Summer 9-1-2014

Identification of Functional Variants in Alzheimer's Disease-Associated Genes

Sheng Chih Jin
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology & Biomedical Sciences
Human and Statistical Genetics

Dissertation Examination Committee:
Alison Goate, Chair
Donald Conrad
Carlos Cruchaga
David Holtzman
John Rice
Ting Wang

Identification of Functional Variants in Alzheimer's Disease-Associated Genes

by

Sheng Chih Jin

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2014

St. Louis, Missouri

TABLE OF CONTENTS

List of Figures and Tables iv

Acknowledgements vi

Abstract of the Dissertation vii

Chapter 1: BACKGROUND AND SIGNIFICANCE

1.1 OVERVIEW OF THIS DISSERTATION 2

1.2 ALZHEIMER'S DISEASE 4

1.3 AMYLOID BETA PLAQUES 5

1.4 NEUROFIBRILLARY TANGLES 5

1.5 ALZHEIMER'S GENETICS 7

1.6 BIOMARKERS FOR AD 12

1.7 ENDOPHENOTYPE GENETICS 12

Chapter 2: TARGETED RE-SEQUENCING OF ALZHEIMER'S DISEASE ASSOCIATED GENES

2.1 ABSTRACT 15

2.2 INTRODUCTION 16

2.3 MATERIALS AND METHODS 17

2.4 RESULTS 23

2.5 DISCUSSION 34

2.6 SUPPLEMENTAL TABLES 39

2.7 SUPPLEMENTAL FIGURES 44

Chapter 3: USE OF CEREBROSPINAL FLUID AS ENDOPHENOTYPES TO FINE-MAP ALZHEIMER'S DISEASE ASSOCIATED GENES

3.1 ABSTRACT 49

3.2 INTRODUCTION 50

3.3 MATERIALS AND METHODS 51

3.4 RESULTS 59

3.5 DISCUSSION 75

3.6 SUPPLEMENTAL TABLES 79

3.7 SUPPLEMENTAL FIGURES 87

Chapter 4: FUNCTIONAL STUDIES: CIS- ACTING EXPRESSION QUANTITATIVE TRAIT LOCI ANALYSES & *TREM2* CELL SURFACE EXPRESSION STUDIES

4.1 ABSTRACT 126

4.2 INTRODUCTION 127

4.3 MATERIALS AND METHODS 128

4.4 RESULTS 135

4.5 DISCUSSION 141

4.6 SUPPLEMENTAL TABLES 147

4.7 SUPPLEMENTAL FIGURES 148

Chapter 5: CONCLUSIONS AND FUTURE DIRECTIONS

5.1 STATE OF ALZHEIMER'S GENETICS PRIOR TO THIS WORK 189

5.2 DISSERTATION WORK 193

5.3 FUTURE DIRECTIONS 196

REFERENCES 203

CURRICULUM VITAE

List of Tables and Figures

Chapter 1: BACKGROUND AND SIGNIFICANCE

Figure 1: Currently known genes or loci affecting Alzheimer's risk 3

Figure 2: Schematic representation of APP 6

Figure 3: Fibrillar neuritic plaques and non-fibrillar diffuse plaques 6

Chapter 2: TARGETED RE-SEQUENCING OF ALZHEIMER'S DISEASE ASSOCIATED GENES

Table 1: Demographic characteristics of re-sequenced European Americans 18

Table 2: Demographic characteristics of re-sequenced African Americans 18

Table 3: Sequence variants found in *PLD3* in the NIA-LOAD, Knight-ADRC and NIA-UK 25

Table 4: Rare *TREM2*-variant association in sequenced samples of EA descent 28

Table 5: Rare *TREM2* variants found in African Americans 29

Table 6: Segregation of rare variants in available family members 31

Table 7: Confirmed variants in GWAS-identified genes 32

Figure 1: Schematic of the pooled-DNA sequencing technique 20

Figure 2: *PLD3* gene-based analysis 24

Figure 3: Schematic representation of PLD3 and the alternative position of the *PLD3* variants 26

Figure 4: Schematic representation of protein structure for TREM2 and for the soluble form of TREM2, location of variants, protein conservation of the mutated positions, and the results of alternative splicing assays. 30

Chapter 3: USE OF CEREBROSPINAL FLUID AS ENDOPHENOTYPES TO FINE-MAP ALZHEIMER'S DISEASE ASSOCIATED GENES

Table 1: Summary characteristics of the CSF study participants 54

Table 2: Summary statistics for most significant SNP in IGAP top loci 63

Table 3: Most significant association with CSF $A\beta_{42}$ level in each IGAP locus 64

Table 4: Most significant association with CSF $ptau_{181}$ level in each IGAP locus 66

Table 5: Results of set-based analyses for each GWAS-identified locus 67

Table 6: Detailed descriptions for SNPs with putative regulatory function 70

Table 7: *TREML2* variants identified by exome-sequencing 73

Figure 1: Regional plots for *CELF1*, *EPHA1*, and *FERMT2* fine-mapping regions 65

Figure 2: Forest plots for rs9381040, rs3747742, and rs75932628 74

Figure 3: Conditional analyses for *TREM2* and *TREML2* variants with CSF $ptau_{181}$ levels 74

Chapter 4: FUNCTIONAL STUDIES: CIS- ACTING EXPRESSION QUANTITATIVE TRAIT LOCI ANALYSES & *TREM2* CELL SURFACE EXPRESSION STUDIES

Table 1: Summary counts of number of individuals, probes, and SNPs tested per brain region 132

Table 2: SNP-Transcript pairs with a suggestive or significant association in GSE15745 and GSE36192 139

Table 3: Association between AD disease status and *C1QTNF4* expression levels in the joint analysis in GSE5281 140

Table 4: Association between AD disease status and *C1QTNF4* expression levels in six brain regions in GSE5281 140

Table 5: Transcript-SNP Pairs with a suggestive or significant association based on the GSE15222 143

Table 6: Association between AD disease status and *C1QTNF4* expression levels in four brain regions in GSE15222 143

Figure 1: Schematic representation of pMXs-hTREM2-IRES-hDAP12 retroviral vector 130

Figure 2: Flow cytometry of NFAT43.1 reporter cells from hTREM2 WT and variants 136

Figure 3: Cis association for *C1QTNF4* expression levels in GSE15745 and GSE36192 138

Figure 4: Linkage disequilibrium analyses for potential eQTLs, CSF top SNP, and IGAP top SN in the *CELF1* region 139

Figure 5: Cis association for *C1QTNF4* expression levels in FCTX, CRBLM, TCTC, and parietal cortex in GSE15222 142

Acknowledgements

First and foremost, I would like to thank my mentor and my role model, Dr. Alison Goate. From day one, she has always been there for me, providing countless hours of support and guidance. Without her, I could not have finished my PhD nor have accomplished so much over the years. Working with Dr. Goate has taught me to grow, personally and professionally, as an independent scientist, and I cannot tell you how grateful I am for the mentorship and bonds that have blossomed throughout my educational journey. You are an amazing scientist, a visionary, and an epitome of a great educator. I simply cannot describe how much I admire you.

I would also like to express my deepest appreciation to my co-mentor, Dr. Carlos Cruchaga, for his valuable and constructive suggestions throughout the planning and development of my thesis research. Without Carlos, no Sequencing works would have been done and no creative ideas would have been given. Additionally, Carlos treated me like his peer and friend, and I respect him like my senior brother.

This thesis project would not have been possible without the support of other members of the Goate and Cruchaga labs. In particular, Celeste Karch, Bruno Benitez, Simon Hsu, and Breanna Cooper, without whom genotyping and functional studies would not have been finished. I would like to thank Dave Carrell, Tara Skorupa, John Budde, Joanne Norton, and Kevin Mayo for their help in preparing samples for my studies and Yefei Cai and Sarah Bertelsen for helping me analyze data over the past few years. I would like to thank my thesis committee members, Dr. David Holtzman, Dr. John Rice, Dr. Donald Conrad, and Dr. Ting Wang for their continued support and guidance over the years. Moreover, much of the TREM2 functional work would have been difficult without the help of Dr. Marco Colonna and Yaming Wang. I absolutely cannot thank them enough.

I would very much like to thank my parents for their love and for their full support in my interest of becoming a scientist. Even though I know you had different expectation of me, thank you both for being so supportive and for letting me do what I love.

Finally, to my dear and loving wife, Pei Wen Hung: my deepest gratitude. Your company and encouragement when I am frustrated and hopeless are much appreciated. I am so grateful to have you stand beside me through hardships and joy. Thank you for everything.

ABSTRACT OF THE DISSERTATION

Identification of Functional Variants in Alzheimer's Disease-Associated Genes

by

Sheng Chih Jin

Doctor of Philosophy in Biology and Biomedical Sciences
(Human and Statistical Genetics)

Washington University in St. Louis, 2014

Professor Alison Goate, Chairperson

Alzheimer's disease (AD) is the most common form of dementia affecting the health of more than 5 million Americans in 2013. Understanding how genetic variants contribute to AD is important to develop effective therapeutics for delaying and eventually curing the disease. Recent sequencing studies have identified rare variants p.V232M in *PLD3* and p.R47H in *TREM2* significantly associated with AD risk. Additionally, genome-wide association studies (GWAS) uncovered common variants in *ABCA7*, *BIN1*, *CD33*, *CD2AP*, *CLU*, *CR1*, *EPHA1*, *MS4A4A*, and *PICALM* that contribute to AD; however, the most-significant variants in these loci are located within non-coding regions and have no direct functional impact on AD pathogenesis. We hypothesized that if *PLD3* and *TREM2* are truly AD risk genes, they will carry additional functional variants that can substantially affect AD risk. Moreover, we hypothesized that GWAS genes carry additional risk alleles across the frequency spectrum so that common, low frequency or rare variants within these genes may contribute to AD risk. We undertook pooled sequencing of exonic and flanking intronic sequence for the aforementioned genes in 3,730 European Americans (EA) (2,082 AD cases and1,648 controls) and 336 African Americans (AA) (204 AD cases and 132 controls). We found rare variants in *PLD3* and *TREM2* are more frequently seen in cases than in controls of EA descent. Single-variant analyses showed that p.M6R and p.A442A in *PLD3* and p.R62H in *TREM2* are significantly associated with AD risk besides p.V232M in *PLD3* and p.R47H in *TREM2*. We found rare variants in *PLD3* ($P_{SKAT\text{-}O}=1.44\times10^{-11}$) and *TREM2* ($P_{SKAT\text{-}O}=5.37\times10^{-7}$) are genome-wide significantly associated with AD. Additionally, we found a significant association for *PLD3* coding variants with AD risk in AA ($P_{SKAT\text{-}O}=1.40\times10^{-3}$). However, we did not find evidence of association for *TREM2* variants with AD risk in AA. We validated 90 coding variants in GWAS-identified genes and bioinformatic analyses implicated that a large proportion of these variants are functional.

The International Genomics of Alzheimer's Project recently performed meta- and gene-wide analyses and identified 23 loci associated with AD risk, of which 13 were novel. However, these loci's role in affecting the

molecular pathways of AD remains unknown. To determine whether these loci are also associated with cerebrospinal fluid (CSF) amyloid-beta 1-42 ($A\beta_{42}$) and phosphorylated tau_{181} ($ptau_{181}$) levels, we combined CSF biomarker datasets from several studies and performed single-variant, set-based, and conditional analyses for each locus. In the *APOE* locus, rs769449 is genome-wide significantly associated with CSF $A\beta_{42}$ and $ptau_{181}$ levels independently of *APOE*-ε2 and *APOE*-ε4 SNPs and the association for CSF $ptau_{181}$ levels was not driven by Aβ metabolism. We found rs7937331, within the *CELF1* fine-mapping region, tags the same signal as the IGAP top SNP (rs62003531) and is significantly associated with CSF $A\beta_{42}$ levels and AD risk. Additionally, rs62003531, located in the intronic region of *FERMT2*, tags the same association as the IGAP top SNP (rs17125944) and is associated with CSF $A\beta_{42}$ levels. None of the SNPs within the IGAP-identified AD risk loci except the *APOE* locus are significantly associated with CSF $ptau_{181}$ levels after multiple test correction. In investigating the potential regulatory functions associated with IGAP top SNPs and CSF top SNPs, most of GWAS top SNPs have no significant regulatory potential and are unlikely to be functional variants for AD risk. However, RegulomeDB predicts that several proxy SNPs in linkage disequilibrium (LD) with rs7937331 may be cis-acting expression quantitative trait loci (eQTLs) for nearby genes. The IGAP study also identified an intergenic polymorphism near *TREML2* suggestively associated with AD risk; however, due to the study design, it was not possible to uncover the underlying functional variant or to determine whether this observed association was driven by the known AD risk allele, *TREM2* p.R47H, or represented a novel locus. We performed analyses using whole-exome sequencing data, CSF biomarker analyses and meta-analyses to demonstrate that the AD risk association is likely driven by a *TREML2* variant p.S144G (rs3747742) independently of *TREM2* p.R47H risk for AD.

Finally, we sought to functionally characterize the effects of novel *TREM2* variants on TREM2 cell surface transport. We transduced a T cell hybridoma cell line with virus containing *TREM2* wild type (WT) and risk variants and measured TREM2 cell surface expression with a TREM2-specific monoclonal antibody. We found cells expressing p.T66M and p.R136W have a robust effect on TREM2 cell surface expression but cells expressing p.R47H and p.R62H are similar to hTREM2 WT. Additionally, since polymorphisms in the *CELF1* fine-mapping region were implicated to be eQTLs for nearby genes, we performed cis-eQTL analysis for mRNA expression levels in several brain regions using four publicly available datasets to identify genetic determinants of gene expression in human brains. We found several *C1QTNF4*-expression-associated SNPs which tag the same signal are in LD with rs7937331, the top CSF $A\beta_{42}$ SNP in the *CELF1* fine-mapping region. Additionally, we found evidence of

differential expression in the *C1QTNF4* transcript between AD cases and controls in human brains. These findings provide additional evidence that genes involved in the inflammatory response play an important role in AD pathogenesis.

Chapter 1

Background and significance

1.1 OVERVIEW OF THIS DISSERTATION

Alzheimer's disease (AD) is a devastating neurodegenerative disease affecting more than 5 million Americans and costing US $200 billion in 2012. The number of AD patients is projected to quadruple in the next 40 years and is becoming a major public health problem worldwide. Early genetic studies have identified disease-causing mutations in the *amyloid precursor protein* (*APP*)[1], *presenilin* 1 (*PSEN1*)[2,3] and *presenilin 2* (*PSEN2*)[2,4], while polymorphisms in *apolipoprotein E* (*APOE*)[5-7] affect risk for developing late onset AD (LOAD), providing a better understanding of AD pathogenesis. Genome-wide association studies (GWAS), a method that can rapidly investigate millions of "common" polymorphisms across the human genome in thousands of individuals, have identified several novel loci influencing LOAD[8-11], yet genetic markers within these loci only explain a small proportion of the genetic phenotypic heritability (i.e. the proportion of phenotypic differences between individuals that is due to genetic differences). The underlying mechanisms involved in LOAD remain largely unknown. Recent studies which used emerging next-generation sequencing (NGS) technologies have analyzed the entire human genome with large sample sizes to identify "rare" coding variants in the *triggering receptor expressed on myeloid 2* (*TREM2*)[12,13] and *phospholipase D3* (*PLD3*)[14] genes as risk factors for AD. Together, these genetic studies support the paradigm that, both common, low-penetrant and rare, high-penetrant alleles contribute to AD pathogenesis (**Figure 1**).

Our recent work has used next-generation sequencing technology to identify rare variants in *APP*, *PSEN1* and *PSEN2* associated with risk for LOAD[15]. We also found rare variants in *APP* along with the microtubule-associated protein tau (*MAPT)* and progranulin (*GRN)*, two genes associated with frontotemporal dementia, in clinically diagnosed AD patients of hispanic descent[16]. These findings clearly highlight the importance of performing deep re-sequencing studies in AD-associated genes to identify additional novel variants that may be pathogenic [15,16]. On the other hand, our previous work has shown that cerebrospinal fluid (CSF) biomarkers can be effectively used as quantitative traits in genetic analyses to not only identify novel genetic markers but also to generate testable hypotheses regarding underlying mechanisms[17-24]. Additionally, these studies clearly demonstrated that using CSF 42 amino acid fragments of amyloid beta ($A\beta_{42}$; decreased in AD) and tau phosphorylated at threonine 181 (a proxy for hyperphosphorylated tau; $ptau_{181;}$ increased in AD) as quantitative traits has increased power relative to qualitative clinical diagnosis to identify common, functional variants in AD-associated genes and

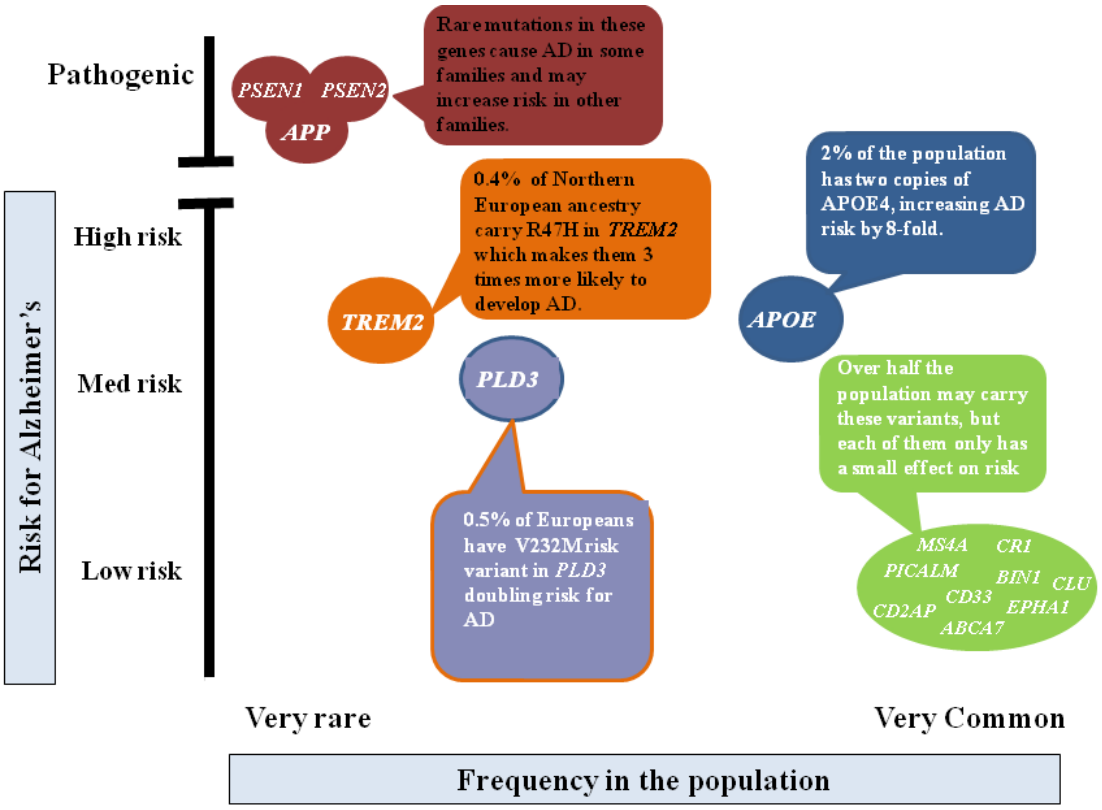


Figure 1. Currently known genes or loci affecting Alzheimer's risk.

to determine whether the identified genetic markers affect AD pathogenesis through an Aβ-dependent or a tau-dependent mechanism[18,19,22,24]. In summary, a combination of next-generation sequencing studies coupled with endophenotype-based association studies can lead to identification of potential functional variants.

The objective of this dissertation is to identify potential functional variants in AD-associated genes and to determine the pathogenic mechanisms associated with these variants. In this chapter, I will give a concise but comprehensive introduction to AD and describe why this dissertation is important to the understanding of AD biology. I will briefly describe what is known about amyloid beta and tau proteins and how they are linked to AD etiology. I will review the known genetic risk factors for AD and their potential roles in AD pathogenesis. I will also summarize a variety of biomarkers that are currently used for measuring Aβ, tau, and neurodegeneration. In Chapter 2, I will describe how I used deep re-sequencing to identify novel functional rare variants in AD-associated genes. Then in Chapter 3, I will describe how I used CSF $A\beta_{42}$ and $ptau_{181}$ levels as quantitative phenotypes to identify novel common variants and further determine whether these potential functional variants affect risk for LOAD

through an Aβ-dependent, a tau-dependent mechanism or another unspecified mechanism (e.g. expression quantitative trait loci [eQTL]). In Chapter 4, I will discuss why I selected variants in *TREM2* for functional studies and how I conducted *in vitro* cell-based assays. Finally in Chapter 5, I will make a final conclusion and discuss future directions and plans. The results of this dissertation could accelerate progress in understanding pathogenic mechanisms associated with AD-implicated genes and may provide molecular targets for treatment.

1.2 ALZHEIMER'S DISEASE

AD -is the most common type of dementia, accounting for approximately 60-70% of all dementia cases[25]. Other types of dementia include Lewy body dementia, Parkinson's disease, frontotemporal dementia, vascular dementia, progressive supranuclear palsy, corticobasal degeneration, Huntington's disease, normal pressure hydrocephalus, Wernicke-Korsakoff syndrome, and Creutzfeldt-Jakob disease. It is estimated that AD affected an estimated 5.2 million Americans in 2013, is the sixth leading cause of death in the USA, and is the only cause of death among the top 10 that cannot be prevented, and is incurable [26]. AD is the third most expensive disease, costing the U.S. government close to $203 billion in 2013[26]. With a global prevalence of 30 million, which is expected to quadruple over the next 40 years, AD is an emergent public health crisis in the 21st century.

