev Cardiol* 2012, **9**: 172-184
124. Katsumata Y, Nelson PT, Ellingson SR, Fardo DW: **Gene-based association study of genes linked to hippocampal sclerosis of aging neuropathology: GRN, TMEM106B, ABCC9, and KCNMB2.** *Neurobiol Aging* 2017, **53**: 193 e117-193 e125
125. Liang WS, Reiman EM, Valla J, Dunckley T, Beach TG, Grover A, Niedzielko TL, Schneider LE, Mastroeni D, Caselli R *et al*: **Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons.** *Proc Natl Acad Sci U S A* 2008, **105**: 4441-4446
126. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31**: 3812-3814
127. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** *Nat Methods* 2010, **7**: 248-249
128. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Ser A Stat Soc* 1995, **57**: 289-300
129. Khera R, Das N: **Complement Receptor 1: disease associations and therapeutic implications.** *Mol Immunol* 2009, **46**: 761-772

130. Rogers J, Li R, Mastroeni D, Grover A, Leonard B, Ahern G, Cao P, Kolody H, Vedders L, Kolb WP *et al*: **Peripheral clearance of amyloid beta peptide by complement C3-dependent adherence to erythrocytes.** *Neurobiol Aging* 2006, **27**: 1733-1739
131. Bralten J, Franke B, Arias-Vasquez A, Heister A, Brunner HG, Fernandez G, Rijpkema M: **CR1 genotype is associated with entorhinal cortex volume in young healthy adults.** *Neurobiol Aging* 2011, **32**: 2106 e2107-2111
132. Bennett ML, Bennett FC, Liddel SA, Ajami B, Zamanian JL, Fernhoff NB, Mulinyawe SB, Bohlen CJ, Adil A, Tucker A *et al*: **New tools for studying microglia in the mouse and human CNS.** *Proc Natl Acad Sci U S A* 2016, **113**: E1738-1746
133. Ma J, Yu JT, Tan L: **MS4A Cluster in Alzheimer's Disease.** *Mol Neurobiol* 2015, **51**: 1240-1248
134. Marambaud P, Dreses-Werringloer U, Vingtdeux V: **Calcium signaling in neurodegeneration.** *Mol Neurodegener* 2009, **4**: 20
135. LaFerla FM: **Calcium dyshomeostasis and intracellular signalling in Alzheimer's disease.** *Nat Rev Neurosci* 2002, **3**: 862-872
136. Desikan RS, Thompson WK, Holland D, Hess CP, Brewer JB, Zetterberg H, Blennow K, Andreassen OA, McEvoy LK, Hyman BT *et al*: **The role of clusterin in amyloid-beta-associated neurodegeneration.** *JAMA Neurol* 2014, **71**: 180-187
137. Yu JT, Tan L: **The role of clusterin in Alzheimer's disease: pathways, pathogenesis, and therapy.** *Mol Neurobiol* 2012, **45**: 314-326
138. Shannan B, Seifert M, Boothman DA, Tilgen W, Reichrath J: **Clusterin and DNA repair: a new function in cancer for a key player in apoptosis and cell cycle control.** *J Mol Histol* 2006, **37**: 183-188
139. Ling IF, Bhongsatiern J, Simpson JF, Fardo DW, Estus S: **Genetics of clusterin isoform expression and Alzheimer's disease risk.** *PLoS One* 2012, **7**: e33923
140. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A: **Widespread monoallelic expression on human autosomes.** *Science* 2007, **318**: 1136-1140
141. Chess A: **Mechanisms and consequences of widespread random monoallelic expression.** *Nat Rev Genet* 2012, **13**: 421-428
142. Yu L, Chibnik LB, Srivastava GP, Pochet N, Yang J, Xu J, Kozubek J, Obholzer N, Leurgans SE, Schneider JA *et al*: **Association of Brain DNA methylation in SORL1, ABCA7, HLA-DRB5, SLC24A4, and BIN1 with pathological diagnosis of Alzheimer disease.** *JAMA Neurol* 2015, **72**: 15-24
143. Kim WS, Guillemin GJ, Glaros EN, Lim CK, Garner B: **Quantitation of ATP-binding cassette subfamily-A transporter gene expression in primary human brain cells.** *Neuroreport* 2006, **17**: 891-896
144. Chan SL, Kim WS, Kwok JB, Hill AF, Cappai R, Rye KA, Garner B: **ATP-binding cassette transporter A7 regulates processing of amyloid precursor protein in vitro.** *J Neurochem* 2008, **106**: 793-804
145. Kim WS, Li H, Ruberu K, Chan S, Elliott DA, Low JK, Cheng D, Karl T, Garner B: **Deletion of Abca7 increases cerebral amyloid-beta accumulation in the J20 mouse model of Alzheimer's disease.** *J Neurosci* 2013, **33**: 4387-4394

146. Gautel M, Zuffardi O, Freiburg A, Labeit S: **Phosphorylation switches specific for the cardiac isoform of myosin binding protein-C: a modulator of cardiac contraction?** *EMBO J* 1995, **14**: 1952-1960
147. Ng D, Pitcher GM, Szilard RK, Sertie A, Kanisek M, Clapcote SJ, Lipina T, Kalia LV, Joo D, McKerlie C *et al*: **Neto1 is a novel CUB-domain NMDA receptor-interacting protein required for synaptic plasticity and learning.** *PLoS Biol* 2009, **7**: e41
148. Shin SM, Zhang N, Hansen J, Gerges NZ, Pak DT, Sheng M, Lee SH: **GKAP orchestrates activity-dependent postsynaptic protein remodeling and homeostatic scaling.** *Nat Neurosci* 2012, **15**: 1655-1666
149. Voglis G, Tavernarakis N: **The role of synaptic ion channels in synaptic plasticity.** *EMBO Rep* 2006, **7**: 1104-1110
150. Fusi S, Drew PJ, Abbott LF: **Cascade models of synaptically stored memories.** *Neuron* 2005, **45**: 599-611
151. Kandel ER: **The molecular biology of memory storage: a dialogue between genes and synapses.** *Science* 2001, **294**: 1030-1038
152. Maher B: **Personal genomes: The case of the missing heritability.** *Nature* 2008, **456**: 18-21
153. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH: **Missing heritability and strategies for finding the underlying causes of complex disease.** *Nat Rev Genet* 2010, **11**: 446-450
154. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A *et al*: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**: 747-753
155. Jonsson T, Stefansson H, Steinberg S, Jonsdottir I, Jonsson PV, Snaedal J, Bjornsson S, Huttenlocher J, Levey AI, Lah JJ *et al*: **Variant of TREM2 associated with the risk of Alzheimer's disease.** *N Engl J Med* 2013, **368**: 107-116
156. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, Cruchaga C, Sassi C, Kauwe JS, Younkin S *et al*: **TREM2 variants in Alzheimer's disease.** *N Engl J Med* 2013, **368**: 117-127
157. Kulldorff M: **Scan Statistics for Geographical Disease Surveillance: An Overview.** In: *Spatial and Syndromic Surveillance for Public Health*. Chichester, UK: John Wiley & Sons, Ltd; 2005.
158. Jin SC, Benitez BA, Karch CM, Cooper B, Skorupa T, Carrell D, Norton JB, Hsu S, Harari O, Cai Y *et al*: **Coding variants in TREM2 increase risk for Alzheimer's disease.** *Hum Mol Genet* 2014, **23**: 5838-5846
159. Finelli D, Rollinson S, Harris J, Jones M, Richardson A, Gerhard A, Snowden J, Mann D, Pickering-Brown S: **TREM2 analysis and increased risk of Alzheimer's disease.** *Neurobiol Aging* 2015, **36**: 546 e549-513
160. Slattery CF, Beck JA, Harper L, Adamson G, Abdi Z, Uphill J, Campbell T, Druyeh R, Mahoney CJ, Rohrer JD *et al*: **R47H TREM2 variant increases risk of typical early-onset Alzheimer's disease but not of prion or frontotemporal dementia.** *Alzheimers Dement* 2014, **10**: 602-608 e604