AD can be clinically classified into 4 stages: pre-clinical, mild, moderate, and severe. The pathologic changes associated with AD start many years before producing symptoms. Preclinical AD is not benign and affected individuals will develop AD symptoms if they live long enough. An individual with preclinical AD is characterized by cognitive decline. A preclinical AD patient may perform completely normal on physical examination and mental status testing. Normally, there are no changes in judgments or abilities to perform tasks in social or work settings. In the mild AD stage, memory loss continues and other cognitive deficits emerge. The mild AD stage is normally characterized by memory loss, attention deficits, spatial disorientation and executive dysfunction. In the moderate stage, signs and symptoms of AD become more conspicuous and widespread. A patient with mild AD has more obvious difficulty with memory and other cognitive functions and lose the abilities to perform activities of daily living[25]. In the final stage, Alzheimer's dementia, almost all cognitive and physical functional are severely deteriorated. Patients need help with all activities of daily living, and lose the ability to respond to their environment, to conduct a conversation, and to control movement.

AD is a devastating neurodegenerative disorder, with progressive stages of cognitive and functional deterioration. AD is often classified based on the age at onset (AAO) and familial aggregation; familial AD (FAD),

the rare form of the disease, usually has an early AAO (less than 65 years of age), follows an autosomal dominant pattern, and accounts for less than 1% of AD patients. LOAD, accounting for > 99% of AD patients, has a late AAO (greater than 65 years of age) and is clinically and genetically complex. Although FAD and LOAD can generally be distinguished by AAO and familial aggregation, both are pathologically characterized by extensive neuronal loss and the presence of extracellular amyloid-beta (Aβ) plaques composed of Aβ protein and intracellular neurofibrillary tangles composed of hyperphosphorylated tau (ptau) protein[25]. Aβ plaques usually occur up to 15 yrs before clinical onset while neurofibrillary tangles occur closer to the time of onset of symptoms.

1.3 AMYLOID BETA PLAQUES

Aβ plaques are fibrillar aggregates of Aβ peptides surrounded by damaged axons and neurites in the extracellular space of brain. Aβ peptides are 35-43 amino acids resulting from proteolytic cleavage of APP[27] (see **Figure 2**). Aβ plaques can be microscopically classified into neuritic plaques and diffuse plaques[28] (see **Figure 3**). The neuritic plaques, also called senile plaques, are fibrillar extracellular deposits with a β-sheet conformation, which can be visualized through immunofluorescence microscopy using Congo red and Thioflavin S stains[28] (see **Figure 3A**). These neuritic plaques are surrounded by degenerating axons and dendrites and contain hypertrophic gliosis and activated astrocytes and microglia. It has been suggested that this inflammatory reactivity results in brain injury even though other evidence shows that glial cells may also have an opposite effect against brain inflammation[29,30]. Diffuse plaques, which can be detected via immunohistochemical techniques, are aggregated in an α-helical (non-fibrillar) structure and represent the majority of Aβ plaques (see **Figure 3B**). Diffuse plaques are less toxic and can be found in brains of cognitively normal elderly individuals, which may represent the early stage of plaque biology in AD brains[31,32].

1.4 NEUROFIBRILLARY TANGLES

Neurofibrillary tangles (NFTs) are composed of intracellular hyperphosphorylated tau protein, a protein that is expressed in neurons, astrocytes and oligodendrocytes[33-36]. Normal tau protein binds to tubulin and stabilizes microtubules, which is essential to intracellular support. In AD brains, tau becomes hyper-phosphorylated, and subsequently results in dissociation from microtubules. The dissociated tau self aggregates into tangles of paired-helical filaments, which can contribute to neuronal dysfunction and other tauopathies. Tau can spread throughout the brain and tau pathology is more strongly correlated with clinical dementia than Aβ pathology. Therefore, the neurologist can determine the stage of the disease based on the pattern of spread of tau pathology.

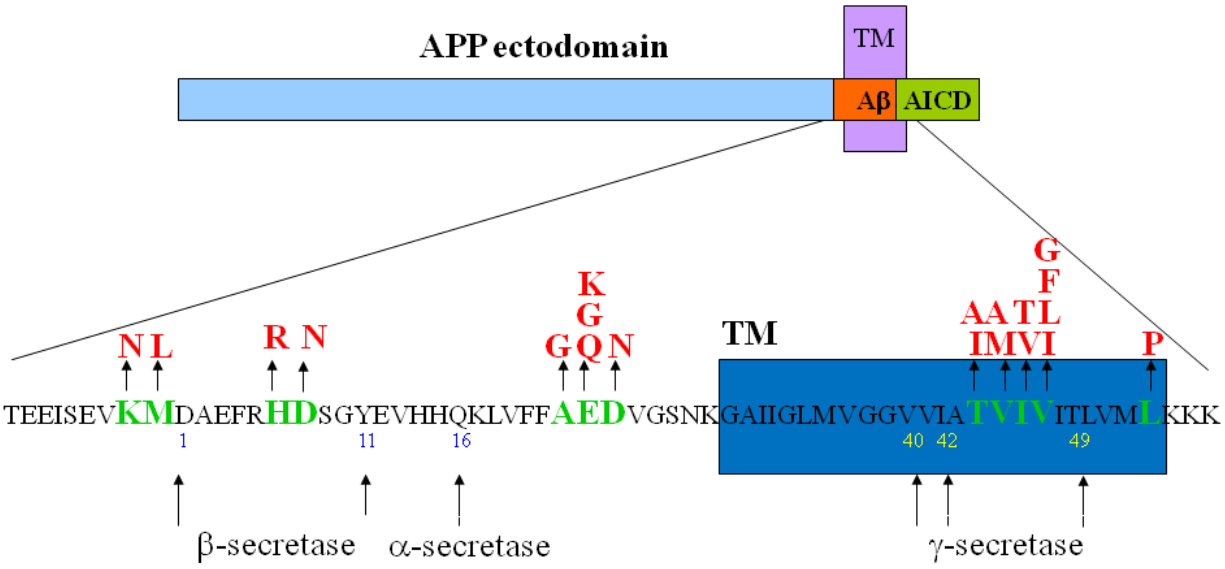


Figure 2. Schematic representation of APP. APP is composed of a large extracellular domain, a single transmembrane (tm)-domain, and a cytosolic tail. The Aβ peptide is derived from sequential cleavage of APP by β-secretase and γ-secretase. Numbers indicate amino acid positions of APP. Arrows point to the cleavage sites of α-secretase, β-secretase, and γ-secretase.

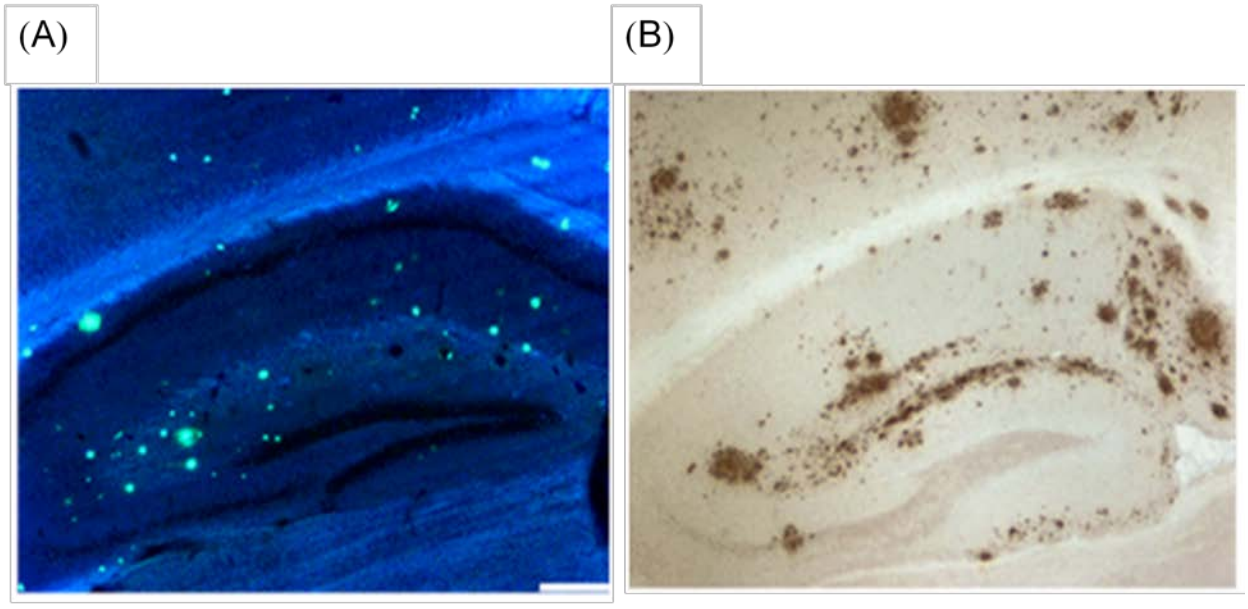


Figure 3. Fibrillar neuritic plaques and non-fibrillar diffuse plaques. High resolution pictures of (A) fibrillar neuritic plaques stained with Thiofalvin and (B) non-fibrillar diffuse plaques detected using anti-Aβ antibody immunohistochemical staining of PDAPP transgenic mice from Fryer et al. J Neurosci 2003.

1.5 ALZHEIMER'S GENETICS

1.5.1 Early dementia genes – *APP*, *PSEN1*, and *PSEN2*

Despite the fact that FAD accounts for less than 1% of AD cases, studies of rare autosomal dominant familial early-onset AD (EOAD) have provided valuable insights into the pathogenesis of AD. Mutations in the *APP*, *PSEN1*, and *PSEN2* genes were initially identified in familial EOAD, with AAO usually between the ages of 30 and 60[1-4]; however, several recent studies have found pathogenic mutations in these genes in late-onset families[15,16]. Most of these mutations are associated with increased Aβ or $A\beta_{42}$ production, resulting in increased Aβ aggregation in the brain[25]. These findings suggested that the Aβ aggregation and misfolding is central to triggering AD pathogenesis and to direct brain damaging[25,27].

APP is an integral membrane protein that is expressed in many tissues and primarily in the synapses of neurons. The primary function of APP is not known even though APP has been implicated to regulate synapse formation[37], neural plasticity[38], and iron export[39]. Mutations in *APP* cause FAD, cerebral amyloid angiopathy (CAA), or a combination of both[40-43]. There are several isoforms of *APP*, including 695, 751, and 770 amino acids in length. The 695- amino acid isoform is highly expressed in neurons but the 751- and 770- amino acid isoforms are primarily expressed in astrocytes[44,45]. Proteolysis of APP primarily occurs at the sites of APP endoproteolysis by α-secretase, β-secretase, and γ-secretase (see **Figure 2**). APP is first proteolyzed within the lumenal domain by α-secretase or β-secretase, resulting in ectodomain shedding and producing membrane-tethered β- or α-C-terminal fragments[46]. Next, the β- and α-C-terminal fragmented are cleaved by γ-secretase within the transmembrane domain to release Aβ and p3 peptides into extracellular space[46]. The majority of APP is cleaved by α-secretase and γ-secretase within the Aβ domain, resulting in nonpathogenic sAPPα and α-C-terminal fragment (CTF). Alternatively, APP can be proteolyzed by β-secretase and γ-secretase to produce Aβ peptides, sAPPβ and β-CTF[46].

Most missense mutations in *APP* are positioned in or near the sites of β-secretase and γ-secretase[1,47,48] cleavage. Mutations close to the C terminus of the Aβ region contribute to increased ratios of $A\beta_{40}$ or $A\beta_{42}$ without affecting total Aβ levels[49,50]. Even though the most abundant Aβ species is $A\beta_{40}$, $A\beta_{42}$ is more fibrillogenic and has been shown to be the first Aβ species deposited in AD[51,52]. The increased $A\beta_{42}$:$A\beta_{40}$ ratio caused by elevated $A\beta_{42}$ production suggests that physiological processing of APP is central to AD pathogenesis. On the contrary, the Swedish FAD mutation (p.K670N/p.M671L), located close to the β-secretase cleavage site, results in an increase in all species of Aβ and is associated with both AD and CAA[53]. The evidence that increased Aβ production contributes

to AD is also strengthened by the fact that people with Down syndrome (Trisomy 21) and individuals in families with an *APP* duplication all develop AD symptoms and neuropathology[54].

Some *APP* mutations that cause FAD and CAA do not affect Aβ production and are positioned in the Aβ peptide. These mutations, C-terminus to the α-secretase cleavage site in *APP*, modify the Aβ sequence by enhancing the propensity of Aβ to oligomerize, fibrillize, or be removed less efficiently[42,43,55]. A recent study described an autosomal recessive pattern of AD inheritance, resulting from the deletion of one amino acid at position 22 in Aβ, resulting in a highly fibrillogenic form of Aß[56,57]. Another study also reported a recessive mutation in *APP*, p.A673V, that significantly increases amyloidogenic β-secretase cleavage[58]. In contrast, a different codon change in the same position (p.A673T) results in a reduction in the formation of amyloidogenic peptides and protects against AD[59]. These studies clearly highlight the importance of beta-amyloid accumulation in AD pathogenesis.

Besides *APP*, mutations in *PSEN1* and *PSEN2* also have been associated with the rare, autosomal dominant form of FAD[2,3]. PSEN1 and PSEN2 are the catalytic component of gamma-secretase. *PSEN1* and *PSEN2* encode highly homologous polytopic transmembrane proteins. *PSEN* mutations can contribute to a selective and significant increase in the total level of $A\beta_{42}$ or the ratio of $A\beta_{42}$ to $A\beta_{4o}$ in the plasma[60]. Transgenic mice with *PSEN* mutations have increased $A\beta_{42}$ in the brain[61]. Additionally, when breeding transgenic mice with *PSEN* mutations with transgenic mice with FAD mutant human *APP*, these mice present accelerated Aβ deposition[62,63]. Primary neurons from *PSEN1* knockout mice produce Aβ less efficiently as a result of decreased γ-secretase activity[64. 64]. Several studies have shown that *PSEN1* mutations cause partial loss of function, resulting in a slight increase in $A\beta_{42}$ production and lower production of $A\beta_{40}$[65,66]. The partial loss of function caused by *PSEN1* mutations is associated with APP cleavage inefficiency and relative increased production of $A\beta_{42}$. This relative increase in $A\beta_{42}$ production elevates the tendency of Aβ to aggregate earlier in the brain.

To summarize, the mechanism by which *APP*, *PSEN1* and *PSEN2* mutations cause the autosomal dominant form of FAD supports the amyloid hypothesis, in which the neurodegenerative process results from an imbalance between Aβ production and Aβ clearance[27]. It also implicated that any genes involved in these pathways may also be associated with risk for developing AD.

1.5.2 Strongest risk factor for late-onset AD – *APOE*

LOAD cases account for more than 95% of AD cases and *APOE* has by far the greatest effects on risk for developing LOAD[5]. *APOE* is located on chromosome 19q13.2 and there are three major isoforms in humans,

ApoE2, ApoE3 and ApoE4, which were resulted from three alleles- *APOE* ε2, *APOE* ε3 and *APOE* ε4. These APOE isoforms differ in sequence by one amino acid at either position 112 or position 158 of the protein. The *APOE* ε4 allele increases the risk for AD while the ε2 allele decreases the risk for AD[6,7]. One APOE ε4 alleles roughly increases AD risk by 3 fold and two alleles by 12 fold[67]. The risk associated with *APOE* ε4 alleles is also associated with age at onset (approximately 5 yrs/ ε4 allele) in a dose-dependent manner[6]. However, about 50% of LOAD patients do not carry the *APOE* ε4 allele, suggesting that other genetic factors also affect LOAD risk.

Several hypotheses have been proposed about how APOE could influence AD pathogenesis. *In vitro* and *in vivo* experiments show that APOE is an Aβ binding molecule which affects not only the clearance of soluble Aβ but also the tendency of Aβ to aggregate by changing Aβ seeding and polymerization [68], both suggesting the potential influence on Aβ aggregation *in vivo*. Neuropathological and neuroimaging studies demonstrated that Aβ deposition begins earlier in ε4 carriers than ε4 non-carriers in both cognitively normal individuals and AD patients[69-71]. Studies have shown that APOE significantly alters the level, amount, and structure of Aβ deposition in the brain using transgenic mouse models of AD in an isoform-dependent manner (ε4>ε3>ε2)[72-75]. Recent studies have reported association between *APOE* genotype and cerebral Aβ deposition, and CSF $A\beta_{42}$ and tau levels[18,22,70]. Overall, these genetic, cellular, mouse and human studies demonstrate that *APOE* affects LOAD risk by influencing Aβ clearance and aggregation.

1.5.3 GWAS versus Sequencing

GWAS and sequencing studies are both powerful tools to discover disease-associated variants and genes for human diseases while there are some divergent aspects between GWAS and sequencing studies. The primary differences between GWAS and sequencing studies lie in the type of variants identified and the genetic-association hypotheses tested. Because the commercial genotyping arrays largely ignore low-frequency and rare variants (lack of systematic catalog of rare variants for assay design), GWAS are mostly used directly (through genotyping or imputation) or indirectly (through linkage disequilibrium) to identify "common" variants associated with disease traits; therefore, GWAS has been characterized as a method to test the common disease - common variant (CDCV) hypothesis. However, the GWAS-identified disease-related variants usually have small effects (odds ratio [OR] usually between 0.8~1.2). Moreover, common variants identified in GWAS usually explain a small fraction of the heritability even when the sample size of the study is huge. These findings implicate that additional functional genetic variants remain unidentified. Contrary to CDCV hypothesis, the alternative common disease – rare variant

(CDRV) hypothesis has been postulated to be able to explain the missing heritability, and sequence-based studies allow scientists to completely investigate "low-frequency" and "rare variants". Even though it remains debatable whether CDCV or CDRV will explain the missing heritability, current genetic studies suggest that CDCV and CDRV are not mutually exclusive and that human diseases are attributed to a combination of common and rare variants[12,76,77].

1.5.4 GWAS-identified AD loci

Recent GWAS have revealed many single nucleotide polymorphisms (SNPs) within several loci (*ABCA7*, *APOE*, *BIN1*, *CASS4*, *CD2AP*, *CD33*, *CELF1*, *CLU*, *CR1*, *EPHA1*, *FERMT2*, *HLA-DRB5/DRB1*, *IGHV1-67*, *INPP5D*, *MEF2C*, *MS4A6A*, *NME8, PICALM*, *PTK2B*, *SLC24A4*, *SORL1*, *TP53INP1*, and *ZCWPW1*) associated with AD risk, providing novel insights into potential biological pathways associated with AD pathogenesis[8,10,11,78]. These findings highlight several previously suspected biological pathways relevant to LOAD, including, amyloid precursor protein metabolism (*SORL1* and *CASS4*), tau metabolism (*CASS4* and *FERMT2*), lipid metabolism (*ABCA7* and *CLU*), innate and adaptive immunity (*CR1*, *CD33*, *HLA-DRB5/DRB1*, *IGHV1-67*, *INPP5D*, *MEF2C*, and *MS4A6A)*; cell adhesion (*EPHA1*), cell migration (*PTK2B*), and intracellular trafficking, (*BIN1*, *CD2AP,* and *PICALM*)[8,79]. Additionally, these results implicated several novel pathways underlying AD, including hippocampal synaptic function (*MEF2C* and *PTK2B*), apoptosis (*TP53INP1*), and cytoskeleton function and axonal transport (*CELF1*, *NME8*, and *CASS4*)[8]. However, some of these GWAS-identified genes may not be the *bona fide* disease-associated genes as many of these loci have multiple genes in the associated region. Therefore, the genuine AD-associated genes within these loci remain unclear.

Even though GWAS have identified several loci associated with AD risk, variants identified in these studies are usually common and non-coding variants that have no direct connection to potential biological mechanisms[8,10,11,78]. Additionally, these identified variants only account for a small proportion of heritability and have small effect sizes (OR between 0.8 and 1.2). One potential explanation is that these tagging GWAS-implicated SNPs are most likely surrogate markers for the underlying causal variants in each locus since tagging SNPs designed by the SNP chips cannot comprehensively capture all of the functional variants. Recent studies tend to support the hypothesis that rare variants could explain some of the missing heritability and have larger effect sizes for common diseases[80-82].

1.5.5 Sequencing-identified genes – *PLD3* and *TREM2*

Due to the significant improvement in NGS technology, whole-genome or whole-exome DNA sequencing have become cost-efficient and high-throughput tools for conducting genetic studies. Recently, two independent groups utilized whole-exome and whole-genome sequencing to uncover a low-frequency variant p.R47H in *TREM2* associated with increased AD risk[12,13]. Similarly, our lab carried out whole-exome sequencing in 14 large LOAD families and identified a rare variant p.V232M in *PLD3* gene segregated within two independent families[14]. Follow-up genotyping in seven independent case-control series demonstrated that p.V232M in *PLD3* increases AD risk by 2- fold[14]. In contrast to GWAS, these studies have used sequencing strategies to identify rare coding variants (ex: p.V232M in *PLD3* and p.R47H in *TREM2*) with significant effects and the consequences of these variants on disease mechanism can be more easily investigated by conducting in vitro or in vivo studies, which can provide more promising targets for therapeutics[12-14].

PLD3 is a non-classical member of the PLD protein family, which is a group of enzymes that hydrolyze phospholipids into fatty acids and lipophilic substances[83]. Unlike PLD1 and PLD2, which were previously reported to be involved in APP trafficking, PLD3 is poorly characterized and has no reported catalytic activity[84]. A recent study using whole-exome sequencing coupled with additional genotyping in large-scale independent case-control cohorts identified a rare variant p.V232M in *PLD3* which increases AD risk by two-fold. PLD3 is highly expressed in brains in the hippocampus, entorhinal cortex, and frontal cortex. Over-expression of PLD3 *in vitro* results in significantly lowered intracellular APP and extracellular $A\beta_{42}$ and $A\beta_{40}$ levels[14]; however, the underlying mechanism remains unclear.