161. Bouchon A, Hernandez-Munain C, Cella M, Colonna M: **A DAP12-mediated pathway regulates expression of CC chemokine receptor 7 and maturation of human dendritic cells.** *J Exp Med* 2001, **194**: 1111-1122
162. Bouchon A, Dietrich J, Colonna M: **Cutting edge: inflammatory responses can be triggered by TREM-1, a novel receptor expressed on neutrophils and monocytes.** *J Immunol* 2000, **164**: 4991-4995
163. Gao L, Jiang T, Yao X, Yu L, Yang X, Li Y: **TREM2 and the Progression of Alzheimer's Disease.** *Curr Neurovasc Res* 2017, **14**: 177-183
164. Benitez BA, Cooper B, Pastor P, Jin SC, Lorenzo E, Cervantes S, Cruchaga C: **TREM2 is associated with the risk of Alzheimer's disease in Spanish population.** *Neurobiol Aging* 2013, **34**: 1711 e1715-1717
165. Pottier C, Wallon D, Rousseau S, Rovelet-Lecrux A, Richard AC, Rollin-Sillaire A, Frebourg T, Campion D, Hannequin D: **TREM2 R47H variant as a risk factor for early-onset Alzheimer's disease.** *J Alzheimers Dis* 2013, **35**: 45-49
166. Cuyvers E, Bettens K, Philtjens S, Van Langenhove T, Gijssels I, van der Zee J, Engelborghs S, Vandebulcke M, Van Dongen J, Geerts N *et al*: **Investigating the role of rare heterozygous TREM2 variants in Alzheimer's disease and frontotemporal dementia.** *Neurobiol Aging* 2014, **35**: 726 e711-729
167. Ruiz A, Dols-Icardo O, Bullido MJ, Pastor P, Rodriguez-Rodriguez E, Lopez de Munain A, de Pancorbo MM, Perez-Tur J, Alvarez V, Antonell A *et al*: **Assessing the role of the TREM2 p.R47H variant as a risk factor for Alzheimer's disease and frontotemporal dementia.** *Neurobiol Aging* 2014, **35**: 444 e441-444
168. Jiao B, Liu X, Tang B, Hou L, Zhou L, Zhang F, Zhou Y, Guo J, Yan X, Shen L: **Investigation of TREM2, PLD3, and UNC5C variants in patients with Alzheimer's disease from mainland China.** *Neurobiol Aging* 2014, **35**: 2422 e2429-2422 e2411
169. Yu JT, Jiang T, Wang YL, Wang HF, Zhang W, Hu N, Tan L, Sun L, Tan MS, Zhu XC *et al*: **Triggering receptor expressed on myeloid cells 2 variant is rare in late-onset Alzheimer's disease in Han Chinese individuals.** *Neurobiol Aging* 2014, **35**: 937 e931-933
170. Miyashita A, Wen Y, Kitamura N, Matsubara E, Kawarabayashi T, Shoji M, Tomita N, Furukawa K, Arai H, Asada T *et al*: **Lack of genetic association between TREM2 and late-onset Alzheimer's disease in a Japanese population.** *J Alzheimers Dis* 2014, **41**: 1031-1038
171. Jin SC, Carrasquillo MM, Benitez BA, Skorupa T, Carrell D, Patel D, Lincoln S, Krishnan S, Kachadoorian M, Reitz C *et al*: **TREM2 is associated with increased risk for Alzheimer's disease in African Americans.** *Mol Neurodegener* 2015, **10**: 19
172. Jiang T, Tan L, Chen Q, Tan MS, Zhou JS, Zhu XC, Lu H, Wang HF, Zhang YD, Yu JT: **A rare coding variant in TREM2 increases risk for Alzheimer's disease in Han Chinese.** *Neurobiol Aging* 2016, **42**: 217 e211-213
173. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB *et al*: **Analysis of protein-coding genetic variation in 60,706 humans.** *Nature* 2016, **536**: 285-291
174. Vardarajan BN, Ghani M, Kahn A, Sheikh S, Sato C, Barral S, Lee JH, Cheng R, Reitz C, Lantigua R *et al*: **Rare coding mutations identified by sequencing of**