TREM2 is a type one transmembrane receptor protein expressed on myeloid cells including microglia, monocyte-derived dendritic cells, osteoclasts and bone-marrow derived macrophages [85,86]. Additionally, protein expression of TREM2 in neurons has been reported[87]. TREM2 transduces its intracellular signaling through DAP12 (TYROBP)[85,86]. Although the natural ligands of TREM2 remain unknown, upon ligand binding, TREM2 associates with DAP12 to mediate downstream signaling. In the brain, TREM2 is primarily expressed on microglia and has been shown to control two signaling pathways: regulation of phagocytosis and suppression of inflammatory reactivity[88-90]. A previous study used microarray and laser microdissection of Aβ plaque-associated areas in an animal model of AD and found that *TREM2* is differentially expressed in Aβ plaque-associated versus Aβ plaque-free tissue [91]. Several studies have shown that homozygous loss-of-function mutations in *TREM2* or *DAP12* are associated with Polycystic lipomembranous osteodysplasia with sclerosing leukoencephalopathy (PLOSL)[92-95].

Recent studies identified a *TREM2* variant p.R47H as a risk factor for LOAD with an OR around 2[12,13], which is similar to the increased AD risk associated with carrying one *APOE* ε4 allele [67]. Several additional rare variants were enriched in AD cases; however, these variants failed to reach statistical significance[12,13,96]. Overall, both GWAS and sequencing studies suggest that inflammatory genes play a key role in mediating AD risk[12,13,97,98].

1.6 BIOMARKERS FOR AD

Due to advances in our understanding of the biomarkers of AD, the accumulation and deposition of Aβ can now be identified by several methods in humans regardless of their clinical status. Amyloid imaging using positron emission tomography (PET) scans has been used to detected presence, quantity, and location of Aβ by binding to fibrillar forms of Aβ in the brain[99,100]. In addition, measurement of $A\beta_{42}$ in CSF is a sensitive and specific method to ascertain the presence or absence of Aβ deposition in the brain [101-103] while CSF tau and $ptau_{181}$ levels are well correlated with the number of neurofibrillary tangles and tangle load in the human brain. Moreover, the fact that CSF tau/$ptau_{181}$ levels increase and CSF $A\beta_{42}$ levels decrease in AD patients indicate their roles as effective biomarkers for AD[22,104]. Functional magnetic resonance imaging (MRI) is able to directly measure task-based activation and resting neural connectivity and indirectly measures neuronal activity and network integrity by utilizing blood-oxygen level dependent (BOLD) T2 techniques. Other useful biomarkers include resting synaptic function via glucose metabolism, visinin-like protein-1 (VILIP-1) in CSF, cortical thinning in parietal and temporal cortices, and hippocampal volume[105]. Novel and highly-effective biomarkers are still urgently needed to help identify individuals at high risk for developing AD.

1.7 ENDOPHENOTYPE GENETICS

Instead of associating genetic markers with case-control status, endophenotype studies correlate genetic markers with quantitative traits in genetic analyses. The main advantages of using quantitative endophenotypes for AD are the increased statistical power and generation of testable hypothesis regarding the biological mechanism associated with genetic variants. Our previous work has shown that use of CSF biomarkers as endophenotypes for AD can lead to the identification of novel variants associated with disease risk, age at onset, and disease progression[17,19,22] as well as validation of AD genetic risk factors[23]. Use of CSF $A\beta_{42}$ and tau/$ptau_{181}$ levels as endophenotypes, which are mechanistically relevant to AD pathology, has identified additional variants and genes that modify Aβ and tau in CSF. Recently, several studies used gene expression as endophenotypes to identify novel

variants and genes affecting quantitative gene expression levels in human brains[106-109]. These data can be subsequently used to design specific functional assays for identified genetic variants to elucidate disease mechanism.

Chapter 2

Targeted re-sequencing of Alzheimer's disease associated genes

The research work in this chapter resulted in the following publications.

1. **Jin SC**, Pastor P, et al. *Pooled-DNA Sequencing Identifies Novel Causative Variants in PSEN1,GRN and MAPT, in A Clinical Early-Onset and Familial Alzheimer's Disease Ibero-American Cohort.* Alzheimer's Research & Therapy. 2012 Aug 20; 4(4): 34.

2. Benitez C BA, Cooper B, Pastor P, **Jin SC**, et al. *TREM2 is Associated with the Risk of Alzheimer's Disease in Spanish Population.* Neurobiology of Aging. 2013 Jun; 34(6): 1711.e15-7.

3. Benitez BA, Karch CM, Cai Y, **Jin SC**, et al. *The PSEN1, p.E318G Variant Increase the Risk of Alzheimer's Disease in APOE-ε4 carriers.* PLoS Genetics. 2013 Aug; 9(8): e1003685.

4. Cruchaga C, Karch CM*****, **Jin SC***, et al. *Rare Coding Variants in Phospholipase D3 Gene Confer Risk for Alzheimer's Disease.* Nature. 2014 Jan. 23; 505 (7484):550-4. doi: 10.1038/nature 12825

5. **Jin SC**, et al. *Coding Variants in TREM2 Increase Risk for Alzheimer's Disease.* Human Molecular Genetics. 2014 Jun 4. Pii: ddu227.

6**. Jin SC**, et al. *Pooled-DNA Sequencing for Elucidation of Genomic Risk Factors/Rare Variants Underlying Alzheimer's Disease.* Systems Biology of Alzheimer's Disease: Methods and Protocols: Springer. (Book chapter)

2.1 ABSTRACT

Recent studies have used DNA sequencing strategies to identify rare variants p.V232M in the *phospholipase D3* (*PLD3*) and p.R47H in the *triggering receptor expressed on myeloid 2* (*TREM2*) significantly associated with risk for AD. Additionally, genome-wide association studies (GWAS) uncovered *ABCA7*, *BIN1*, *CD33*, *CD2AP*, *CLU*, *CR1*, *EPHA1*, *MS4A4A*, and *PICALM* as AD risk loci; however, the most-significant variants in these loci are located within non-coding genomic regions and have no direct functional impact on AD pathogenesis. Thus, we hypothesized that if *PLD3* and *TREM2* are *bona fide* AD risk genes they will carry additional functional variants that can substantially affect risk for AD. Additionally, we hypothesized that GWAS-identified risk genes carry additional risk alleles across the frequency spectrum so that common, low frequency or rare variants within these genes may contribute to AD risk. To test these hypotheses, we performed deep re-sequencing of exonic and flanking intronic sequence for the aforementioned genes in order to identify novel functional variants and test for association with AD risk. Our analyses showed that rare variants in *PLD3* and *TREM2* are more frequently seen in cases than in controls of European American (EA) descent (*TREM2*: 6.7% in cases and 2.7% in controls; *PLD3*: 8.0% in cases and 3.1% in controls). Single-variant analyses showed that p.M6R ($p=0.02$; odds ratio [OR]=7.73 [1.09~61]) and p.A442A ($p=3.78\times10^{-7}$; OR=2.21 [1.58~2.8]) in *PLD3* and p.R62H ($p=2.36\times10^{-4}$; OR=2.36 [1.47~3.80]) in *TREM2* are significantly associated with AD risk in addition to p.V232M in *PLD3* and p.R47H in *TREM2*. Gene-based tests demonstrated that *PLD3* ($P_{SKAT-O}=1.44\times10^{-11}$; OR=2.75 [2.05~3.68]) and *TREM2* are genome-wide significantly associated with AD ($P_{SKAT-O}=5.37\times10^{-7}$; OR=2.55 [1.62~3.87]). The associations of *PLD3* and *TREM2* rare variants with AD risk are still highly significant after excluding p.V232M ($P_{SKAT-O}=1.5\times10^{-8}$; OR=2.58 [1.87~3.57]) and p.R47H ($P_{SKAT-O}=7.72\times10^{-5}$; OR=2.47 [1.62~3.87]) respectively, which suggests that additional *PLD3* and *TREM2* variants affect AD risk. Additionally, we found a significant association for *PLD3* coding variants with AD risk in African Americans ($P_{SKAT-O}=1.40\times10^{-3}$; OR=5.48 [1.77~16.92]). However, we did not find evidence of association for *TREM2* variants with AD risk in African Americans at both gene-level and SNP-level. We also validated 90 coding variants in GWAS-identified genes, 66 of which were not annotated in the Exome Variant Server, which consists of whole-exome sequencing (WES) data in 2,203 African American and 4,300 European American unrelated individuals. Bioinformatic analyses predict that 56 of the confirmed variants (62.2%) are damaging. Nucleotide conservation analyses suggest that 57 of the validated variants are under evolutionary constraint (a GERP score>2), which suggests that a large proportion of

rare coding variants in GWAS-identified genes are potentially functional. Together, these findings suggest that deep re-sequencing is an effective strategy to identify additional functional variants in AD-associated genes.

2.2 INTRODUCTION

In the past decade, enormous progress has been made in mapping genetic variants associated with complex human disease, one of which is the development of GWAS. GWAS have been successfully used to identify thousands of novel susceptibility loci for hundreds of disease traits. However, most of identified loci are located within non-coding genomic regions or are far from discovered genes. The susceptibility loci are found to harbor common variants which have very weak effect sizes and only explain a small proportion of phenotypic heritability. Moreover, it has been challenging to identify underlying functional variants due to their linkage disequilibrium (LD) with other variants. The fact that many of these susceptibility loci lie within gene-dense regions makes it difficult to identify the *bona fide* causal genes. Even though GWAS results can potentially reveal genes not previously suspected in disease etiology, the difficulties in replicating previous results and in translating these results into targets for downstream functional experiments are vivid

Due to the dramatic technological developments in next-generation sequencing (NGS), high-throughput sequencing of targeted genomic regions of the human genome in thousands of individuals in a single run is now cheap and feasible. Recent findings favor the rare variant-common disease hypothesis by which the combination of the effects of rare variants could explain a large proportion of the phenotypic heritability[81,82,110]. A previous study involving our group used whole genome, whole-exome and targeted Sanger sequencing to identify a rare variant p.R47H in *TREM2,* that increases AD risk by 2-fold[12]. In our recent work, Dr. Carlos Cruchaga performed whole-exome sequencing (WES) in 14 large Late onset AD (LOAD) families and follow-up genotyping of the candidate variants in independent large LOAD case-control series[14]. A rare variant p.V232M in *PLD3* segregated with disease status in two independent families and doubled risk for AD in seven independent case-control series with a total of more than 11,000 cases and controls of European American descent[14]. These sequencing-based studies clearly demonstrated that rare coding variants contribute significantly to AD and may account for the genetic heritability that was not fully explained by common variants. In order to test the hypothesis that additional functional variants in *TREM2*/*PLD3* significantly influence AD risk, we performed deep re-sequencing of *TREM2*/*PLD3* coding regions in European Americans and African Americans to identify functional rare variants that are associated with AD risk.

Recent GWAS identified nine novel genes, including *ABCA7*, *BIN1*, *CD2AP*, *CD33*, *CR1*, *CLU*, *EPHA1*, *MS4A4A*, and *PICALM*, as AD susceptibility loci[9-11,78]. However, the most-significant SNPs in these genes had very small effect size and only explained 50% of the phenotypic heritability for AD, which has an estimated overall heritability of up to 80%[111]. Recent studies support the paradigm that both common, low-penetrant and rarer, high penetrant alleles exist in the same gene contributing to disease risk[77,80,112,113]. Hence, we hypothesized that rare coding variants within GWAS-identified loci have stronger effects on AD risk compared to most-significantly associated common variants found in GWAS[10,11,78]. Our previous work has shown that the pooled-DNA sequencing strategy can effectively identify novel rare variants in genes of interest associated with AD risk[15,16,114-116]. In these studies, we used the same method and demonstrated that NGS-identified and GWAS-identified genes contain additional rare coding variants and that rare coding variants in NGS-identified genes (*PLD3* and *TREM2*) significantly increase AD risk. Further analysis of the underlying mechanisms by which these variants alter protein function could provide important insights into AD pathogenesis.

2.3 MATERIALS AND METHODS

2.3.1 Participants and Study Design

The Institutional Review Board (IRB) at Washington University School of Medicine in Saint Louis approved the study. Written informed consent was obtained from participants or their collateral source by the Clinical Core of the Charles F. and Joanne Knight Alzheimer's Disease Research Center (Knight-ADRC). The approval number for the Knight-ADRC Genetics Core is 93-0006.

2.3.2 Knight-ADRC study

The Knight-ADRC samples included 1,082 LOAD cases and 706 cognitively normal controls of EA descent, and 149 African American (AA) AD cases and 87 controls, matched for age, gender and ethnicity. These individuals were evaluated by the Clinical Core of the Knight-ADRC and a blood sample was drawn for genetics studies. Cases received a clinical diagnosis of AD dementia in accordance with standard criteria, and dementia severity was determined with the Clinical Dementia Rating (CDR)[117], with higher scores being associated with more severe cognitive decline. Controls underwent the same assessment but were cognitively normal (CDR=0). The Knight-ADRC samples were recruited without enrichment based on family history. **Tables 1 and 2** summarize the demographic characteristics of re-sequenced samples of EA descent and of AA descent respectively

Table 1. Demographic characteristics of re-sequenced individuals of European American descent

	Knight-ADRC cases	NIA-LOAD cases	Controls
N	1,082	1,000	1,648
Age ± SD (Range)	72.65±9.17 (44-103)	71.77±6.98 (48-98)	76.88±9.00 (50-105)
% Female	57.72	64.86	60.12
% *APOE*-ε4 Positive	55.86	76.21	29.25

Sample size (N), mean, standard deviation and range for age in years, percentage of female subjects, and percentage of subjects that carry at least one *APOE*-ε4 allele for the Knight-ADRC AD cases, NIA-LOAD AD cases, and cognitively-normal elderly controls from both studies. We sequenced *ABCA7*, *BIN1*, *CD33*, *CD2AP*, *CR1*, *EPHA1*, *MS4A4A*, *PICALM*, *PLD3* and *TREM2* in these samples.

Table 2. Demographic characteristics of re-sequenced individuals of African American descent

	Knight-ADRC cases	NIA-LOAD cases	Controls
N	149	55	132
Age ± SD (Range)	75.31±8.57 (51-94)	70.75±7.00 (51-84)	73.37±7.66 (60-94)
% Female	77.18	75	68.18
% *APOE*-ε4 Positive	58.9	62.5	39.39

Sample size (N), mean, standard deviation and range for age in years, percentage of female subjects, and percentage of subjects that carry at least one *APOE*-ε4 allele for the Knight-ADRC AD cases, NIA-LOAD AD cases, and cognitively-normal elderly controls from both studies. We sequenced *ABCA7*, *BIN1*, *CD33*, *CD2AP*, *CR1*, *EPHA1*, *MS4A4A*, *PICALM*, *PLD3* and *TREM2* in these samples.

2.3.3 NIA-LOAD study

The NIA-LOAD study case control series consists of one affected individual from each of 1,000 families multiply affected by AD and 942 healthy unrelated controls of EA descent and 55 AD cases and 45 controls of AA descent, with no family history of dementia in first-degree relatives. All AD cases were diagnosed with AD dementia using criteria equivalent to the National Institute of Neurological and Communication Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) for probable AD[118]. All NIA-LOAD AD cases had a family history of AD. Probands were required to have a diagnosis of definite or probable AD and a sibling with definite, probable or possible AD with a similar age at onset. A third biologically-related family member (first, second or third degree) was also recruited, regardless of cognitive status. We screened one individual from each family by selecting the youngest affected family member with the most definitive diagnosis (i.e. individuals with autopsy confirmation were chosen over those with clinical diagnosis only. Written informed

consent was obtained from all participants, and the study was approved by local IRB committees. See **Tables 1 and 2** for the demographic characteristics of re-sequenced samples.

2.3.4 Deep re-sequencing and sequencing analysis

We used a pooled-DNA sequencing strategy as previously described[15,16,97,119]. A schematic representation of the technique was shown in **Figure 1.** Equimolar amounts of individual DNA samples were pooled together after being measured using Quant-iT^{TM} PicoGreen (Invitrogen) reagent. Pools with 100 ng of DNA from 94 individuals were made. Protein coding regions of the *ABCA7*, *BIN1*, *CD2AP*, *CD33*, *CR1*, *CLU*, *EPHA1*, *MS4A4A*, *PICALM*, *PLD3*, and *TREM2* genes were amplified by the PCR using specific primers and Pfu Ultra high-fidelity polymerase (Agilent). We used the web-based Primer3 tool (http://bioinfo.ut.ee/primer3) for the design of PCR primers. To guarantee sufficient coverage of each desired exon, a minimum of 50bp of flanking sequences on either side was required for the primer design. An average of 20 diploid genomes (0.14 ng DNA) per individual was used as the input. PCR products were cleaned using QIAquick PCR (Qiagen) purification kits, quantified using Quant-iT PicoGreen reagent and ligated in equimolar amounts using T4 Ligase and T4 Polynucleotide Kinase. After ligation, concatenated PCR products were randomly sheared by sonication and prepared for sequencing on an Illumina HiSeq 2000 according to the manufacturer's specifications. The pCMV6-XL5 amplicon (1,908 bp) was included in the reaction as a negative control. The positive controls contained 10 different constructs (p53 gene) with synthetically engineered mutations at an assigned frequency of one mutated copy per 188 normal copies were amplified and pooled with the PCR products. Paired-end reads (150 bp) were aligned to the reference sequence (GRCh37/hg19) using SPLINTER[120]. SPLINTER uses the positive controls to estimate sensitivity and specificity for variant calling[120]. The wild-type-to-mutant ratio in the positive control was similar to the relative frequency expected for a single mutation in one pool (1 chromosome mutated in 94 samples = 1/188). SPLINTER uses the negative controls (first 900 bp) to model the error rates across the 102 bp Illumina reads and to create an error model from each sequencing run of the machine. Based on the error model (see **Figure S1** for an example), SPLINTER calculates a p-value for the probability that a predicted variant is a true positive. A p-value at which all mutants in the positive controls were identified was defined as the cutoff to estimate the sensitivity and specificity. All mutants included as part of the amplified positive control vectors were found upon achieving .30-fold coverage at mutated sites (sensitivity=100%) and only ~80 sites in the 1,908 bp negative control vector were

predicted to be polymorphic (specificity=~95%). The variants with a p-value below this cutoff value were considered for follow-up genotyping.

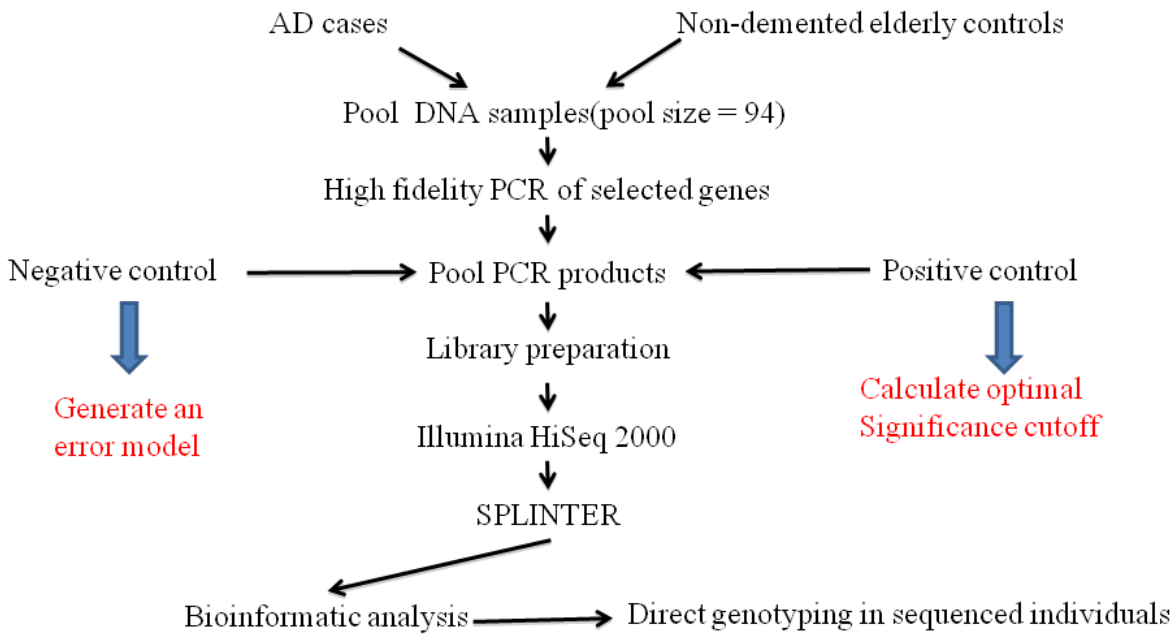


Figure 1: Schematic of the pooled-DNA sequencing technique. Pooled-DNA sequencing was performed in DNA pools of 94 individuals to identify rare coding variants by using Illumina HiSeq 2000. The SPLINTER software was used to call the variants. High-confident variants were selected for Sequenom or KASPar genotyping in all sequenced samples. SPLINTER, short indel prediction by large deviation inference and nonlinear true frequency estimation by recursion.

2.3.5 SNP genotyping and segregation with disease

All rare missense or splice site variants identified by SPLINTER were validated by directly genotyping all sequenced individuals using Sequenom iPLEX or KASPar genotyping systems as described previously[14-16]. To avoid potential batch/plate effects, genotyping was repeated with heterozygous cases or controls that were randomly assigned in the plates. The genotype call rate of these SNPs was >98%. We genotyped confirmed variants in all available family members to determine whether these variants segregate with disease status.

2.3.6 Population structure

A principal component analysis (PCA) was conducted to infer genetic structure of individuals who have GWAS data available using the EIGENSTRAT software as previously described[121]. When GWAS data is not available, self-reported ethnicity was used for consideration of inclusion in the final analysis.

2.3.7 Statistical analyses

We used the Fisher's exact test to test for association between AD risk and each genetic variant in *TREM2* and *PLD3* using PLINK [122]. For the gene-based association, we tested for association between the confirmed set of variants in *TREM2* and *PLD3* and AD risk using SKAT-O conducted using R package SKAT[123]. The gene-level significance threshold is defined by type-I error rate divided by the number of human genes ($0.05/20{,}000=2.4\times10^{-6}$). For the family-based association analysis, we used the Fisher's exact test to determine whether any *TREM2* variants are associated with disease status within families. We did not perform single-variant and gene-based analyses for GWAS-identified genes because the direct genotyping has not been completed.