- Alzheimer disease genome-wide association studies loci.** *Ann Neurol* 2015, **78**: 487-498
175. Wetzel-Smith MK, Hunkapiller J, Bhangale TR, Srinivasan K, Maloney JA, Atwal JK, Sa SM, Yaylaoglu MB, Foreman O, Ortmann W *et al*: **A rare mutation in UNC5C predisposes to late-onset Alzheimer's disease and increases neuronal cell death.** *Nat Med* 2014, **20**: 1452-1457
 176. Sims R, van der Lee SJ, Naj AC, Bellenguez C, Badarinarayan N, Jakobsdottir J, Kunkle BW, Boland A, Raybould R, Bis JC *et al*: **Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease.** *Nat Genet* 2017, **49**: 1373-1384
 177. Logue MW, Schu M, Vardarajan BN, Farrell J, Bennett DA, Buxbaum JD, Byrd GS, Ertekin-Taner N, Evans D, Foroud T *et al*: **Two rare AKAP9 variants are associated with Alzheimer's disease in African Americans.** *Alzheimers Dement* 2014, **10**: 609-618 e611
 178. Vardarajan BN, Zhang Y, Lee JH, Cheng R, Bohm C, Ghani M, Reitz C, Reyes-Dumeyer D, Shen Y, Rogaeva E *et al*: **Coding mutations in SORL1 and Alzheimer disease.** *Ann Neurol* 2015, **77**: 215-227
 179. Jakobsdottir J, van der Lee SJ, Bis JC, Chouraki V, Li-Kroeger D, Yamamoto S, Grove ML, Naj A, Vronskaya M, Salazar JL *et al*: **Rare Functional Variant in TM2D3 is Associated with Late-Onset Alzheimer's Disease.** *PLoS Genet* 2016, **12**: e1006327
 180. Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, Harari O, Norton J, Budde J, Bertelsen S *et al*: **Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease.** *Nature* 2014, **505**: 550-554
 181. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C, Milosavljevic A, Gibbs RA, Yu F: **An integrative variant analysis suite for whole exome next-generation sequencing data.** *BMC Bioinformatics* 2012, **13**: 8
 182. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20**: 1297-1303
 183. Yao Z, Petschnigg J, Ketteler R, Stagljar I: **Application guide for omics approaches to cell signaling.** *Nat Chem Biol* 2015, **11**: 387-397
 184. Hasin Y, Seldin M, Lusic A: **Multi-omics approaches to disease.** *Genome Biol* 2017, **18**: 83
 185. Zavar C, Plant TD, Schirra C, Konnerth A, Neumcke B: **Cell-type specific expression of ATP-sensitive potassium channels in the rat hippocampus.** *J Physiol* 1999, **514 (Pt 2)**: 327-341
 186. Shi NQ, Ye B, Makielski JC: **Function and distribution of the SUR isoforms and splice variants.** *J Mol Cell Cardiol* 2005, **39**: 51-60
 187. Nichols CG, Singh GK, Grange DK: **KATP channels and cardiovascular disease: suddenly a syndrome.** *Circ Res* 2013, **112**: 1059-1072
 188. Palacios R, Gazave E, Goni J, Piedrafita G, Fernando O, Navarro A, Villoslada P: **Allele-specific gene expression is widespread across the genome and biological processes.** *PLoS One* 2009, **4**: e4150

VITA

EDUCATION

Ph.D. Social Medicine, Hokkaido University, Hokkaido, Japan December 2005
M.S. Physics, Kanazawa University, Ishikawa, Japan March 1997
B.S. Physics, Kanazawa University, Ishikawa, Japan March 1995

PROFESSIONAL EXPERIENCE

Graduate Research Assistant September 2013 – December 2017
Department of Biostatistics, University of Kentucky, KY, USA

Visiting Researcher September 2011 – August 2013
Layton Aging and Alzheimer's Disease Center,
Oregon Health and Science University, OR, USA

Assistant Professor June 2011 – September 2011
Center of Residency and Fellowship Program, University Hospital,
Faculty of Medicine, University of the Ryukyus, Okinawa, Japan

Assistant Professor June 2006 – June 2011
Department of Public Health and Hygiene, Graduate School of Medicine,
University of the Ryukyus, Okinawa, Japan

Visiting Researcher October 2005 – June 2006
Research Center for Zoonosis Control,
Hokkaido University, Hokkaido, Japan

Graduate Research Assistant April 2001 – March 2004
Department of Health for Senior Citizens, Graduate School of Medicine,
Hokkaido University, Hokkaido, Japan