2.3.8 Bioinformatic analysis

The EVS (http://evs.gs.washington.edu/EVS), SeattleSeq Annotation (http://snp.gs.washington.edu/SeattleSeqAnnotation137/) and the Ensembl Genome Database (http://useast.ensembl.org/index.html) were used to annotate the rare variants. SIFT (http://sift.jcvi.org/) and Polyphen2 (http://genetics.bwh.harvard.edu/pph2/) algorithms were used to predict the functional effect of the identified variants. The Uniprot database (http://www.uniprot.org/) was used to extract and perform alignment of the protein sequences across different species. To determine the effect of the p.A442A variant in *PLD3* on splicing we used the ESEfinder (http://rulai.cshl.edu/tools/ESE). Multiple sequence alignment was performed by ClustalW2, and the PLD3 orthologues were downloaded from Ensembl.

2.3.9 Gene expression and alternative splicing analyses

Parietal lobes of 82 AD case brains and 39 individuals without dementia were acquired from the Knight-ADRC for validating *PLD3* gene expression and alternative splicing. Parietal lobes from EA autopsy-confirmed AD (N = 2) case brains were selected for validating *TREM2* alternative splicing. All subjects signed and provided the hospital autopsy form. If the participant does not provide future consent before death, the DPOA (durable power of attorney) or next of kin provide it after death. The Washington University IRB reviewed the protocol operated by Knight-ADRC Neuropathology Core and determined the study was exempt from approval. RNA was extracted from brain tissue using an RNeasy kit (Qiagen) following the manufacture's protocol. cDNA was synthesized from the extracted RNAs (10 ug) by the PCR using the High-Capacity cDNA Reverse Transcriptase kit (ABI).

Total *PLD3* gene expression levels were analyzed by real-time PCR, using an ABI-7900 real-time PCR system. TaqMan assays were used to quantify *PLD3* mRNA levels. Primers and TaqMan probe for the reference

gene, *GAPDH*, were designed over exon-exon boundaries, using Primer Express software v3 (ABI). Cyclophilin A (ABI: 4326316E) was also used as a reference gene. Each real-time PCR run included within-plate triplicates and each experiment was performed at least twice for each sample.

To evaluate alternative slicing changes due to *PLD3* p.A442A variant, we selected eight p.A442A carriers as well as eight CDR-, age-, *APOE*-, and PMI-matched individuals to analyze the expression level of exon 11 containing transcripts, the exon in which the p.A442A variant is located. Real-time PCR assays were used to quantify *PLD3* exon 7 (forward primer, 5'-GCAGCTCCATCCCATCAACT-3'; reverse, 5'-CTTGGTTGTAGCGGGTGTCA-3'), exon 8 (forward primer, 5'-CTCAACGTGGTGGACAATGC-3'; reverse, 5'-AGTGGGCAGGTAGTTCATGACA-3'), 9 (forward primer, 5'-ACGAGCGTGGCGTCAAG-3'; reverse, 5'-CATGGATGGCTCCGAGTGT-3'), 10 (forward primer, 5'-GGTCCCCGCGGATGA-3'; reverse, 5'-GGTTGACACGGGCATATGG-3') and 11 (first pair of primers: forward primer, 5'-CCAGCTGGAGGCCATTTTC-3'; reverse, 5'-TGTCAAGGTCATGGCTGTAAGG-3'; second pair forward primer, 5'-GCTGCTGGTGACGCAGAAT-3'; reverse, 5'-AGTCCCAGTCCCTCAGGAAAA-3'). Two pairs of primers were designed for exon 11 as an internal control. SYBR-green primers were designed using Primer Express software, v3 (ABI). Each real-time PCR run included within-plate duplicates and each experiment was performed at least twice for each sample. Real-time data were analysed using the comparative Ct method. Only samples with a standard error of, 0.15% were analyzed. The Ct values for exon 11 were normalized with the Ct value for the exons 7–10. The relative exon 11 levels for the p.A442A carriers versus the non-carriers were compared using a *t*-test.

According to Ensembl, *TREM2* encodes three alternative transcripts (ENST00000373113, ENST00000373122, and ENST00000338469). To evaluate *TREM2* alternative splicing and determine whether these transcripts exist in the human brain, cDNA isolated from parietal lobes of two Alzheimer's disease brains were amplified using PCR with Pfu (Agilent) enzyme. The PrimerQuest Design Tool (Integrated DNA Technology) was used to design primers spanning exon junctions. PCR primers include a forward primer located at the junction between exons 3 and 4 and a reverse primer located in exon 4 of the longest transcript ENST00000373113 (See **Table S1** and **Fig S2** for primers and the expected amplicon lengths). For the transcript ENST00000373122, we designed a unique forward primer, which only exists in this transcript, located spanning the exon3-exon4 junction and a reverse primer located in exon 4 to amplify the sequence (See **Table S1** for designed primers and the expected length). For the transcript ENST00000338469, a forward primer located across the exon3-exon5 junction and a

reverse primer located in exon 5 were used to amplify this transcript (See **Table S1** and **Fig S2** for primers and the expected amplicon lengths). Each PCR reaction contained 7.5 µl of PerfeCTa SYBR Green FastMix (Quanta Biosciences), 720 nM forward and reverse primers, and 15 ng of cDNA in a final volume of 15 µl. The reaction mix was incubated using a program as follows: (1) 45°C for 2 min; (2) 95°C for 2 min; (3) 95°C for 15 sec; (4) 60°C for 1 min; (5) repeat steps 3-4 for 40 cycles. The resulting PCR product was run on a 2% agarose gel and visualized on a Syngene Imaging system.

2.4 RESULTS

2.4.1 *PLD3* sequencing

To identify risk variants in *PLD3*, we performed deep re-sequencing of *PLD3* coding regions in 2,363 cases and 2,027 controls of EA descent, and 130 cases and 172 controls of AA descent (see **Figure 2**). Fourteen variants (p.M6R, p.P76A, p.T136M, p.K228R, p.V232M, p.N236S, p.N284S, p.C300Y, p.A325T, p.Q406H, p.T426A, p.G452E, p.G454C, and p.R488C; **Table 3 and Figure 3**) were observed more frequently in cases than in controls, including nine variants (p.P76A, p.T136M, p.K228R, p.N284S, p.A325T, p.Q406H, p.T426A, p.G454C, and p.R488C) that were unique to cases (a total of 16 carriers; **Table 3 and Figure 3**). The gene-based burden analysis resulted in a genome-wide significant association of carriers of *PLD3* coding variants among Alzheimer's disease cases (7.99%) compared to controls (3.06%; $p=1.44\times10^{-11}$; OR=2.75, 95% CI=2.05~3.68; **Figure 2**). When the p.V232M variant was excluded, the association remained highly significant, still passing genome-wide multiple test correction ($p=1.58\times10^{-8}$; OR=2.58, 95% CI=1.87~3.57; **Figure 2 and Table S2**), indicating that there are additional variants in *PLD3* that increase risk for AD independent of p.V232M. There were two additional highly conserved variants (see **Figure S3**), that were nominally associated with LOAD risk: p.M6R (p=0.02; OR=7.73, 95% CI=1.09~61; **Table 3**), and p.A442A ($p=3.78\times10^{-27}$; OR=2.12, 95% CI=1.58~2.83; **Table 3**).The p.A442A variant was included in the gene-based analysis because our bioinformatics and functional analyses indicate that this variant affects splicing and gene expression (see **Figures S3** and **S4**). After excluding both p.V232M and p.A442A from the model, the gene-based association remained highly significant ($p=1.61\times10^{-3}$; OR=2.86, 95% CI=1.62~5.06; **Figure 2 and Table S2**), which indicates other risk variants in *PLD3* remained undiscovered. In order to identify which of the remaining variants affect risk for AD, functional studies will be required.

If the association for *PLD3* with AD risk is real, it is possible that rare coding variants in *PLD3* in other populations will also increase risk for AD. We therefore sequenced *PLD3* in 302 AA AD cases and controls (see

Figure 2). Both the p.V232M and p.A442A variants were found in AD cases but not controls, and the p.A442A variant showed a significant association with AD risk in African Americans (p = 0.03; **Table 3**). There was also a significant association with LOAD risk at the gene level ($p=1.4\times10^{-3}$; OR=5.48, 95% CI=1.77~16.92; **Figure 2**). This consistent evidence of association with AD risk, at the single-nucleotide polymorphism (SNP) and gene level in two different populations strongly supports *PLD3* as an AD risk gene.

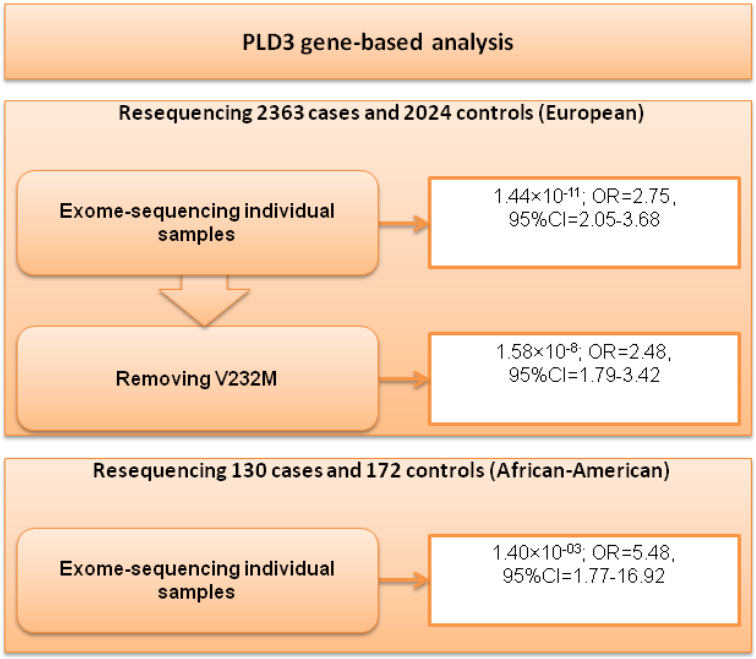


Figure 2. Gene-based analyses for *PLD3* in European and African Americans. SKAT-O analysis was performed to evaluate the gene-level association for *PLD3* variants. CI, confidence interval; OR, odds ratio.

Table 3. Sequence variants found in *PLD3* in the NIA-LOAD, Knight-ADRC and NIA-UK datasets.

Chr. position	AA		NIA LOAD	Knight ADRC	NIA-UK	total	MAF %	p-value	OR (95% CI)	EVS MAF%	SIFT	Polyphen
40872407	**M6R**	**CA**	0	8	1	9	0.19	**0.02**	7.73	NP	tolerated	deleterious
		CO	0	1	0	1	0.02		(1.09-61)			
40872764	**S63G**	**CA**	3	1	0	4	0.08	0.74	0.68	0.16	tolerated	neutral
		CO	5	0	0	5	0.12		(0.18-2.55)			
40872803	**P76A**	**CA**	3	1	0	4	0.08	0.12	NA	0.03	tolerated	benign
		CO	0	0	0	0	0.00					
40873764	**T136M**	**CA**	0	1	0	1	0.02	0.54	NA	NP	tolerated	deleterious
		CO	0	0	0	0	0.00					
40876055	**H197Y**	**CA**	0	1	0	1	0.02	0.49	0.85	NP	damaging	benign
		CO	0	1	0	1	0.02		(0.05-13-7)			
40877584	**K228R**	**CA**	1	1	1	3	0.06	0.25	NA	NP	damaging	deleterious
		CO	0	0	0	0	0.00					
40877595	**V232M**	**CA**	29	16	1	46	0.99	**$1.05x10^{-5}$**	3.99	0.48	damaging	deleterious
		CO	8	2	0	10	0.25		(2.01-7.94)			
40877608	**N236S**	**CA**	0	2	0	2	0.04	0.40	1.71	0.01	damaging	deleterious
		CO	0	1	0	1	0.02		(0.15-18.91)			
40877752	**N284S**	**CA**	0	1	0	1	0.02	0.54	NA	NP	tolerated	deleterious
		CO	0	0	0	0	0.00					
40880407	**C300Y**	**CA**	2	3	0	5	0.10	0.46	2.14	0.09	tolerated	deleterious
		CO	1	0	1	2	0.04		(0.41-11.06)			
40880481	**A325T**	**CA**	0	1	0	1	0.02	0.54	NA	NP	damaging	deleterious
		CO	0	0	0	0	0.00					
40883725	**Q406H**	**CA**	1	0	0	1	0.02	0.54	NA	NP	tolerated	neutral
		CO	0	0	0	0	0.00					
40883783	**T426A**	**CA**	1	0	0	1	0.02	0.54	NA	NP	tolerated	neutral
		CO	0	0	0	0	0.00					
40883911	**G435V**	**CA**	0	0	0	0	0.00	0.46	NA	0.02	damaging	deleterious
		CO	1	0	0	1	0.02					
40883933	**A442A**	**CA**	48	35	12	95	2.09	**$1.08x10^{-5}$**	2.31	1.59	-	-
		CO	17	12	7	36	0.90		(1.56-3.41)			
40883956	**Q450L**	**CA**	0	0	0	0	0.00	0.46	NA	NP	tolerated	neutral
		CO	0	0	1	1	0.02					
40883962	**G452E**	**CA**	4	6	0	10	0.21	0.16	2.86	0.09	tolerated	deleterious
		CO	0	2	1	3	0.07		(0.78-10.4)			
40883967	**G454C**	**CA**	0	1	0	1	0.02	0.54	NA	NP	damaging	deleterious
		CO	0	0	0	0	0.00					
40884037	**D477G**	**CA**	0	1	0	1	0.02	0.49	0.42	0.02	damaging	deleterious
		CO	0	1	0	1	0.02		(0.04-4.72)			
40884069	**R488C**	**CA**	0	3	0	3	0.06	0.25	NA	0.02	damaging	deleterious
		CO	0	0	0	0	0.00					
total		**CA**	1106	1114	143	2363						
total		**CO**	928	913	183	2024						

The coding region of *PLD3* was sequenced in 2,363 AD cases and 2,024 controls (see materials and methods) from the Knight-ADRC, NIA-LOAD and the NIA-UK datasets. The table shows the coding variants identified as well as the number of carriers in each dataset. The minor allele frequency (MAF) in cases and in controls, the p-value and the OR for the association with case-control status is shown. The MAF of the identified variants in the Exome Variant Server (EVS) is shown. We also used SIFT and Polyphen to predict the impact of the non-synonymous changes on protein function. NA: not available. NP: not present

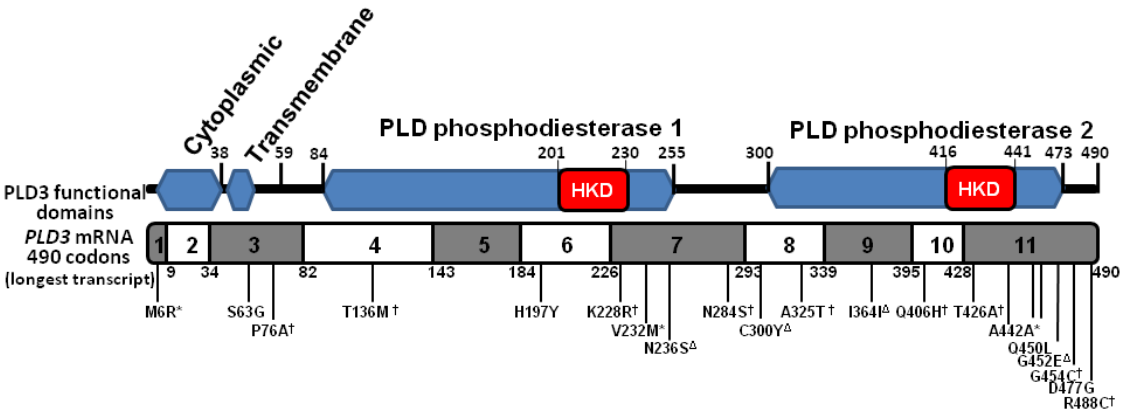


Figure 3. Schematic representation of *PLD3* and the relative position of the *PLD3* variants. PLD3 has two PLD phosphodiesterase domains, which contain an HKD signature motif (H-X-K-X(4)-D-X(6)-G-T-X-N, where X represents any amino acid residue) The scheme also shows the exon composition of the longest *PLD3* mRNA and the position of the variants found in this study. *Variants significantly associated with Alzheimer's risk. † Variants found only in Alzheimer's disease cases. Δ Variants that are more frequency in Alzheimer's disease cases than in controls.

2.4.2 *TREM2* sequencing

TREM2 was sequenced in 2,082 AD cases and 1,648 cognitively-normal elderly controls of EA descent and 204 AD cases and 132 controls of AA descent using pooled-DNA sequencing (see **Tables 1 and 2**). In European Americans, pooled sequencing identified sixteen rare variants in *TREM2*, six of which were not identified in the recent studies[12,13,96]: p.R52H, p.R136W, p.E151K, p.W191X, p.E202D, and p.H215Q (see **Table 4 and Figure 4A**). Nine variants (p.R52H, p.T66M, p.R136W, p.R136Q, p.H157Y, p.W191X, p.E202D, p.H215Q and p.T223I) were only found in AD cases (a total of 13 carriers; **Table 4**), three of which were not reported in the Exome Variant Server (EVS) database (p.R136W, p.E220D and p.H215Q). The protein sequence conservation analysis suggests that p.R47H, p.R52H, p.R62H, p.T66M, p.D87N, p.T96K, p.E151K, p.H157Y, p.L211P, and p.T223I are particularly conserved across species (see **Figure 4B**). For the single-variant analyses, we replicated the association of p.R47H with AD risk ($p=9.17\times10^{-4}$; OR=2.63 [1.44~4.81]; **Table 4**). The minor allele (T) of a second variant, p.R62H was also significantly associated with increased AD risk ($p=2.36\times10^{-4}$; OR=2.36 [1.47~3.80]; **Table 4 and Figure 4A**) after multiple test correction. In African Americans, a total of eleven coding variants were identified in

TREM2, four of which were not identified in previous studies (p.A105V, p.E151K, p.W191X, and p.E202D; **Table 5 and Figure 4A**). None of these *TREM2* variants, including the confirmed risk factor p.R47H, were significant associated with AD risk which may due to our small AA sample size (see **Table 5)**.

To determine whether *TREM2* rare variants collectively contribute to AD risk, we performed a gene-based association test using the optimal SNP-set sequence kernel association test (SKAT-O). In European Americans, gene-based association testing for *TREM2* achieved genome-wide significance ($P_{SKAT-O}=5.37\times10^{-7}$; OR=2.55 [1.80~3.67]; **Table 4**) and remained highly significant after excluding p.R47H ($P_{SKAT-O}=7.72\times10^{-5}$; OR=2.47 [1.62~3.87]; **Table S3**), the only confirmed risk factor for AD, in *TREM2*. This result demonstrates that additional rare variants in *TREM2* contribute to AD risk. The cumulative carrier frequency of all *TREM2* variants is 6.7% (139 out of 2,082) in AD cases and 2.7% (45 out of 1,648) in cognitively normal elderly controls of EA descent. However, we did not identify significant association for *TREM2* coding variants with AD risk ($P_{SKAT-O}=1$; OR=0.73 [0.44~1.23]; **Table 5).** The cumulative carrier frequency of *TREM2* variants is 25.4% (52 out of 204) in AD cases and 31.8% (42 out of 132) in controls of AA descent.

Next we used the NIA-LOAD family series to test whether *TREM2* variants are associated with disease status within families. We found that p.R47H and p.R62H were more frequently found in AD cases than in controls (Fisher's exact $p=4.65\times10^{-2}$ and 6.87×10^{-3} for p.R47H and p.R62H respectively; **Table 6**) after directly genotyping all samples individually from 13 and 21 independent families respectively. Other variants were either too rare or the families were not sufficiently large to provide statistical evidence of association with disease within and across families (see **Table 6 and Figure S4)**. These results strongly support p.R62H as a risk factor for AD in addition to p.R47H.

Two of the identified *TREM2* (p.W191X and p.E202D, found only in AD cases) variants are only located in the coding region of the predicted shortest transcript (ENST00000338469), encoding a soluble form of TREM2 (sTREM2). However, it remains unclear whether sTREM2 results from alternative splicing or sequential cleavage of the transmembrane form of TREM2 molecules. To confirm the existence of this alternative transcript, we performed polymerase chain reaction (PCR) on cDNA from 2 human brains using transcript-specific primers to amplify each isoform based on the Ensembl database. Gel electrophoresis analysis confirmed that there are at least three distinct *TREM2* transcripts: ENST00000373113, ENST00000373122, and ENST00000338469 expressed in the parietal

Table 4. Rare *TREM2*-variant association in sequenced samples of EA descent*

AA Change	SNP	CHR:BP (hg19)	Guerreiro et al. [#]		Cuyvers et al. [@]		AD Cases		Controls				
			P[#]	OR (95% CI)[#]	P[@]	OR (95% CI)[@]	No. of Cases	No. of Carriers	No. of Controls	No. of Carriers	P[!]	OR (95% CI)	PolyPhen
All variants											5.37×10^{-7}[&]	2.55 (1.80-3.67)	
p.Q33X	rs104894002	6:41129295	0.25	NA	NA	NA	2050	1	1611	1	1	0.81 (0.05-12.97)	NA
p.R47H	rs75932628	6:41129252	<0.001	4.5 (1.7-11.9)	0.08	3.01 (0.83-10.94)	2050	46	1616	14	9.17×10^{-4}	2.63 (1.44-4.81)	Damaging
p.R52H	rs374851046	6:41129237	NA	NA	NA	NA	2077	1	1642	0	1	NA	Damaging
p.R62H	rs143332484	6:41129207	0.5	0.8 (0.5-1.4)	0.08	1.54 (0.96-2.49)	2050	68	1618	24	2.36×10^{-4}	2.36 (1.47-3.80)	Benign
p.T66M	rs201258663	6:41129195	0.5	NA	NA	NA	2052	1	1622	0	1	NA	Damaging
p.D87N	rs142232675	6:41129133	0.02	NA	NA	NA	2051	9	1619	4	0.41	1.84 (0.56-5.96)	Damaging
p.T96K	rs2234253	6:41129105	0.72	1.4 (0.3-6.0)	NA	NA	2044	2	1609	2	1	0.81 (0.11-5.77)	Damaging
p.R136W	NA	6:41127606	NA	NA	NA	NA	2003	2	1562	0	0.51	NA	NA
p.R136Q	rs149622783	6:41127605	1	1.8 (0.1-28.6)	NA	NA	2047	1	1623	0	1	NA	Benign
p.E151K	rs79011726	6:41127561	NA	NA	NA	NA	2077	0	1642	1	0.45	0	Damaging
p.H157Y	rs2234255	6:41127543	0.36	NA	NA	NA	2052	3	1610	0	1	NA	Damaging
p.W191X	rs2234258	6:41126429	NA	NA	NA	NA	1816	1	1440	0	1	NA	NA
p.E202D	NA	6:41126395	NA	NA	NA	NA	2077	1	1642	0	1	NA	NA
p.L211P	rs2234256	6:41126655	0.56	0	NA	NA	2043	2	1605	2	1	0.81 (0.11-5.76)	Benign
p.H215Q	NA	6:41126642	NA	NA	NA	NA	2001	1	1560	0	1	NA	NA
p.T223I	rs138355759	6:41126619	NA	NA	NA	NA	2077	2	1642	0	0.51	NA	Benign

* NA represents not applicable. [#] These values were derived from Table 2 in Guerreiro et al[12]. [@] These values were derived from Table 1 in Cuyvers et al[96]. [!] The Fisher's exact test was used to calculate the p values using the default commands in PLINK. [&] This p value summarizes the gene-based association of the identified SNP set and was estimated using the SKAT-O algorithm.