HONORS AND AWARDS

McCullers Scholar 2017
Mu Sigma Rho Statistical Honor Society, Elected 2015
Best Performance Epidemiology and Biostatistics Ph.D. Comprehensive Student Exam,
Department of Biostatistics, University of Kentucky 2015

Student Travel Award,	
Genetic Analysis Workshop (GAW) 19	2014
Research Encouraging Award,	
Japan Personal Computer Application Technology Society	2008

PUBLICATIONS

1. Gall J, Chen J, **Katsumata Y**, Fardo DW, Wang WX, Artiushin S, Price D, Anderson S, Patel E, Zhu H, Nelson PT. Detergent insoluble proteins and inclusion body-like structures immunoreactive for PRKDC/DNA-PK/DNA-PKcs, FTL, NNT, and AIFM1 in the amygdala of cognitively impaired elderly persons. *Journal of Neuropathology and Experimental Neurology*. 2017 (in press).
2. Marottoli FM, **Katsumata Y**, Koster KP, Thomas R, Fardo DW, Tai LM. Peripheral inflammation, *APOE4* and amyloid-beta interact to induce cognitive and cerebrovascular dysfunction. *ASN Neuro*. 2017 (in press).
3. Fardo DW, **Katsumata Y**, Kauwe JS, Deming Y, Harari O, Cruchaga C; Alzheimer's Disease Neuroimaging Initiative., Nelson PT. CSF protein changes associated with hippocampal sclerosis risk gene variants highlight impact of *GRN/PGRN*. *Experimental Gerontology*. 2017; 90: 83-89.
4. **Katsumata Y**, Nelson PT, Ellingson SR, Fardo DW. Gene-based association study of genes linked to hippocampal sclerosis of aging neuropathology: *GRN*, *TMEM106B*, *ABCC9*, and *KCNMB2*. *Neurobiology of Aging*. 2017; 53: 193.e17-193.e25.
5. Ozaki T, **Katsumata Y**, Arai A. The use of psychotropic drugs for behavioral and psychological symptoms of dementia among residents in long-term care facilities in Japan. *Aging & Mental Health*. 2017; 21: 1248-1255.
6. Arai A, Ozaki T, **Katsumata Y**. Behavioral and psychological symptoms of dementia in older residents in Long-Term Care facilities in Japan: A cross-sectional study. *Aging & Mental Health*. 2017; 21: 1099-1105.
7. Ighodaro ET, Abner EL, Fardo DW, Lin AL, **Katsumata Y**, Schmitt FA, Kryscio RJ, Jicha GA, Neltner JH, Monsell SE, Kukull WA, Moser DK, Appiah F, Bachstetter AD, Van Eldik LJ; Alzheimer's Disease Neuroimaging Initiative (ADNI), Nelson PT.

- Risk factors and global cognitive status related to brain arteriolosclerosis in elderly individuals. *Journal of Cerebral Blood Flow & Metabolism*. 2017; 37: 201-216.
8. **Katsumata Y**, Fardo DW. On Combining Family- and Population-based Sequencing Data. *BMC Proceedings*. 2016; 10 (Suppl 7): 175-179.
 9. Nelson PT, **Katsumata Y**, Nho K, Artiushin SC, Jicha GA, Wang WX, Abner EL, Saykin AJ, Kukull WA; Alzheimer's Disease Neuroimaging Initiative (ADNI), Fardo DW. Genomics and CSF analyses implicate thyroid hormone in hippocampal sclerosis of aging. *Acta Neuropathologica*. 2016; 132: 841-858.
 10. Webber KH, Casey EM, Mayes L, **Katsumata Y**, Mellin L. A comparison of a behavioral weight loss program to a stress management program: A pilot randomized controlled trial. *Nutrition*. 2016; 32: 904-909.
 11. Allen GI, Amoroso N, Anghel C, Balagurusamy V, Bare CJ, Beaton D, Bellotti R, Bennett DA, Boehme KL, Boutros PC, Caberlotto L, Caloian C, Campbell F, Chaibub Neto E, Chang YC, Chen B, Chen CY, Chien TY, Clark T, Das S, Davatzikos C, Deng J, Dillenberger D, Dobson RJ, Dong Q, Doshi J, Duma D, Errico R, Erus G, Everett E, Fardo DW, Friend SH, Fröhlich H, Gan J, St George-Hyslop P, Ghosh SS, Glaab E, Green RC, Guan Y, Hong MY, Huang C, Hwang J, Ibrahim J, Inglese P, Iyappan A, Jiang Q, **Katsumata Y**, Kauwe JS, Klein A, Kong D, Krause R, Lalonde E, Lauria M, Lee E, Lin X, Liu Z, Livingstone J, Logsdon BA, Lovestone S, Ma TW, Malhotra A, Mangravite LM, Maxwell TJ, Merrill E, Nagorski J, Namasivayam A, Narayan M, Naz M, Newhouse SJ, Norman TC, Nurtdinov RN, Oyang YJ, Pawitan Y, Peng S, Peters MA, Piccolo SR, Praveen P, Priami C, Sabelnykova VY, Senger P, Shen X, Simmons A, Sotiras A, Stolovitzky G, Tangaro S, Tateo A, Tung YA, Tustison NJ, Varol E, Vradenburg G, Weiner MW, Xiao G, Xie L, Xie Y, Xu J, Yang H, Zhan X, Zhou Y, Zhu F, Zhu H, Zhu S; Alzheimer's Disease Neuroimaging Initiative. Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. *Alzheimer's & Dementia*. 2016; 12: 645-653.
 12. Aono J, Ikeda S, **Katsumata Y**, Higashi H, Ohshima K, Ishibashi K, Matsuoka H, Watanabe K, Hamada M. Correlation between plaque vulnerability of aorta and coronary artery: an evaluation of plaque activity by direct visualization with angioscopy. *International Journal of Cardiovascular Imaging*. 2015; 31: 1107-1114.

13. **Katsumata Y**, Mathews M, Abner EL, Jicha GA, Caban-Holt A, Smith CD, Nelson PT, Kryscio RJ, Schmitt FA, Fardo DW. Assessing discriminant ability, reliability, and comparability of multiple short forms of the Boston Naming Test in an Alzheimer's Disease Center cohort. *Dementia and Geriatric Cognitive Disorders*. 2015; 39: 215-227.
14. Dodge HH, **Katsumata Y**, Zhu J, Mattek N, Bowman M, Gregor M, Wild K, Kaye JA. Characteristics associated with willingness to participate in a randomized controlled behavioral clinical trial using home-based personal computers and a webcam. *Trials*. 2014; 23: 508.
15. Downer B, Estus S, **Katsumata Y**, Fardo DW. Longitudinal trajectories of cholesterol from midlife through late life according to apolipoprotein E allele status. *International Journal of Environmental Research and Public Health*. 2014; 11: 10663-10693.
16. Obayashi Y, Arai A, Liu Y, **Katsumata Y**, Kono K, Uchida H, Masaki M, Aoki M, Takahashi A, Yamada T, Kamei C, Sugiura S, Kaeriyama M, Tamahiro H. A survey on awareness of the Nitobe College among the first-year students of Hokkaido University, 2013. *Journal of Higher Education and Lifelong Learning*. 2014; 21: 61-68.
17. **Katsumata Y**, Todoriki H, Higashiuesato Y, Yasura S, Ohya Y, Craig Willcox D, Dodge HH. Very old adults with better memory function have higher low-density lipoprotein cholesterol levels and lower triglyceride to high-density lipoprotein cholesterol ratios: KOCO A project. *Journal of Alzheimer's Disease*. 2013; 34: 273-279.
18. **Katsumata Y**, Todoriki H, Higashiuesato Y, Yasura S, Craig Willcox D, Ohya Y, Willcox BJ, Dodge HH. Metabolic syndrome and cognitive decline among the oldest old in Okinawa: In search of a mechanism. The KOCO A project. *Journals of Gerontology Series A-Biological Sciences and Medical Sciences*. 2012; 67: 126-134.
19. **Katsumata Y**, Arai A, Ishida K, Tomimori M, Lee RB, Tamashiro H. Which categories of social and lifestyle activities moderate the association between negative life events and depressive symptoms among community-dwelling older adults in Japan? *International Psychogeriatrics*. 2012; 24: 307-315.