Table 5. Rare *TREM2* variants found in African Americans

Variant	SNP	Position	AD Cases		Controls		P[!]	OR (95% CI)	SIFT	PolyPhen
			No. of Cases	No. of Carriers	No. of Controls	No. of Carriers				
All variants							1[&]	0.73 (0.44-1.23)		
p.R47H	rs75932628	6:41129252	203	0	131	1	0.39	0	Tolerated	Damaging
p.R62H	rs143332484	6:41129207	203	1	131	0	1	NA	Tolerated	Benign
p.T66M	rs201258663	6:41129195	203	1	132	0	1	NA	Damaging	Damaging
p.D87N	rs142232675	6:41129133	203	1	131	1	1	0.64 (0.04-10.35)	Tolerated	Damaging
p.T96K	rs2234253	6:41129105	200	45	131	38	0.19	0.72 (0.45-1.14)	Damaging	Damaging
p.A105V	rs145080901	6:41129078	203	1	131	1	1	0.64 (0.04-10.35)	Damaging	Damaging
p.E151K	rs79011726	6:41127561	204	1	132	0	1	NA	NA	Damaging
p.H157Y	rs2234255	6:41127543	203	1	132	0	1	NA	Damaging	Damaging
p.W191X	rs2234258	6:41126429	183	12	122	8	1	1 (0.40-2.48)	NA	NA
p.E202D	NA	6:41126395	204	0	132	1	0.39	0	Damaging	NA
p.L211P	rs2234256	6:41126655	201	46	131	38	0.29	0.78 (0.49-1.22)	Tolerated	Benign

* NA represents not applicable. [&] This p value summarizes the gene-based association of the identified SNP set and was estimated using the SKAT-O algorithm

[!] The Fisher's exact test was used to calculate the p values using plink.

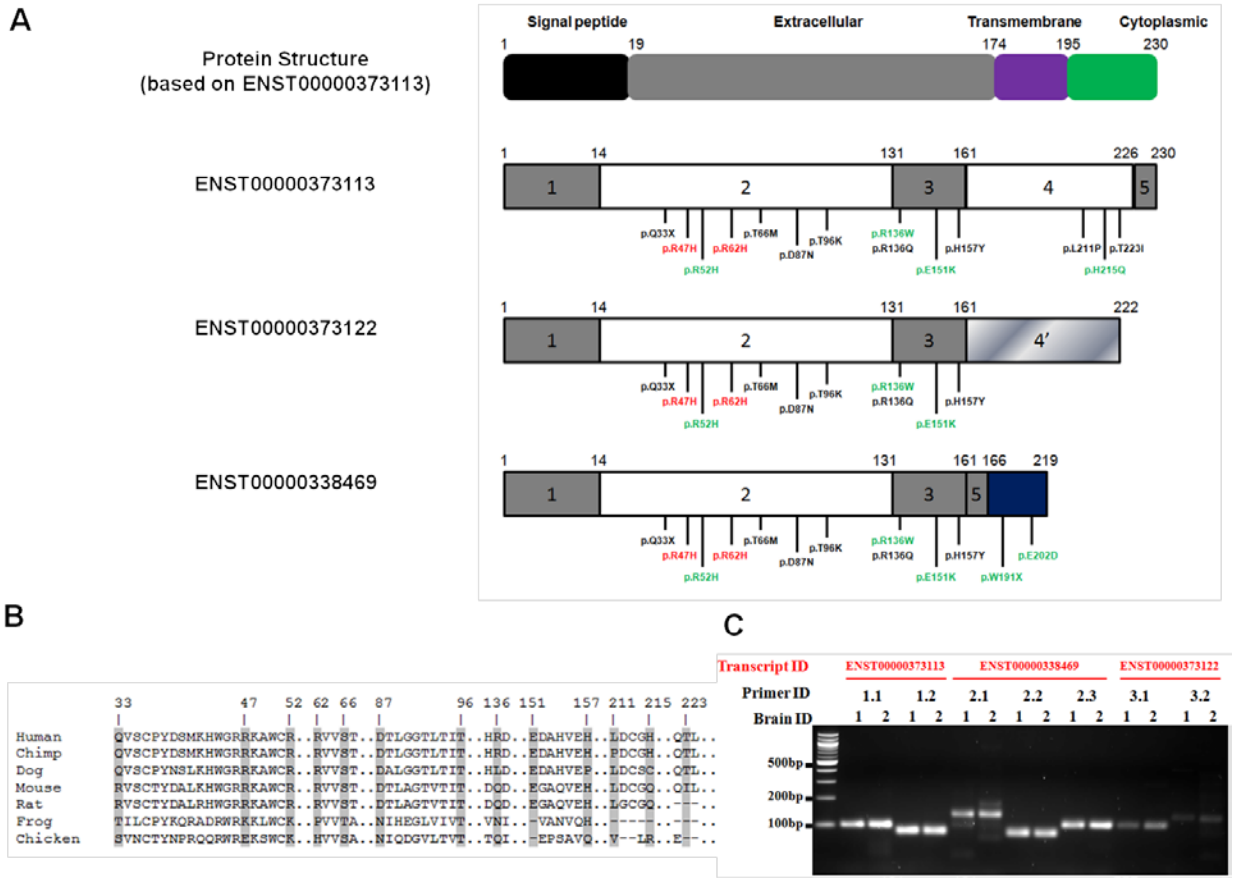


Figure 4. Schematic representation of protein structure for TREM2 and for the soluble form of TREM2, location of variants, protein conservation of the mutated positions, and the results of alternative splicing assays. (A) The top panel shows the protein structure of TREM2 (based on ENST00000373113), a type-I transmembrane receptor that is encoded by a gene containing 5 exons. The isoform ENST00000373122 encodes a different protein coding sequence after exon 3 (gradient fill rectangle) compared to ENST00000373113. The soluble form of TREM2 (ENST00000338469) lacks exon 4, which encodes the transmembrane domain, and contains a coding region after exon 5 (texture fill rectangle). Figures shown below include the structure of three different *TREM2* isoforms, the location of confirmed variants in the most common *TREM2* transcript (ENST00000373113), and the location of confirmed variants only in the *sTREM2* transcript (ENST00000338469). Most of the variants in the transmembrane form of TREM2 are located in the extracellular domain with three exceptions, located in the cytoplasmic tail. We identified two variants that are located near the C-terminus of the soluble form of TREM2. (B) The protein conservation analysis of confirmed *TREM2* variants. Variants are shown with an arrow identifying the corresponding amino acid position. Protein sequences were downloaded from UniProt. The entries used for each species are as follows: Q9NZC2 (Human), Q99NH8 (Mouse), D3ZZ89 (Rat), H2QSZ0 (Chimp), F7CW35 (Frog), Q2YHU4 (Chicken), and E2RP46 (Dog). (C) Results of alternative splicing validation. PCR was performed to amplify the cDNA of two AD cases (Brain ID =1 and 2) extracted from autopsy brain tissue obtained from the Knight-ADRC. ENST00000373113, ENST00000338469, and ENST00000373122 were amplified using 7 different primer pairs designed to specifically amplify one of the three transcripts (Primer ID=1.1 and 1.2 for ENST00000373113; Primer ID=2.1, 2.2 and 2.3 for ENST00000338469; Primer ID=3.1 and 3.2 for ENST00000373122). The amplicon length is 100 bp for 1.1, 84 bp for 1.2, 135 bp for 2.1, 81 bp for 2.2, 104 bp for 2.3, 103 bp for 3.1, and 127 bp for 3.2. The gel electrophoresis analysis clearly shows the presence of three distinct isoforms in the cDNA extracted from brains of two AD cases.

cortex of human brain (see **Figure 4C**). The variant p.W191X is predicted to result in a premature stop codon in the ENST00000338469 transcript; however, the impact of this variant on AD pathogenesis remains unknown, due to the rarity of the allele (1/1,816 cases).

Table 6. Segregation of rare variants in available family members

			Affected		Unaffected		
Variant	# of families	Status	Carriers	Non-carriers	Carriers	Non-carriers	P
p.Q33X	1	Nasu Hakola variant	0	2	1	1	1
			NA	75±0	56	58	
p.R47H	13	Confirmed risk factor	15	4	4	7	$4.65\times10^{-2*}$
			70.8±7.9	72.3±10.7	74±4.5	78.7±7.7	
p.R62H	21	Previously identified	18	11	11	28	$6.87\times10^{-3*}$
			71.8±5.7	67.9±7.7	71.5±11.4	71.4±8.8	
p.D87N	2	Previously identified	2	0	2	3	4.29×10^{-1}
			81.8±1.4	NA	71±1.4	74.3±9.0	
p.H157Y	1	Previously identified	2	1	0	0	1
			67±5.7	85	NA	NA	
p.H215Q	1	Novel variant	1	1	0	0	1
			79	70	NA	NA	
p.T223I	1	Novel variant	1	0	1	0	1
			62	NA	73	NA	

Family-based association analysis was performed for variants when samples from family members of the probands were available. The same variant was genotyped to test whether the rare allele is associated with disease status. Variants, number of families performed, variant type, the number of affected carriers, non carriers and un-affected carriers, non-carriers, the average and standard deviation of age at onset (years) for the affected individuals and the average and standard deviation of age at last assessment (years) for the unaffected individuals were shown. All of the confirmed carriers only carried one rare allele. Novel variants were not identified in previous studies[12,13]. A two-tailed Fisher's exact test was used to determine evidence of segregation for each variant. * denotes significant association. NA represents not applicable.

2.4.3 GWAS-identified genes sequencing

Deep re-sequencing in the coding regions of GWAS-identified genes was undertaken in the same AD cases and cognitively-normal elderly controls as *TREM2* sequencing (see **Tables 1 and 2**). We have performed direct genotyping and validated 90 non-synonymous variants in these genes (40 in *ABCA7*, 4 in *BIN1*, 7 in *CD2AP*, 7 in *CD33*, 7 in *CLU*, 11 in *EPHA1*, 4 in *MS4A4A*, and 10 *PICALM*, **Table 7**), 66 of which were not annotated in the EVS database. I did not validate any variants in *CR1* due to the fact that this genomic region is highly repetitive, which leads to low specificity and a below-average coverage. SIFT and PolyPhen algorithms both provide predictions for the possible impact of an amino acid change on the protein structure, which suggests that 56 out of

90 (62.2%) non-synonymous variants are damaging. Additionally, 57 out of 90 (63.3%) variants have a GERP conservation score equal or greater than two (the greater the score the greater the level of evolutionary constraint inferred to be adding on the site). Therefore, based on protein change and evolutionary constraint predictions, a large proportion of these validated variants are likely to be functional. We will finish direct genotyping and segregation analyses in the coming months and SNP-level and gene-level association testing will be conducted to determine the effects on AD risk.

Table 7. Confirmed variants in GWAS-identified genes

Gene	AA Change	SNP	CHR:BP (hg19)	Function	PolyPhen	SIFT	GERP Score	In EVS?
ABCA7	p.M8I	NA	19:1041384	missense	DAMAGING	TOLERATED	4.78	No
ABCA7	p.L9M	NA	19:1041385	missense	DAMAGING	DAMAGING	-3.59	No
ABCA7	p.L101R	rs201665195	19:1041971	missense-near-splice	DAMAGING	DAMAGING	4.07	Yes
ABCA7	p.A172T	NA	19:1042760	missense	TOLERATED	TOLERATED	0.615	No
ABCA7	p.D245Y	rs377552810	19:1043193	missense	DAMAGING	DAMAGING	-1.93	Yes
ABCA7	p.G272R	NA	19:1043356	missense	DAMAGING	TOLERATED	2.2	No
ABCA7	p.R281H	rs375028339	19:1043384	missense	DAMAGING	DAMAGING	0.571	Yes
ABCA7	p.W284X	NA	19:1043393	stop-gained	NA	NA	4.25	No
ABCA7	p.A676T	rs59851484	19:1047336	missense	DAMAGING	TOLERATED	4.88	Yes
ABCA7	p.E708X	NA	19:1047506	stop-gained	NA	NA	3.64	No
ABCA7	p.R907Q	NA	19:1051189	missense	TOLERATED	TOLERATED	4.37	No
ABCA7	p.A913V	NA	19:1051207	missense	TOLERATED	TOLERATED	-0.686	No
ABCA7	p.A914S	rs111940546	19:1051209	missense	TOLERATED	TOLERATED	3.4	Yes
ABCA7	p.V960A	rs146811533	19:1051502	missense	DAMAGING	TOLERATED	4.41	Yes
ABCA7	p.R1026C	rs141322593	19:1052054	missense	DAMAGING	DAMAGING	2.14	Yes
ABCA7	p.T1060A	NA	19:1052243	missense	TOLERATED	TOLERATED	-6.12	No
ABCA7	p.K1179N	NA	19:1054069	missense	TOLERATED	TOLERATED	-0.7	No
ABCA7	p.R1210Q	NA	19:1054243	missense	TOLERATED	TOLERATED	-2.97	No
ABCA7	p.R1218H	NA	19:1054267	missense	TOLERATED	TOLERATED	-5.43	No
ABCA7	p.R1236C	NA	19:1054320	missense	DAMAGING	TOLERATED	2.59	No
ABCA7	p.R1385Q	rs144595576	19:1055299	missense	DAMAGING	TOLERATED	3.68	Yes
ABCA7	p.G1415R	rs374185879	19:1056069	missense	DAMAGING	DAMAGING	2.75	Yes
ABCA7	p.R1434C	rs137888610	19:1056126	missense	DAMAGING	DAMAGING	0.991	Yes
ABCA7	p.V1487G	rs200825702	19:1056372	missense	TOLERATED	DAMAGING	3.29	No
ABCA7	p.R1489X	NA	19:1056377	stop-gained	NA	NA	-0.428	No
ABCA7	p.R1505H	rs113269196	19:1056426	missense	TOLERATED	TOLERATED	-3.6	Yes
ABCA7	p.V1598M	NA	19:1057340	missense	DAMAGING	DAMAGING	3.65	No
ABCA7	p.L1625M	NA	19:1057421	missense	DAMAGING	DAMAGING	2.61	No
ABCA7	p.W1628S	NA	19:1057916	missense	DAMAGING	DAMAGING	4.22	No
ABCA7	p.M1634V	NA	19:1057933	missense	DAMAGING	DAMAGING	3.2	No

ABCA7	p.M1634V	NA	19:1057933	missense	DAMAGING	DAMAGING	3.2	No
ABCA7	p.E1679X	NA	19:1058154	stop-gained	NA	NA	-0.843	No
ABCA7	p.N1797H	NA	19:1058928	missense	TOLERATED	DAMAGING	3.14	No
ABCA7	p.M1807I	NA	19:1059042	missense	TOLERATED	TOLERATED	-6.97	No
ABCA7	p.L1813W	rs200308069	19:1059059	missense	DAMAGING	DAMAGING	4.45	No
ABCA7	p.R1932C	rs114787084	19:1063624	missense	DAMAGING	DAMAGING	0.096	Yes
ABCA7	p.R1976S	NA	19:1063837	missense	DAMAGING	DAMAGING	2.3	No
ABCA7	p.C1992F	NA	19:1064183	missense	DAMAGING	DAMAGING	3.25	No
ABCA7	p.L2022P	NA	19:1064950	missense	DAMAGING	DAMAGING	4.24	No
ABCA7	p.A2075V	NA	19:1065109	missense	TOLERATED	TOLERATED	3.73	No
BIN1	p.S392I	NA	2:127808487	missense-near-splice	DAMAGING	DAMAGING	3.92	No
BIN1	p.T307M	NA	2:127816621	missense	TOLERATED	TOLERATED	4.01	No
BIN1	p.G244R	NA	2:127819725	missense	DAMAGING	TOLERATED	5.13	No
BIN1	p.G238S	rs372072916	2:127819743	missense	TOLERATED	TOLERATED	2.75	Yes
CD2AP	p.G47R	NA	6:47471150	missense	DAMAGING	DAMAGING	4.69	No
CD2AP	p.Q207R	NA	6:47541878	missense	DAMAGING	TOLERATED	6.08	No
CD2AP	p.S233G	NA	6:47541955	missense	TOLERATED	TOLERATED	3.57	No
CD2AP	p.G266S	NA	6:47544326	missense	TOLERATED	TOLERATED	2.26	No
CD2AP	p.P341Q	NA	6:47548613	missense	DAMAGING	TOLERATED	5.42	No
CD2AP	p.K436R	NA	6:47567069	missense	DAMAGING	TOLERATED	5.92	No
CD2AP	p.S510C	NA	6:47574012	missense-near-splice	DAMAGING	DAMAGING	4.53	No
CD33	p.R91C	NA	19:51728707	missense	TOLERATED	DAMAGING	-6.9	No
CD33	G156Tfs*5	NA	19:51729103	frameshift	NA	NA	NA	Yes
CD33	p.G156S	rs201342074	19:51729106	missense	DAMAGING	DAMAGING	3.08	Yes
CD33	p.A183V	NA	19:51729188	missense	DAMAGING	DAMAGING	0.739	No
CD33	p.F243L	rs11882250	19:51729594	missense	TOLERATED	TOLERATED	2.88	Yes
CD33	p.S324G	rs200505271	19:51742818	missense	TOLERATED	TOLERATED	-1.3	No
CD33	p.H339R	rs374286140	19:51742864	missense	DAMAGING	DAMAGING	1.47	Yes
CLU	p.T445A	NA	8:27456140	missense	TOLERATED	TOLERATED	-4.69	No
CLU	p.E387A	NA	8:27457457	missense	DAMAGING	DAMAGING	5.62	No
CLU	p.V385I	rs368729462	8:27457464	missense	TOLERATED	TOLERATED	-11.2	Yes
CLU	p.T203I	rs41276297	8:27462662	missense	TOLERATED	TOLERATED	1.03	Yes
CLU	p.R234H	rs201670453	8:27462725	missense	DAMAGING	TOLERATED	0.175	Yes
CLU	p.D226N	NA	8:27462750	missense	DAMAGING	TOLERATED	3.76	No
CLU	p.L58V	NA	8:27468073	missense	DAMAGING	TOLERATED	2.37	No
EPHA1	p.R891Q	NA	7:143090788	missense	DAMAGING	DAMAGING	5.16	No
EPHA1	p.L820Q	NA	7:143091330	missense	DAMAGING	DAMAGING	4.58	No
EPHA1	p.R801Q	NA	7:143091387	missense	DAMAGING	DAMAGING	4.5	No
EPHA1	p.R534Q	NA	7:143095027	missense	DAMAGING	TOLERATED	4.7	No
EPHA1	p.V514I	NA	7:143095088	missense	TOLERATED	DAMAGING	2.72	No
EPHA1	p.A477V	rs373049955	7:143095448	missense	TOLERATED	TOLERATED	0.833	Yes
EPHA1	p.P460L	rs202178565	7:143095499	missense	DAMAGING	TOLERATED	5.08	No
EPHA1	p.G408S	NA	7:143095808	missense	DAMAGING	TOLERATED	5.09	No

EPHA1	p.D363E	NA	7:143095941	missense	DAMAGING	DAMAGING	1.46	No
EPHA1	p.Q347R	NA	7:143095990	missense	TOLERATED	DAMAGING	5.09	No
EPHA1	p.R82H	NA	7:143098604	missense	DAMAGING	DAMAGING	4.99	No
MS4A4A	M26K	rs368779495	11:60059733	missense	TOLERATED	TOLERATED	-5.16	No
MS4A4A	M76K	NA	11:60064695	missense	DAMAGING	DAMAGING	1.62	No
MS4A4A	A88T	rs183720563	11:60064730	missense	TOLERATED	TOLERATED	-8.13	No
MS4A4A	P96L	NA	11:60064755	missense	TOLERATED	TOLERATED	2.99	No
PICALM	p.M26K	rs368779495	11:60059733	missense	TOLERATED	TOLERATED	-5.16	Yes
PICALM	p.M76K	NA	11:60064695	missense	DAMAGING	DAMAGING	1.62	No
PICALM	p.A88T	rs183720563	11:60064730	missense	TOLERATED	TOLERATED	-8.13	No
PICALM	p.P96L	NA	11:60064755	missense	TOLERATED	TOLERATED	2.99	No
PICALM	p.G518S	rs369070232	11:85692249	missense	DAMAGING	NA	5.27	Yes
PICALM	p.P411A	rs34013602	11:85707896	missense	DAMAGING	NA	1.57	Yes
PICALM	p.L249I	NA	11:85722091	missense	TOLERATED	NA	2.63	No
PICALM	p.P72T	NA	11:85742568	missense	DAMAGING	NA	5.52	No
PICALM	p.L42V	NA	11:85779699	missense	TOLERATED	NA	4.61	No
PICALM	p.P36L	NA	11:85779716	missense	DAMAGING	NA	4.73	No

* NA represents not applicable. SIFT (http://sift.jcvi.org/) and PolyPhen2 (http://genetics.bwh.harvard.edu/pph2/) algorithms were used to predicted the functional impact of these variants on protein structure. The presence of these variants in the Exome Variant Server (http://evs.gs.washington.edu/EVS/) was also annotated for reference.