20. **Katsumata Y**, Todoriki H, Yasura S, Dodge HH. Timed up and go test predicts cognitive decline among healthy adults aged 80 and older in Okinawa: KOCOA project. *Journal of the American Geriatrics Society*. 2011; 59: 2188-2189.
21. **Katsumata Y**, Arai A, Tomimori M, Ishida K, Lee RB, Tamashiro H. Fear of falling and falls self-efficacy and their relationship to higher-level competence among community-dwelling senior men and women in Japan. *Geriatrics and Gerontology International*. 2011; 11: 282-289.
22. Dodge HH, **Katsumata Y**, Todoriki H, Yasura S, Willcox DC, Bowman GL, Willcox B, Leonard S, Clemons A, Oken BS, Kaye JA, Traber MG. Comparisons of plasma/serum micronutrients between Okinawan and Oregonian Elderly: A Pilot Project. *Journals of Gerontology Series A-Biological Sciences and Medical Sciences*. 2010; 65: 1060-1067.
23. Arai A, Ishida K, Tomimori M, **Katsumata Y**, Grove JS, Tamashiro H. Association between lifestyle activity and depressed mood among home-dwelling older people: A community-based study in Japan. *Aging & Mental Health*. 2007; 1: 547-555.
24. Zheng KC, Todoriki H, **Katsumata Y**, Gao WM, Keohavong P. Analysis of p53 and K-ras mutations in patients with chronic bronchitis using laser capture microdissection microscope and mutation detection. *Ningen Dock*. 2007; 21: 66-68.
25. **Katsumata Y**, Arai A, Tamashiro H. Contribution of falling and being homebound status to subsequent functional changes among the Japanese elderly living in a community. *Archives of Gerontology and Geriatrics*. 2007; 45: 9-18.
26. **Katsumata Y**, Arai A, Tamashiro H. Nonlinear association of higher-level functional capacity with the incidence of falls in Japan. *American Journal of Physical Medicine and Rehabilitation*. 2006; 85: 688-693.
27. **Katsumata Y**, Arai A, Ishida K, Tomimori M, Denda K, Tamashiro H. Gender differences in the contributions of the risk factors to depressive symptoms among the elderly persons dwelling in a community, Japan. *International Journal of Geriatric Psychiatry*. 2005; 20: 1084-1089.
28. Arai A, **Katsumata Y**, Konno K, Tamashiro H. Sociodemographic Factors associated with Incidence of Dementia in Senior Citizens of a Small Town in Japan. *Care Management Journals*. 2004; 5: 159-165.

29. Konno K, **Katsumata Y**, Arai A, Tamashiro H. Functional status and active life expectancy among senior citizens in a small town in Japan. *Archives of Gerontology and Geriatrics*. 2004; 38: 153-166.
30. **Katsumata Y**, Arai A, Konno K, Tamashiro H. Difference in terminology for referencing aging between generations. *Journal of Public Health Practice*. 2004; 68: 578-582 (in Japanese).
31. Tamashiro H, **Katsumata Y**, Arai A, Konno K, Nakazawa H, Usami K, Rambelli R. Past, present, and future HIV/AIDS. *Japanese Journal of Urological Surgery*. 2003; 67: 637-641 (in Japanese).
32. Arai A, **Katsumata Y**, Konno K, Tamashiro H. Status quo in barrier-free environment for wheelchair users in Sapporo. *Journal of Public Health Practice*. 2003; 67: 637-641 (in Japanese).
33. Arai A, **Katsumata Y**, Konno K, Ohta K, Ohtomo K, Kimura S, Takahashi M, Dobata T, Machida K. From a view of the wheelchair users--a training report of the students on barrier-free environment in Sapporo. *Hokkaido Journal of Medical Science*. 2002; 77: 107-110 (in Japanese).
34. **Katsumata Y**, Arai A, Kishi R, Tamashiro H. The latest information of the schools of public health in USA: with reference to a proposed school of public health in Japan at the 21st century. *Japanese Journal of Public Health*. 2001; 48: 298-303 (in Japanese).
35. Kagohashi T, Takarada T, Hasegawa H, **Katsumata Y**, Kuga M, Okamoto H, Kaneko H, Ishihara Y. Electrical resistivity and thermoelectric power of a quasi-one-dimensional Nb₃Te₄ single crystal inserted with mercury: Hg_xNb₃Te₄. *Journal of Physics: Condensed Matter*. 1999; 11: 6373-6384.