DISCUSSION

This work provides extensive genetic evidence that *PLD3* and *TREM2* are AD risk genes: genome-wide significant evidence that rare variants in *PLD3* and *TREM2* increase risk for AD in multiple data sets and two populations. Moreover, our later functional studies confirm that PLD3 affects APP processing in a manner that is consistent with increased risk of AD[14]. The studies on *TREM2*[12,13] and this project indicate that next-generation sequencing technology can help uncover additional low-frequency and rare variants associated with AD.

TREM2 is a type one transmembrane receptor protein expressed on myeloid cells including microglia, monocyte-derived dendritic cells, osteoclasts and bone-marrow derived macrophages[85,86]. Additionally, protein expression of TREM2 in neurons has been reported[87]. TREM2 transduces its intracellular signaling through DAP12 (TYROBP)[85,86]. Although the natural ligands of TREM2 remain unknown, upon ligand binding, TREM2 associates with DAP12 to mediate downstream signaling. In the brain, TREM2 is primarily expressed on microglia and has been shown to control two signaling pathways: regulation of phagocytosis and suppression of inflammatory reactivity[88-90]. A previous study used microarray and laser microdissection of beta amyloid (Aβ) plaque-associated

areas in an animal model of AD and found that *TREM2* is differentially expressed in Aβ plaque-associated versus Aβ plaque-free tissue[91].

Homozygous loss-of-function mutations in *TREM2* were initially associated with an autosomal recessive form of early-onset dementia, Polycystic lipomembranous osteodysplasia with sclerosing leukoencephalopathy (PLOSL), also known as Nasu-Hakola disease, in Swedish and Norwegian families[92]. Subsequently, mutations in *TREM2* were found worldwide in PLOSL patients from different countries and ethnic origins[92-95]. PLOSL patients carrying different *TREM2* variants exhibit a similar clinical phenotype with respect to the neurologic and skeletal abnormalities[92-95]. The clinical spectrum of disease associated with *TREM2* variants was expanded after the identification of three patients from a Lebanese family carrying mutations in *TREM2* that exhibited early-onset dementia without skeletal symptoms (bone cysts)[124]. Additional *TREM2* variants were also found in three Turkish probands with frontotemporal-like dementia without any bone-associated symptoms[12].

Recently, two independent studies reported that a heterozygous rare variant in *TREM2* p.R47H is significantly associated with AD, with an odds ratio similar to that of an individual carrying one *APOE* ε4 allele[12,13]. Subsequently, the association of p.R47H with AD risk was replicated in Spanish and French populations[125,126]. Several studies also found that *TREM2* variants are associated with Parkinson's disease, frontotemporal dementia (FTD) and amyotrophic lateral sclerosis[98,127-136]. Thus, *TREM2* variants produce a heterogeneous disease state that ranges in clinical phenotypes from skeletal abnormalities to neurodegeneration.

To our knowledge, this study is the largest deep re-sequencing study to date which aims to identify novel rare coding variants in *TREM2.* Recently, a study has sequenced *TREM2* coding regions in a Belgian population and found additional coding variants in *TREM*[96]. Even though an enrichment of *TREM2* variants in both AD and FTD patients compared to controls was reported, none of the rare variants were individually significant[96]. In our study, sixteen *TREM2* rare coding variants were observed in European Americans*,* including two variants (p.R47H and p.R62H) that were significantly associated with AD risk and six novel variants that were not found in previous studies[12,13,96]. The minor alleles of p.R47H ($p=9.17\times10^{-4}$; OR=2.63 [1.44~4.81]) and p.R62H ($p=2.36\times10^{-4}$; OR=2.36 [1.47~3.80]) were associated with increased AD risk after multiple test correction. After adjusting for *APOE* ε2 and ε4 alleles in the logistic regression, the association for p.R47H and p.R62H only changed slightly and remained significant ($p=5.91\times10^{-3}$; OR=2.48 [1.30~4.75] for p.R47H; $p=8.08\times10^{-4}$; OR=2.36 [1.43~3.90] for

p.R62H; **Table S5**), which suggests that p.R47H and p.R62H affect AD risk independently of *APOE* ε2 and ε4 alleles. The gene-based test for *TREM2* remained highly significant even after dropping p.R47H, suggesting that additional variants in *TREM2* influence AD risk. After excluding both p.R47H and p.R62H, the gene-based p-value is 0.09 (see **Table S3**), suggesting that most of the statistical significance for the gene-based association comes from these two variants. However, the OR for the gene-based analyses when these two variants were removed was 2.95 (see **Table S3**), suggesting that additional very- low-frequency variants may have a larger effect size for AD risk than p.R47H and p.R62H. This observation is also supported by the fact that nine out of sixteen *TREM2* variants are identified in 13 AD cases and no controls. The lack of association after excluding p.R47H and p.R62H is likely due to the rarity of the other variants and a lack of statistical power. Moreover, we did not identify any rare variants significantly associated with AD risk in African Americans as the sample size was too small. Additional AA samples are required to help determine whether *TREM2* variants influence the susceptibility to AD in the AA population. Recently, a study has shown that loss of one copy of TREM2 had no effect on Aβ biology but it altered the morphological phenotype of plaque-associated microglia[137]. In order to better understand the mechanism by which these *TREM2* variants affect AD pathogenesis, functional studies are urgently needed.

We also evaluated the impact on the analysis of excluding individuals who could not be included in the PCA owing to a lack of GWAS data. After removing individuals without GWAS data, a total of 1,724 AD cases and 1,437 controls were included in the analyses. The single-variant association changed slightly but still surpassed the multiple test threshold: p.R47H ($p=2.99\times10^{-3}$; OR=2.53 [1.35~4.76]; **Table S6**) and p.R62H ($p=3.25\times10^{-4}$; OR=2.54 [1.49~4.35]; **Table S6**). The gene-based association for *TREM2* reduced slightly ($P_{SKAT-O}=6.81\times10^{-6}$; OR=2.56 [1.74~3.83]; **Table S6**) and was no longer genome-wide significant (2.5×10^{-6}). These results suggest that the SNP-level and gene-level significant associations using data including individuals with/without GWAS data are not false positives due to population substructure.

We also identified two variants, p.W191X and p.E202D, which are predicted to be located only in the coding region of the shortest transcript (ENST00000338469), encoding a soluble form of TREM2 (sTREM2) according to Ensembl. A soluble isoform of TREM2 protein has been described as a transcript that encodes a soluble form of *TREM2* [138,139]. Extracellular TREM2 could be derived from the sTREM2 alternative transcript, a posttranslational cleavage product, or a combination of both. A previous study has described the presence of soluble

TREM2 protein in human cerebrospinal fluid (CSF) and serum[140]. Furthermore, CSF levels of sTREM2 were found to be elevated in patients with multiple sclerosis[140]. Experimental data suggests that soluble TREM1 results from sequential cleavage of the transmembrane form of this related protein[140,141] and so TREM2 may be cleaved in a similar fashion. In this study, we showed the presence of cDNA corresponding to the predicted sTREM2 transcript in brain tissue from AD cases (see **Figure 4C**). Together, this study provides evidence of the presence of sTREM2 mRNA in the human brain. The p.W191X variant introduces a nonsense mutation into this sTREM2 transcript at codon 191. It is unclear whether this would result in a truncated protein or removal of the mutant mRNA by nonsense mediated decay.

Despite that fact that GWAS have been very successful in pinpointing the regions of the human genome that associate with AD, the underlying causal variants and genes remain unclear as the LD with other SNPs prevents the identification of the functional variants and most of these susceptibility loci are located in gene-dense regions. For instance, a recent meta-GWAS analysis discovered an intergenic signal for rs981040, located approximately 5.5 kb downstream from *TREML2* and 24 kb upstream from *TREM2*[8] but it was impossible to determine whether this GWAS signal is driven by *TREM2* or *TREML2* due to its study design (a meta analysis), leaving the underlying causal variants and genes unknown. A recent study performed bioinformatic analyses and predicted that several top GWAS SNPs may affect protein binding or mRNA expression of a gene target using data from the ENCODE Project Consortium, but evidence of regulatory function are primarily in blood or cancer cells[142]. Recently, two independent studies reported that gene expression of some GWAS-identified genes were altered in AD brains and some of these GWAS-identified genes harbor cis-variants affecting gene expression in human brains[143,144]. Nonetheless, another independent study assessed the regional expression in the human brain but did not find any of these GWAS loci had eQTLs explaining the association[109]. Even though most of these findings did not yield a very strong association and not yet replicated in other studies, these results suggested that some of these GWAS-identified variants and genes can affect AD pathogenesis through regulating gene expression in the human brains[109,143,144].

Alternatively, Holton et al. also found that coding variability may explain the *ABCA7* association but common coding variability does not explain any of the other loci, which indicates rare coding variants within these GWAS loci can also contribute to AD risk[109]. Our recent work has performed deep re-sequencing of *APP*, *PSEN1*,

and *PSEN2,* three genes cause early-onset familial AD, in LOAD families in order to identify rare coding variants[15,16]. We found that the impact of rare coding variants in *APP*, *PSEN1*, and *PSEN2* on LOAD risk is more than previously estimated[15,16]. These findings suggest that rare variants in these genes (*APP*, *PSEN1*, and *PSEN2*) can account for a portion of the LOAD cases that cannot be detected by GWAS. In this study, we have confirmed 90 rare coding variants, 66 of which were not present in the EVS database. 56 out of 90 (62.2%) confirmed variants are predicted to be damaging in terms of changes to protein structures and evolutionary constraint, which recapitulates that rare coding variants within these GWAS loci can play a key role in AD risk association.

In conclusion, to our knowledge, this study is the largest deep re-sequencing study to date which aims to uncover novel rare coding variants in *PLD3*, *TREM2*, and GWAS-identified genes. This study demonstrated that there multiple rare coding variants exist in *PLD3* and *TREM2* and that cumulatively these variants are associated with AD risk. The analysis of GWAS genes also identified many rare and novel coding variants, which need to be tested for association with disease risk. Together, these data support the paradigm that both common, low penetrant and rarer, higher penetrant alleles exist in the same gene in common diseases (ex: coronary heart disease and type 1 diabetes)[80,145]. These findings can provide better targets for downstream functional studies and may eventually lead to effective therapies.

SUPPLEMENTAL TABLES

Table S1. Information for primers designed for TREM2 alternative splicing assays

Primer ID	Amplicon Length	Amplified Transcript ID	Forward Primer	Reverse Primer
1.1	100	ENST00000373113	5'-GCATCTCCAGGAGCCTCT-3'	3'-CTGGCTGCTAGAATCTTGATGAG-5'
1.2	84	ENST00000373113	5'-GATGCTGGAGATCTCTGGTT-3'	3'-CAAGAGGCTCCTGGAGATG-5'
2.1	135	ENST00000338469	5'-AGCATCTCCAGGGCTGA-3'	3'-CTCTTGCCAGAGCAGAACAA-5'
2.2	81	ENST00000338469	5'-GATGCTGGAGATCTCTGGTT-3'	3'-CTCAGCCCTGGAGATGC-5'
2.3	104	ENST00000338469	5'-ATCTCCAGGGCTGAGAGACAC-3'	3'-GGTGGCCAAGTGGCAAGTAT-5'
3.1	103	ENST00000373122	5'-CATCTCCAGGCCATCTCAAG-3'	3'-AGGAGGAGAAGGATGGAAGT-5'
3.2	127	ENST00000373122	5'-AAGGTCCTGGTGGAGGT-3'	3'-CTTGAGATGGCCTGGAGATG-5'

Primer IDs, amplicon lengths, corresponding transcript IDs, forward- and reverse-primer sequences were listed. Each primer set was uniquely designed to amplify the corresponding transcript. The PrimeQuest Design Tool (Integrated DNA Technology) was used to design the primers.

Table S2. Comparison of the gene-based analysis including all *PLD3* coding variants or only variants predicted to be deleterious

	Benign + deleterious		Only deleterious	
	P	OR (CI)	P	OR (CI)
All variants	1.44×10^{-11}	2.75 (2.05-3.68)	2.52×10^{-12}	2.86 (2.10-3.88)
Excluding p.V232M	1.58×10^{-8}	2.58 (1.87-3.57)	2.95×10^{-8}	2.54 (1.81-3.57)
Excluding p.A442 and p.V232M	1.61×10^{-3}	2.86 (1.62-5.06)	5.88×10^{-5}	3.20 (1.59-6.45)

Gene-based analyses were performed by SKAT-O. Variants that were predicted to be benign by both SIFT and Polyphen were removed for the second analysis

Table S3. Gene-based analyses for *TREM2*

	AD Cases	Controls		
All variants including p.R47H			P	OR
Non- Carriers	1,943	1,603	5.37×10^{-7}	2.55 (1.80-3.67)
Carriers	139	45		
Excluding p.R47H				
Non- Carriers	1,943	1,603	7.72×10^{-5}	2.47 (1.62-3.87)
Carriers	93	31		
Excluding p.R47H and p.R62H				
Non- Carriers	1,943	1,603	0.09	2.95 (1.23-8.09)
Carriers	25	7		

Results of SKAT-O analyses including all the coding variants with/without p.R47H and without p.R47H and p.R62H were presented.

Table S4. *TREM2* segregation data for each independent family with available DNA samples

Variant	Anonymous family ID	Status	Affected		Unaffected	
			Carriers	Non-carriers	Carriers	Non-carriers
p.Q33X	1	Nasu Hakola variant	0	2	1	1
			NA	75±0	56	58
p.R47H	2	Confirmed risk factor	1	1	0	0
			80	60	NA	NA
p.R47H	3	Confirmed risk factor	1	0	0	1
			81	NA	NA	84
p.R47H	4	Confirmed risk factor	0	1	1	0
			NA	85	79	NA
p.R47H	5	Confirmed risk factor	1	0	1	0
			68	NA	68	NA
p.R47H	6	Confirmed risk factor	1	0	0	1
			67	NA	NA	68
p.R47H	7	Confirmed risk factor	1	0	1	0
			66	NA	74	NA
p.R47H	8	Confirmed risk factor	2	0	0	1
			61±1.4	NA	NA	74
p.R47H	9	Confirmed risk factor	1	1	0	1
			80	68	NA	77
p.R47H	10	Confirmed risk factor	2	0	0	1
			71.5±16.3	NA	NA	73
p.R47H	11	Confirmed risk factor	1	1	0	1
			78	76	NA	89
p.R47H	12	Confirmed risk factor	2	0	0	0
			67.5±3.5	NA	NA	NA
p.R47H	13	Confirmed risk factor	0	0	1	1
			NA	NA	75	86
p.R47H	14	Confirmed risk factor	2	0	0	0
			71±1.4	NA	NA	64
p.R62H	15	Previously identified	1	0	0	3
			83	NA	NA	78.3±3.5
p.R62H	16	Previously identified	1	1	0	1
			69	63	NA	55
p.R62H	17	Previously identified	2	0	0	1
			70±0	NA	NA	72
p.R62H	18	Previously identified	0	0	1	2
			NA	NA	83	76±4.2

p.R62H	19	Previously identified	1	0	0	4
			70	NA	NA	74.8±15.5
p.R62H	20	Previously identified	0	3	1	8
			NA	76.7±5.9	91	65.5±4.2
p.R62H	21	Previously identified	0	2	1	0
			NA	61±11.3	77	NA
p.R62H	22	Previously identified	1	0	1	1
			65	NA	53	73
p.R62H	23	Previously identified	2	0	0	1
			72.5±3.5	NA	NA	80
p.R62H	24	Previously identified	0	1	1	1
			NA	70	79	86
p.R62H	25	Previously identified	2	0	0	1
			71±8.5	NA	NA	68
p.R62H	26	Previously identified	1	1	0	0
			72	65	NA	NA
p.R62H	27	Previously identified	1	0	1	0
			79	NA	70	NA
p.R62H	28	Previously identified	1	0	1	0
			65	NA	58	NA
p.R62H	29	Previously identified	1	0	0	1
			76	NA	NA	64
p.R62H	30	Previously identified	1	0	0	1
			62	NA	NA	71
p.R62H	31	Previously identified	1	0	1	0
			79	NA	68	NA
p.R62H	32	Previously identified	1	0	1	0
			71	NA	67	NA
p.R62H	33	Previously identified	1	0	0	1
			75	NA	NA	78
p.R62H	34	Previously identified	0	1	1	0
			NA	66	79	NA
p.R62H	35	Previously identified	0	2	1	2
			NA	65.5±4.9	62	70.5±3.5
p.D87N	36	Previously identified	2	0	0	1
			81±1.4	NA	NA	64
p.D87N	37	Previously identified	0	0	2	2
			NA	NA	71±1.4	79.5±0.7
p.H157Y	38	Previously identified	2	1	0	0
			67±5.7	85	NA	NA
p.H215Q	39	Novel variant	1	1	0	0

			79	70	NA	NA
p.T223I	40	Novel variant	1	0	1	0
			62	NA	74	0

Variants, anonymous family ID, variant status, the number of affected carriers, non-carriers and un-affected carriers, non-carriers, the average and standard deviation of age at onset (years) for the affected individuals and the average and standard deviation of age at last assessment for the unaffected individuals were listed.

Table S5. Unadjusted and adjusted analyses for *TREM2* p.R47H and p.R62H variants

	Unadjusted*		Adjusted[&]	
Variant	P	OR (95% CI)	P	OR(95% CI)
p.R47H	9.17×10^{-4}	2.63 (1.44-4.81)	5.91×10^{-3}	2.48 (1.30-4.75)
p.R62H	2.36×10^{-4}	2.36 (1.47-3.80)	8.08×10^{-4}	2.36 (1.43-3.90)

*A Fisher's exact test was used to evaluate the case-control association with p.R47H and p.R62H. [&] Logistic regression was used to evaluate case-control association adjusting for *APOE* ε2 and ε4 status. Analyses were performed using default commands in PLINK[122].

Table S6. Rare *TREN2* variant association in sequenced samples with GWAS data*

AA Change	SNP	CHR:BP (hg19)	AD Cases		Controls		P[!]	OR (95% CI)
			No. of Cases	No. of Carriers	No. of Controls	No. of Carriers		
All variants							6.81×10^{-6}[&]	2.56 (1.74-3.83)
p.Q33X	rs104894002	6:41129295	1700	1	1406	1	1	0.83 (0.05-13.33)
p.R47H	rs75932628	6:41129252	1701	39	1414	13	2.99×10^{-3}	2.53 (1.35-4.76)
p.R52H	rs374851046	6:41129237	1724	1	1433	0	1	NA
p.R62H	rs143332484	6:41129207	1700	55	1415	19	3.25×10^{-4}	2.54 (1.49-.4.35)
p.T66M	rs201258663	6:41129195	1701	1	1419	0	1	NA
p.D87N	rs142232675	6:41129133	1701	7	1417	3	0.36	1.96 (0.51-7.59)
p.T96K	rs2234253	6:41129105	1696	0	1406	2	0.21	0
p.R136W	NA	6:41127606	1671	1	1364	0	1	NA
p.R136Q	rs149622783	6:41127605	1697	1	1419	0	1	NA
p.E151K	rs79011726	6:41127561	1724	0	1433	1	0.46	0
p.H157Y	rs2234255	6:41127543	1701	1	1406	0	1	NA
p.W191X	rs2234258	6:41126429	1513	1	1281	0	1	NA
p.E202D	NA	6:41126395	1724	1	1433	0	1	NA
p.L211P	rs2234256	6:41126655	1693	0	1402	2	0.21	0
p.H215Q	NA	6:41126642	1669	1	1363	0	1	NA
p.T223I	rs138355759	6:41126619	1724	2	1433	0	0.5	NA

A total of 1,724 AD cases and 1,437 controls were used for analysis.*NA represents not applicable. [!]The Fisher's exact test was used to calculate the p values using the default commands in PLINK. [&]This p value summarizes the gene-based association of the identified SNP set and was estimated using the SKAT-O algorithm.

SUPPLEMENTAL FIGURES

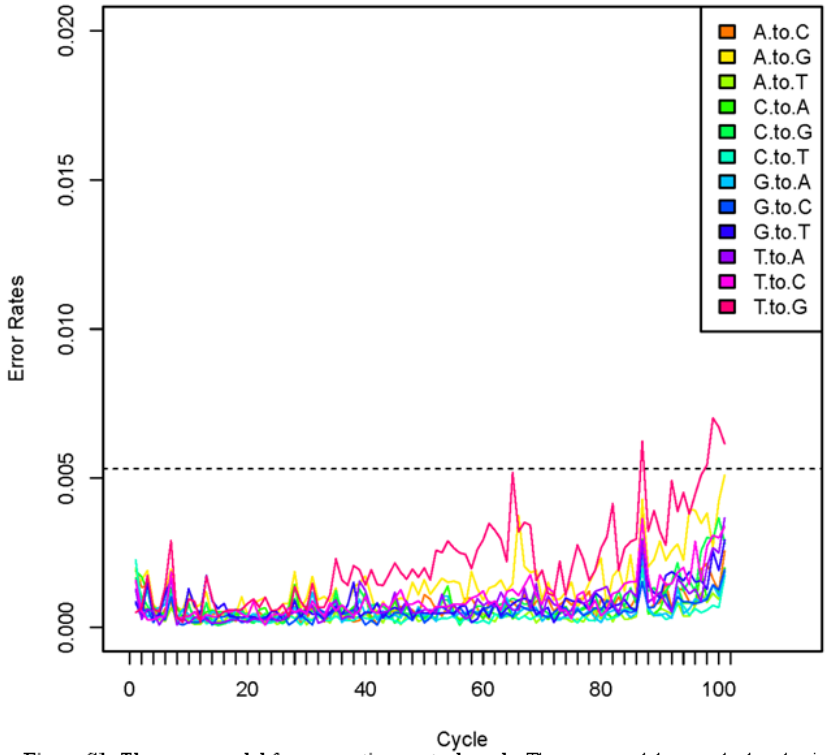


Figure S1: The error model from negative control reads. The error model was calculated using the negative control reads with a read length 102 bp (cycle) generated from Illumina HiSeq 2500. The x-axis represents the sequencing cycle and the y-axis represents the negative-control error rate. The dash line indicates the error-rate cutoff defined by one mismatch divided by the number of chromosome (which is 2 times the number of individuals in a pool). The sequencing cycle with an error rate exceeding the cutoff will be excluded from the SPLINTER analyses. Based on this error model, the first 98 cycles were used for SPLINTER analyses while skipping cycles 65 and 87.

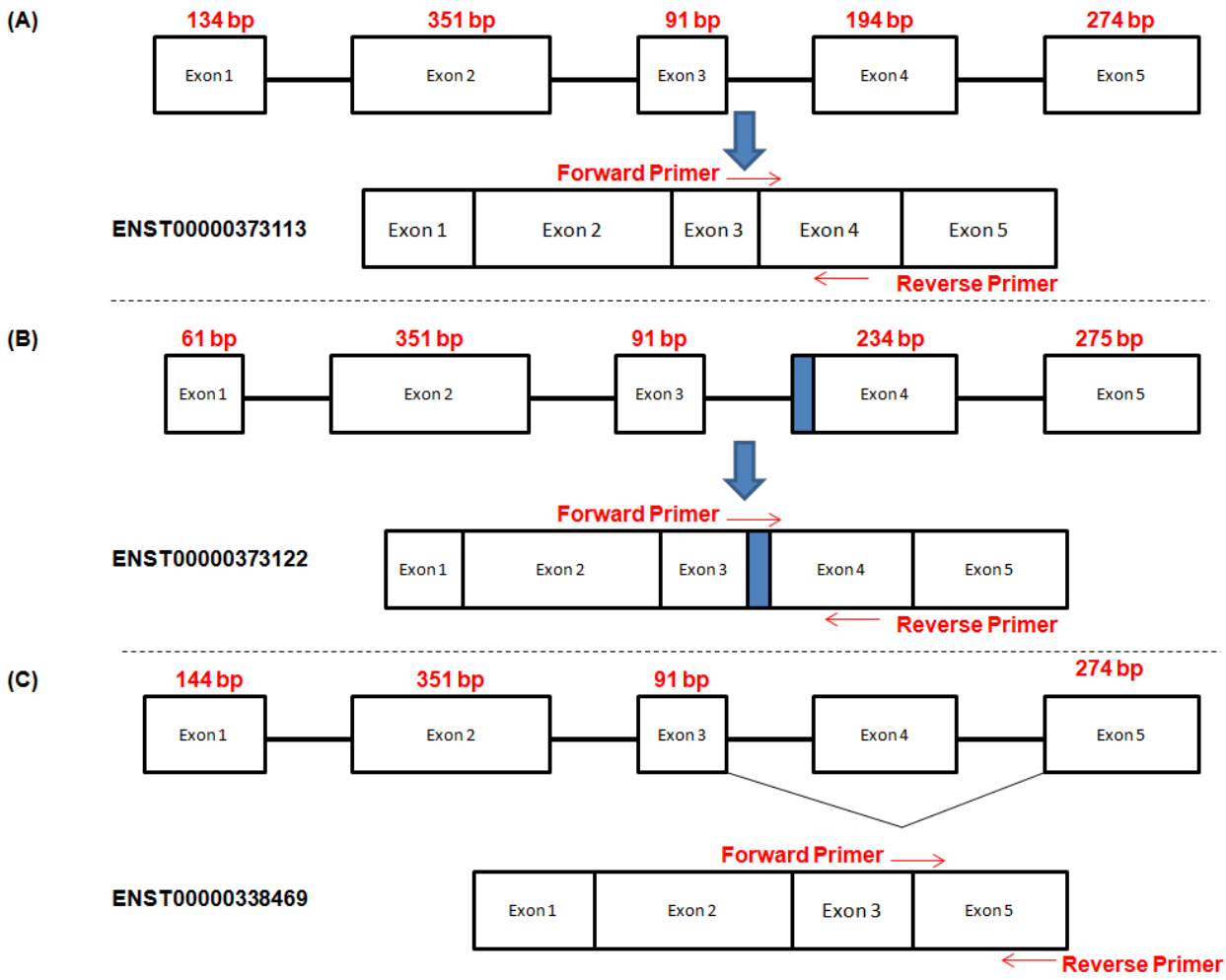


Figure S2. DNA and mRNA transcripts of TREM2. The top panel of (A)-(C) represents the genomic structure of each transcript and the bottom panel of (A)-(C) represents the corresponding mRNA transcript. Red arrows indicate the location where the isoform-specific primers are located. The length of each exon is indicated above each exon box. ENST00000373113 is the longest transcript of TREM2, which contains 5 exons. The designed forward and reverse primers for ENST00000373113 are located in exon3-exon4 junction and exon 4. ENST00000373122 has an additional proportion of exon4 relative to ENST00000373113. The designed forward and reverse primers for ENST00000373122 are located in exon3-exon4 junction and exon 4. ENST00000338469 lacks the exon4 which encodes the transmembrane domain. The forward and reverse primers for amplifying ENST00000338469 are located in exon3-exon5 junction and exon 5.

Figure S3. Multiple Sequence alignment of PLD3 amino acid sequences among homologous genes. The p.M6R (A), p.V232M (B) and p.A442A (C) variants are highlighted in yellow. The HKD signature motif (expanded to H-x-K-x(4)-D-x(6)-G-T-x-N in this subfamily, where x represents any amino acid residue) is also noted in the figures

A. p.M6R

```
ENSP00000387050/Human         MKPKLMYQELKVPAEEPANELPMNEIEAWKAAEKKARWVLLVLILAVVGFGALMT
ENSGGOP00000010739/Gorilla    MKPKLMYQELKVPAEEPANELPMNEIEAWKAAEKKARWVLLVLILAVVGFGALMT
ENSPPYP00000011181/Pongo      MKPKLMYQELKVPAEEPANELPMNEIEAWKAAEKKARWVLLVLILAVVGFGALMT
ENSMMUP00000031602/macaca     MKPKLMYQELKVPAEEPANELPMNEIEAWKAAEKKARWVLLVLILAVVGFGALMT
ENSBTAP00000041666/BosT       MKPKLMYQELKVPAEEPASELPMNEIEAWKAAEKKARWVLLVLILAVVGFGALMT
ENSCAFP00000007996/canis      MKPKLMYQELKVPAEEPASELPMNEIEAWKAAEKKARWVLLVLILAVVGFGALMT
ENSMUSP00000112942/MusMus     MKPKLMYQELKVPVEEPAGELPLNEIEAWKAAEKKARWVLLVLILAVVGFGALMT
ENSRNOP00000054004/rattus     MKPKLMYQELKVPVEEPAGELPMNEIEAWKAAEKKARWVLLVLILAVVGFGALMT
ENSEEUP00000013322/C.Ele      MKTKSLLLSHSIVAIVAVIITTAIWLT TYFVAV----NPNINNNGGQVINNYSNN
FBpp0297826/D.melano          MPEYKKLEDQESDVENANRTTVQN-TATVQDAGEGQRQAAGQQAGQ-MVT VSLFM
ENSDARP00000083110/DanioR     MKSDIPYEKMVDV-ELSR--------GEGHHGSQKYYRCLIVLTCITTVLLILLSL
```

HKD signature motif

B. p.V232M

```
ENSP00000387050/Human         LQSGAQVRMVDMQKLTHGVLHTKFWVVDQTHFYLGSANMDWRSLTQVKELGVVMYNCSCL
ENSGGOP00000010739/Gorilla    LQSGAQVRMVDMQKLTHGVLHTKFWVVDQTHFYLGSANMDWRSLTQVKELGVVMYNCSCL
ENSPPYP00000011181/Pongo      LQSGAQVRMVDMQKLTHGVLHTKFWVVDQTHFYLGSANMDWRSLTQVKELGVVMYNCSCL
ENSMMUP00000031602/macaca     LQSGAQVRMVDMQKLTHGVLHTKFWVVDQTHFYLGSANMDWRSLTQVKELGVVMYNCSCL
ENSBTAP00000041666/BosT       LQSGAQVRMVDMQKLTHGVLHTKFWVVDQTHFYLGSANMDWRSLTQVKELGVVMYNCSCL
ENSCAFP00000007996/canis      LQSGAQVRMVDMQKLTHGVLHTKFWVVDQTHFYIGSANMDWRSLTQVKELGVVMYNCSCL
ENSMUSP00000112942/MusMus     LQSGAQVRMVDMQKLTHGVLHTKFWVVDQTHFYLGSANMDWRSLTQVKELGVVMYNCSCL
ENSRNOP00000054004/rattus     LQSGAQVRMVDMQKLTHGVLHTKFWVVDQTHFYLGSANMDWRSLTQVKELGVVMYNCSCL
ENSEEUP00000013322/C.Ele      LQSXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXVKELGVVMYNCSCL
FBpp0297826/Drosophila        GAAEVVSISFPKYFGSGVLHTKLWVVDNKHFYLGSANMDWRALTQ-VKEMGVLVQNCPEL
ENSDARP00000083110/DanioR     ATGAEVRGVDLQSITGGILHTKLWVVDKKHVYLGSANMDWRSLTQ-VKEVGVALEDCGCL
```

HKD signature motif

C. **p.A442A**

```
ENSP00000387050/Human         VTERATYIGTSNWSGNYFTETAGTSLLVTQNGRGGLRSQLEAIFLRDWDSPYSHDLDTSA
ENSGGOP00000010739/Gorilla    VTERATYIGTSNWSGNYFTETAGTSLLVTQNGRGGLRSQLEAIFLRDWDSPYSHDLDTSA
ENSPPYP00000011181/Pongo      VTERATYIGTSNWSGNYFTETAGTSLLVTQNGRGGLRSQLEAIFLRDWDSPYSHDLDTSA
ENSMMUP00000031602/macaca     VTERATYIGTSNWSGNYFTETAGTSLLVMQNGRGSLRSQLEAIFLRDWDSPYSHDLDASA
ENSBTAP00000041666/BosT       VTERATYIGTSNWSGSYFTETAGTSLLVTQNGRGGLRSQLEAVFLRDWDSPYSHDLDAAA
ENSCAFP00000007996/canis      VTERATYIGTSNWSGSYFTETAGTSLLVTQNGRGGLRSQLEAVFLRDWDSPYSHDLDTSA
ENSMUSP00000112942/MusMus     VTERASYIGTSNWSGSYFTETAGTSLLVTQNGHGGLRSQLEAVFLRDWESPYSHDLDTSA
ENSRNOP00000054004/rattus     VTERTTYIGTSNWSGSYFTETAGTSLLVTQNGHGGLRSQLEAVFLRDWESPYSHNLDTSA
ENSEEUP00000013322/C.Ele      VTESAAYIGTSNWSSDYWQYTAGIGIVIRADDFTSKSKLVQQFTSVFERDWSSTYTIPLL
FBpp0297826/Drosophila        VTDRVAYIGTSNWSGDYFTDTAGIGLVLSETFETETTNTLRSDLRNVFERDWNSKYATPL
ENSDARP00000083110/DanioR     VTDQVAYIGTSNWSGDYFVNTAGSALVVNQTSASASSTVQEQLQAVFERDWESAYSTDIN
```

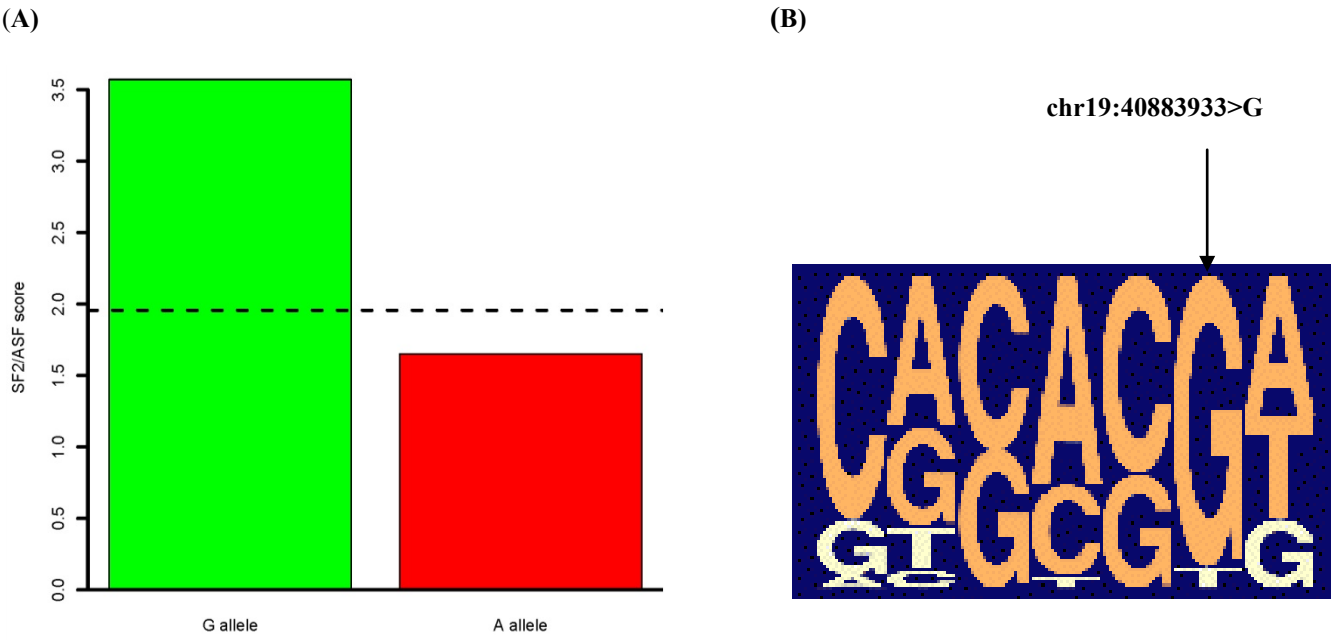


Figure S4: The *PLD3*- p.A442A variant modifies a splicing enhance binding site. A) ESEfinder predicts that the chr19:40883933 G>A change (p.A442A) disrupts a splicing enhancer binding site by making the sequence less similar to the consensus for the human SR protein SF2/ASF. The WT (chr19:40883933>G) allele has a prediction score of 3.57 for the SRSF1 protein. The A- allele has a prediction score of 1.X, below the threshold for SRSF1 binding**. B)** Conservation of the chr19:40883933>G allele. The chr19:40883933>G is the most conserved nucleotide of the neighboring nucleotides.

Chapter 3

Use of cerebrospinal fluid as endophenotypes to fine-map Alzheimer's disease associated genes

The research work in this chapter resulted in the following publications.

1. Cruchaga C, Kauwe JS, Harari O, **Jin SC**, et al. *GWAS of Cerebrospinal Fluid Tau Levels Identifies Risk Variants for Alzheimer's Disease.* Neuron. 2013 Apr 24; 78(2): 256-58.

2. Benitez BA*, **Jin SC***, et al. *Missense Variant in TREML2 Protects Against Alzheimer's Disease.* Neurobiology of Aging. 2013 Dec 21. doi: 10.1016/j.neurobiolaging.2013.12.010. (*Equal contribution)

3.1 ABSTRACT

The International Genomics of Alzheimer's Project (IGAP) performed meta- and gene-wide analyses and identified 23 loci associated with Alzheimer's disease (AD) risk, of which 13 were novel. However, the mechanisms by which most of these loci affect the molecular pathways leading to AD remain unknown. To determine whether these loci are also associated with cerebrospinal fluid (CSF) biomarker levels, we combined CSF biomarker datasets from several studies and performed single-variant, set-based, and conditional analyses for each locus. In the *APOE* locus, rs769449 is genome-wide significantly associated with CSF amyloid-beta 1-42 ($A\beta_{42}$) and phosphorylated tau_{181} ($ptau_{181}$) levels. Furthermore, as reported before in a smaller dataset the association between rs769449 and CSF $ptau_{181}$ levels is only partially explained by differences in CSF $A\beta_{42}$ levels. We also revealed that the association of rs769449 with CSF $A\beta_{42}$ and $ptau_{181}$ levels is independent of *APOE*-ε2 and *APOE*-ε4 genotypes, which suggests that another *APOE* risk variant is present in addition to known *APOE*-ε2 and *APOE*-ε4 alleles. We found evidence of association ($p<0.05$) for the top IGAP single nucleotide polymorphisms (SNPs), rs4147929 (*ABCA7*), rs17125944 (*FERMT2*), and rs35349669 (*INPP5D*) with CSF $A\beta_{42}$ levels, and for rs6656401 (*CR1*), rs17125944 (*FERMT2*), rs190982 (*MEF2C*), rs10792832 (*PICALM*), rs28834970 (*PTK2B*), and rs11218343 (*SORL1*) with CSF $ptau_{181}$ levels. Our locus-specific analyses suggested that after multiple test correction, rs7937331, within the *CELF1* fine-mapping region, is significantly associated with CSF $A\beta_{42}$ levels and AD risk, and tags the same signal as the IGAP top SNP, rs10838725. Additionally, rs62003531, located in the intronic region of *FERMT2*, tags the same association as the IGAP top SNP (rs17125944) and is associated with CSF $A\beta_{42}$ levels. The association for the *CELF1* and *FERMT2* fine-mapping regions with CSF $A\beta_{42}$ levels was confirmed in set-based analyses. None of the SNPs within the IGAP-identified AD risk loci except the *APOE* locus are significantly associated with CSF $ptau_{181}$ levels after multiple test correction. When investigating the potential regulatory functions associated with IGAP top SNPs and CSF top SNPs, most of GWAS top SNPs have no significant regulatory potential and are unlikely to be the functional variants for AD risk. However, RegulomeDB predicts that several proxy SNPs in linkage disequilibrium (LD) with rs7937331 in *SLC39A13* may be cis-acting expression quantitative trait loci (eQTLs) for nearby genes and are located in transcription factor binding sites. In summary, our results suggest that AD risk variants may not necessarily be associated with CSF biomarker levels, and that GWAS-identified non-coding variants may affect AD risk through regulatory mechanisms. The IGAP study also identified an intergenic polymorphism near *TREML2* suggestively ($p<10^{-6}$) associated with AD risk; however, due to the study

design, it was not possible to uncover the underlying functional variant or to determine whether this observed association was driven by the known AD risk allele, *TREM2* p.R47H, or represented a novel locus. Here, we conducted comprehensive analyses using whole-exome sequencing (WES) data, CSF biomarker analyses and meta-analyses (16,254 cases and 20,052 controls) from several independent cohorts to demonstrate that the AD risk association is likely driven by a *TREML2* missense variant p.S144G (rs3747742) and that this association is independent of *TREM2* p.R47H risk for AD.

3.2 INTRODUCTION

Recently, IGAP performed meta-analysis of genome-wide association studies (GWAS) and gene-wide analysis in more than 74,000 individuals and identified 23 loci (*ABCA7*, *APOE*, *BIN1*, *CASS4*, *CD2AP*, *CD33*, *CELF1*, *CLU*, *CR1*, *EPHA1*, *FERMT2*, *HLA-DRB5/DRB1*, *IGHV1-67*, *INPP5D*, *MEF2C*, *MS4A6A*, *NME8, PICALM*, *PTK2B*, *SLC24A4*, *SORL1*, *TP53INP1*, and *ZCWPW1*) significantly associated with risk for LOAD[8], 13 of which were novel. However, these loci contain multiple genes within the associated region and therefore, the suggested genes may not contain the underlying functional variants. Additionally, the most-significant polymorphisms within these loci were located in non-coding regions and thus there is no clear functional impact of these variants linked to AD pathogenesis[8]. It is also possible that multiple functional variants reside within the same locus affecting AD risk independently. Endophenotype-based analyses along with conditional analyses can help to identify the potential drivers of the associations, infer the underlying mechanisms associated with AD, and determine whether there are multiple independent genetic variants affecting AD pathogenesis.

AD is characterized pathologically by intracellular neurofibrillary tangles and extracellular amyloid plaques. Previous studies have shown that CSF $A\beta_{42}$ is a useful biomarker for plaques[104] and CSF $ptau_{181}$ levels are well correlated with the number of neurofibrillary tangles and tangle load in the human brain[146]. Moreover, the fact that CSF $ptau_{181}$ levels increase and CSF $A\beta_{42}$ levels decrease in AD patients indicate their roles as effective biomarkers for AD[22,104]. Changes in CSF biomarkers in cognitively normal people are also strong indicators of future cognitive decline[147]. Our previous work has used CSF $A\beta_{42}$ and $ptau_{181}$ levels as endophenotypes to validate AD genetic risk factors[23], to generate hypotheses regarding the mechanism by which AD risk variants contribute to AD development[18], and to find novel variants associated with age at onset and disease progression[17,19]. Similarly, in this study we used CSF $A\beta_{42}$ and $ptau_{181}$ levels as endophenotypes to determine whether the top polymorphisms

identified in the IGAP GWAS meta-analysis are also associated with CSF biomarker levels and to test whether there are multiple independent signals within the same locus affecting AD. This information can also provide clues to the mechanism of the disease-associated variants. Furthermore, since the most significant genetic markers within these loci are located outside of protein coding regions, we utilized RegulomeDB (http://regulomedb.org), a database including experimental datasets and computational algorithms, to predict regulatory potential for these variants.

The IGAP study also identified an intergenic polymorphism (rs9381040; $p<6.3\times10^{-7}$) located 5.5 Kb downstream from *TREML2* and 2.2 Kb upstream from *TREM2*[8] suggestively associated with AD risk[8]. The *TREM* and *TREM-like* receptor genes clustered on chromosome 6p.21 have different patterns of LD among them[22,148]. This genomic region has previously been implicated in genetic risk for AD[12,13,98,130-135]. A low frequency missense variant in *TREM2* (p.R47H, minor allele frequency [MAF]=0.003), identified through sequencing studies was reported to substantially increase risk for AD[12,13,125]. Single nucleotide polymorphisms (SNPs) in this region were also found to be associated with CSF $ptau_{181}$ levels[22]. Because of the design of the IGAP study (a meta-analysis) and the low frequency of the p.R47H variant in *TREM2*, it was not possible to determine whether the GWAS signal (rs9381040) was independent of the *TREM2*-p.R47H variant (rs75932628). Here, we analyzed WES data to identify the most likely functional variant in *TREML2* responsible for the GWAS signal and tested whether this signal is independent of the *TREM2*-p.R47H variant.

3.3 MATERIALS AND METHODS

3.3.1 CSF samples and data cleaning

CSF samples were obtained from the Charles & Joanne Knight Alzheimer's Disease Research Center (Knight-ADRC) (N=748), the Alzheimer's Disease Neuroimaging Initiative (ADNI) Study (ADNI-1) (N=382), ADNI-2 Study (N=377), the Mayo Clinic (N=441), Lund University (Swedish) (N=294), University of Pennsylvania (Penn) (N=181), University of Washington (UW) (N=323), and Saarland University (German) (N=102). Individuals were diagnosed with dementia of the Alzheimer's type (DAT) according to the National Institute of Neurological and Communication Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA)[118]. Control individuals were evaluated using the same criteria and showed no symptoms of cognitive impairment. All participants provided written informed consent and the ethics committee approved the study (IRB ID #: 201105364).

CSF was collected in a standardized manner[101]. Biomarker measurements within each study were conducted using internal standards and controls to achieve consistency and reliability. However, differences in the measured values between studies were observed which are likely due to differences in the antibodies and technologies used for quantification (standard ELISA with Innotest for Knight-ADRC, UW, Swedish, German, and Mayo versus Luminex with AlzBia3 for ADNI-1, ADNI-2, and Penn), ascertainment, and/or in handling of the CSF after collection. The differences in the number of freeze thaw cycles prior to analysis may also introduce some variation. To normalize CSF biomarker distributions, we log-transformed the values and then subtracted the mean of the transformed values for each dataset. The standardized CSF measurements were normally distributed after transformation and thus the standardized CSF biomarker measurements were used for the following analyses (see **Figure S1).**

We identified unanticipated duplicates and cryptic relatedness using pair-wise genome-wide estimates of proportion identity by descent (IBD) using the PLINK program[122]. When duplicate samples or a pair of samples with cryptic relatedness was identified, priority was given to samples with higher call rates. In order to control for population substructure, a principal component analysis (PCA) was conducted using the EIGENSTRAT software[121]. HapMap samples (CEU: CEPH Europeans from Utah; JPT: Japanese in Tokyo; YRI: Yoruba in Ibadan, Nigeria) were included in the analyses in order to remove outliers and confirm self-reported ethnicity. Samples were excluded if not within the CEU cluster (see **Figure S2 A-B**). We also checked whether the gender was discordant by analyzing the X-chromosome SNPs using PLINK. Individuals were removed if the recorded gender did not match the gender reported by our analyses. Individuals with an age under 45 were also removed. After these initial quality control (QC) steps, we applied several inclusion criteria for selection of cohorts into the final analyses in order to attain stable analytical results. Cohorts were excluded if they met any of the following conditions: (1) sample size less than 200 after QC, (2) an r-squared (R^2) value for *APOE*-ε4 association less than 10%, (3) sample truncation based on specific biomarker cutoffs. A total of 2,036 individuals from Knight-ADRC, ADNI-1, ADNI-2, Mayo, and UW passed QC filters and were finally analyzed. **Table 1** summarizes the sample characteristics used in the final analyses.

Samples were genotyped on either the Illumina 660 Chip (N = 614) or the Omniexpress Chip (N = 1,422). Samples genotyped on the Illumina 660 Chip included 183 from Knight-ADRC, 74 from UW and 351 from ADNI-

1. Samples genotyped on the Illumina Omniexpress included 431 from the Knight-ADRC, 216 from UW, 349 from ADNI-2 and 426 from the Mayo Clinic. Standard procedures were used to determine *APOE* genotypes (rs7412 and rs429358) as previously described[149]. Stringent quality thresholds were applied to the genotype data. SNPs were dropped if they fulfilled any one of the following criteria: i) genotyping success rate<98% per SNP or per individual; ii) Hardy-Weinberg equilibrium (HWE) ($p<1\times10^{-6}$); iii) MAF < 0.01. QCs were carried out separately based on genotyping chips (Illumina 660 versus Omniexpress).

After removing low quality SNPs and individuals, genotype imputation was performed using the Impute2 program[150] with haplotypes derived from 1,000 Genomes Project (released June 2012). Genotype imputation was performed separately based on the genotype platform used (Illumina Omniexpress versus 660K chip). SNPs with an info-score quality of less than 0.3 reported by Impute2, with a MAF < 0.05 or out of HWE were removed. A total of 4,185,256 imputed and directly-genotyped SNPs and 2,036 individuals were used for final analyses. After quality control filtering, a total of 2,036 individuals (614 from Knight-ADRC, 357 from ADNI-1, 349 from ADNI-2, 426 from Mayo, and 290 from UW) were used for final analyses (see **Table 1**).

In order to determine whether IGAP intergenic signal rs9381040 is independent of the *TREM2* p.R47H variant, rs9381040 and *TREML2* p.S144G (rs3747742), a coding variant in *TREML2* in tight LD with rs9381040, were extracted from the GWAS data[22], and confirmed by direct genotyping. The *TREM2* p.R47H was genotyped using KASPaR genotyping assay (LGC Genomics), as previously described[15,16].

3.3.2 Exome-sequencing data from Knight-ADRC

Enrichment of coding exons and flanking intronic regions was performed using a solution hybrid selection method with the SureSelect human all exon 50Mb kit (Agilent Technologies, Santa Clara, CA, USA) following the manufacturer's standard protocol on 46 unrelated AD cases and 39 unrelated controls from the Knight-ADRC. This was performed by the Genome Technology Access Center at Washington University in St. Louis (https://gtac.wustl.edu/). The captured DNA was sequenced by paired-end reads on the HiSeq 2000 sequencer (Illumina, San Diego, CA, USA). Raw sequence reads were aligned to the reference genome National Center for Biotechnology Information (NCBI) 36/hg18 by using Novoalign (Novocraft Technologies, Selangor, Malaysia). Base and/or SNP calling was performed using SNP SAMtools[151]. SNP annotation was carried out using version 5.07

Table 1. Summary characteristics of the CSF study participants

	Knight-ADRC	ADNI-1	ADNI-2	Mayo	UW
N	614	357	349	426	290
Age Mean ± SD (range)	70.49 ± 8.85 (46-91)	77.88 ± 6.97 (58-93)	72.84 ± 7.40 (55-91)	78.72 ± 6.38 (50-95)	66.91 ± 10.45 (45-88)
APOE ε4+ (%)	39.51	51.54	38.4	25.12	47.59
CDR 0 (%)	71.99	26.89	31.81	78.17	58.97
Male (%)	45.11	61.06	55.87	60.33	47.93
$Ptau_{181}$ Mean ± SD	66.79 ± 36.11	33.78 ± 17.64	23.67 ± 10.13	23.07 ± 10.48	60.57 ± 30.50
$Aß_{42}$ Mean ± SD	613.69 ± 272.27	168.84 ± 55.55	215.93 ± 74.03	331.29 ± 122.11	135.62 ± 40.80

Sample size, age in years at lumbar puncture (LP), the percentage of the subjects that carry at least one *APOE ε4* allele, the percentage of subjects with Clinical Dementia Rating (CDR) = 0, the percentage of male participants, and the mean in pg/ml, standard deviation, and range for CSF $ptau_{181}$ and $Aß_{42}$ for samples from Washington University Charles F. & Joanne Knight Alzheimer's Disease Research Center (Knight-ADRC), Alzheimer's Disease Neuroimaging Initiative Study (ADNI-1), ADNI2 Study (ADNI-2), Mayo Clinic (Mayo) and the University of Washington (UW) are shown.

of SeattleSeq Annotation server (http://snp.gs.washington.edu/SeattleSeqAnnotation138/)[152]. On average, 95% of the exome had > 8 fold coverage.

3.3.3 UK- National Institute on Aging (UK-NIA) dataset

A description of the UK-NIA dataset can be found in Guerreiro et al[12]. This dataset includes whole-exome sequencing data from 143 AD cases and 186 elderly controls individuals without dementia. All individuals were of European descent.

3.3.4 Alzheimer's Disease Genetic Consortium methods

Data used in the preparation of this article were obtained from the Alzheimer's Disease Genetic Consortium (ADGC), which includes 2,247 individuals from Adult Changes in Thought study, 4,325 individuals from Alzheimer's Disease Center, 413 individuals from ADNI, 1,256 individuals from Multi-Site Collaborative Study for Genotype-Phenotype Associations in Alzheimer's Disease, 2,087 individuals from University of Pittsburgh, 279 individuals from Oregon Health &Science University, 1,614 individuals from National Institute on Aging Late-onset AD, 2,263 individuals from University of Miami, 588 individuals from Multi-Institutional Research on Alzheimer's Genetic Epidemiology, 1,049 individuals from the Ruth University Religious Orders Study/Memory and Aging Project, 1,210 individuals from Translational Genomics Research Institute series, 481 individuals from Washington University in St. Louis, and 1,880 individuals from Mayo Clinic. To control for population substructure, we performed PCA using the EIGENSTRAT software[121] and a total of 19,673 (10,067 AD cases and 9,606 controls) individuals of European ancestry were used for final analysis (see **Figure S2 C-D**). A description of the sample included in the study as well as the methods used can be found in Naj et al[11]. Genome-wide imputation was performed per cohort using MACH software with HAPMAP phase 2 (release 22) CEPH Utah pedigrees reference. Imputation quality was set at $r^2 \geq 0.50$. A multivariate logistic regression was performed to evaluate the association between genetic markers and risk for AD adjusting for age, gender, population substructure, and study-specific effects.

3.3.5 Genetic and Environmental Risk for Alzheimer's Disease Consortium

Data were obtained from the Genetic and Environmental Risk for Alzheimer's Disease (GERAD) Consortium. The imputed GERAD sample comprised of 3,177 AD cases and 974 healthy elderly (age<70 yrs)

control subjects with available age and gender data. Cases and screened control subjects were recruited by the Medical Research Council (MRC) Genetic Resource for AD (Cardiff University; Institute of Psychiatry, London; Cambridge University; Trinity College Dublin), the Alzheimer's Research UK Collaboration (University of Nottingham; University of Manchester; University of Southampton; University of Bristol; Queen's University Belfast; the Oxford Project to Investigate Memory and Aging, Oxford University); Washington University in St. Louis; Medical Research Council PRION Unit, University College London; London and the South East Region AD Project, University College London; Competence Network of Dementia, and Department of Psychiatry, University of Bonn, Germany; the National Institute of Mental Health AD Genetics Initiative. A number of 6,129 control subjects were drawn from large existing cohorts with available GWAS data, including the 1,958 British Birth Cohort (http://www.b58cgene.sgul.ac.uk), the KORA F4 Study, and the Heinz Nixdorf Recall Study. All AD cases met criteria for either probable (National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's disease and Related Disorders Association [NINCDSADRDA], Diagnostic and Statistical Manual of Mental Disorders [DSM-IV]) or definite (Consortium to Establish a Registry for Alzheimer's Disease [CERAD]) AD. All elderly controls were screened for dementia using the MMSE or ADAS-cog. Additionally, individuals with a Braak score of 2.5 or lower were also included as controls. Genotypes from all cases and controls were previously included in the AD GWAS by Harold et al[78]. Imputation of the dataset was performed using IMPUTE2 and the 1,000 Genomes (http://www.1000genomes.org/) Dec 2010 reference panel (NCBI build 37.1). The imputed data was then analyzed using logistic regression including covariates for country of origin, gender, age, and first three principal components were obtained with EIGENSTRAT (EIGENSOFT 4.2)[153] software based on individual genotypes (best guess) for the GERAD study participants.

3.3.6 European Alzheimer's Disease Initiative Consortium

All AD cases were ascertained by neurologists from Bordeaux, Dijon, Lille, Montpellier, Paris, Rouen, and were identified as French Caucasian[154,155]. Clinical diagnosis of probable AD was established according to the DSM-III-R and NINCDS-ADRDA criteria. Control subjects were selected from the 3C Study[155]. This cohort is a population-based, prospective (7-year follow-up) study of the relationship between vascular factors and dementia. It has been carried out in 3 French cities: Bordeaux (southwest France), Montpellier (southeast France), and

Dijon (central eastern France). A sample of non-institutionalized subjects over 65yrs was randomly selected from the electoral rolls of each city. Between January 1999 and March 2001, 9,686 subjects meeting the inclusion criteria agreed to participate. After recruitment, 392 subjects withdrew from the study. Thus, 9,294 subjects were finally included in the study (2,104 in Bordeaux, 4,931 in Dijon, and 2,259 in Montpellier). Genomic DNA samples of 7,200 individuals were transferred to the French Centre National de Génotypage. At the end, we removed 308 samples because they were found to be first- or second- degree relatives of other study participants, or were assessed non-European descent based on genetic analysis using methods described in Heath et al[156]. In this final sample, at 7 years of follow-up, 459 individuals suffered from AD with 97 prevalent and 362 incident cases. These AD cases were included as cases in the European Alzheimer's disease initiative (EADI) discovery dataset. We retained the other individuals as control subjects (N=6,017). After individual sample collection (2,243 AD cases and 6,017 controls), the imputation was performed using 1,000 Genomes multi-ethnic data (1,000 Genomes phase 1 integrated variant set release v3) as a reference panel. Imputation was performed in 2 steps: pre-phasing with SHAPEIT (v2), followed by imputation with IMPUTE2.

3.3.7 Gene and SNP selection

For the locus-wide analysis, we selected 23 genes for the final analyses, including *ABCA7*, *APOE, BIN1*, *CD2AP*, *CD33*, *CASS4, CELF2, CLU, CR1, EPHA1, FERMT2, HLA-DRB5/DRB1, IGHV1-67, INPP5D, MEF2C, MS4A6A, NME8, PICALM, PTK2B, SLC24A4, SORL1, TP53INP1,* and *ZCWPW1,* which were recently identified by the IGAP study[8]. We extracted, imputed, and directly-genotyped genetic markers within the fine mapping regions of each locus (see **Table S1**) and tested for association with CSF $A\beta_{42}$ and $ptau_{181}$ levels. The fine mapping region for each locus was defined by the closest recombination hot spots in either direction from the coding region based on the estimated recombination rate from the HapMap samples (released March 2008) (see **Table S1**). Since the coordinates of the HapMap data were based on NCBI build 36 (UCSC hg18), we lifted them over to NCBI build 37 (UCSC hg19) using the UCSC Genome Browser (http://genome.ucsc.edu/). We then calculated the multiple test threshold for the fine-mapping regions of each locus by implementing the simpleM[157] approach in R (version 3.0.1).

3.3.9 Statistical analysis

For the analyses of CSF biomarker data, we first performed multivariate linear regression for association between standardized CSF $ptau_{181}$ and $A\beta_{42}$ levels and genetic variants in IGAP-identified loci adjusting for age,

gender, the first three principal components (PCs) and sites (coded as dummy variables) using PLINK[122]. Analyses were performed conditioning on the most significant SNP within the locus to determine whether other SNPs within the region represent independent associations after adjusting for covariates. Set-based analyses were then used for each IGAP-identified locus to evaluate the effects of all of the SNPs within each locus on CSF $A\beta_{42}$ or $ptau_{181}$ levels in PLINK[122] with default parameters by performing 10,000 permutation tests.

For the analyses of case-control data in *TREM2*/*TREML* regions, we performed multivariate logistic regression to evaluate the association between genetic markers and risk for LOAD adjusting for age, gender, population substructure, and study effects using PLINK[122]. Conditional analysis was performed to identify additional independent signals by conditioning on the top case-control GWAS hits. We first estimated the odds ratios (OR) for SNPs across cohorts. These models calculate crude OR and confidence intervals from counts of heterozygotes in case and control subjects in each study. Then we performed a fixed-effect model to combine the odds ratios from study-specific estimates into a summary measure. No multiple-testing correction was used in our analyses. The heterogeneity of effects was evaluated using the Woolf test for heterogeneity[158]. Meta-analysis was conducted using the META package (http://www.stats.ox.ac.uk/wjsliu/meta.html) in R (version 3.0.1).

3.3.10 Power calculation

We estimated the power by conducting an overall F test in a one-way, three-group analysis of variance to calculate using proc power in SAS. We calculated the effect size, which was measured in fold-difference between the means by assigning statistical power equal to 80% for minor allele frequencies from 0.1 to 0.5 and Type I error equal to 0.05 and 0.00001(most stringent single SNP multiple testing correction) assuming HWE (see **Table S2**).

3.3.11 Genome partitioning

The program GCTA (Genome-wide Complex Trait Analysis) was used to estimate the proportion of phenotypic variance explained by all imputed/genotyped SNPs, SNPs in the *APOE* gene, and SNPs that pass multiple test correction[159].

3.3.12 Bioinformatic analysis

The SeattleSeq Annotation server was used to annotate the variants. The RegulomeDB database (http://www.regulomedb.org) was used to investigate potential effects of associated variants on regulatory functions

based on data from the ENCODE Project[160]. Different types of information are integrated in RegulomeDB, which includes: (a) ENCODE transcription factor ChIP-seq, histone ChIP-seq, Formaldehyde-Assisted Isolation of Regulatory Elements, and DNase 1 hypersensitive site data (The ENCODE Project Consortium 2012),(b) transcription factor ChIP-seq data from NCBI Sequence Read Archive, and (c) a large collection of eQTL, DNase I sensitivity quantitative trait loci (dsQTL), and ChIP-exo (an advanced ChIP method to specifically identify binding sites of DNA-bound proteins at almost single-nucleotide resolution) data totaling 962 data sources covering over 100 tissues and cell lines. With around 60 million annotations, RegulomeDB provides users with a straightforward way to classify variants of interest and thus allow generation of hypotheses regarding the underlying mechanism of selected variants, esp. non-coding variants. A score ranging from 1 to 6 is assigned to a variant with a lower score representing a higher likelihood of affecting binding and expression of a gene target[160].

3.4 RESULTS

3.4.1 Sample characteristics

For the CSF datasets, sample collection, genotype imputation, and quality controls (QC) for the genotype and phenotype data were conducted as described in **MATERIALS AND METHODS.** A total of 4,185,256 imputed and directly-genotyped SNPs and 2,036 individuals passed QC filters (see **Table 1**). In order to define the fine-mapping regions, we considered the recombination rate from HapMap samples and finally selected a total of 12,109 SNPs within the GWAS-identified loci for further analysis[8]. The multiple test correction threshold for each locus was calculated using the simpleM algorithm[157] (see **Table S1**). The final analyses included 2,036 individuals of European American (EA) descent from Knight-ADRC, ADNI-1, ADNI-2, Mayo, and UW. The ethnicity of each individual was confirmed by PCA before inclusion in the final analyses (see **Figure S2 A-B)**. CSF collection and biomarker measurements within each study were conducted using standardized protocols, internal standards and controls to achieve consistency and reliability. However, differences in the measured values between studies were observed which are likely due to differences in the antibodies and technologies used for quantification, ascertainment, and in handling of the CSF samples after collection (see **Table 1**), e.g. differences in the number of freeze thaw cycles prior to analysis may introduce some variation. To normalize CSF biomarker distributions, we log-transformed the values and then subtracted the mean of the transformed values for each dataset. The

standardized CSF values were normally distributed after transformation; these CSF values were used for the final analyses (see **Figure S1**).

3.4.2 Effect of the *APOE* locus on CSF $A\beta_{42}$ and $ptau_{181}$ levels

Since *APOE* has by far the greatest effect on LOAD risk and has been reported to affect CSF $A\beta_{42}$ and $ptau_{181}$ levels in previous studies[22,161,162], we examined the association between *APOE* variants and CSF $A\beta_{42}$ and $ptau_{181}$ levels for reference purposes. The association for the *APOE* locus with both biomarker levels was significant when analyzing each study separately (see **Table S3**). When conducting a joint-analysis for the combined dataset, rs769449, was significantly associated with both CSF $A\beta_{42}$ and $ptau_{181}$ levels (MAF=17.88%; $p=7.60\times10^{-79}$ and 3.31×10^{-32} respectively; **Table S4**). To evaluate whether the association for rs769449 with CSF $ptau_{181}$ was driven by differences in $A\beta_{42}$ levels, CSF $A\beta_{42}$ levels were included in the regression modeling. The association for rs769449 with CSF $ptau_{181}$ levels became less significant but still exceeded the genome-wide significance cutoff ($p=9.91\times10^{-11}$), which suggests that rs769449 mediates differences in tau biology independently of Aβ pathology, supporting our previous findings[22]. To examine whether the association for rs769449 is independent of *APOE*-ε2 and *APOE*-ε4 SNPs, we performed analysis by including *APOE*-ε2 and *APOE*-ε4 as covariates. The association for rs769449 with CSF $A\beta_{42}$ and $ptau_{181}$ levels remained highly significant ($p=9.21\times10^{-11}$and 4.42×10^{-5} respectively), which indicates that another independent signal in the *APOE* locus contributes to the CSF biomarker association in addition to the known *APOE* ε2 and ε4 alleles. Genome partition analyses suggest the overall phenotypic variance for CSF $A\beta_{42}$ and $ptau_{181}$ levels explained by all genotyped and imputed variants is approximately 31.64% and 10.40% respectively. The variability for CSF $ptau_{181}$ levels is slightly higher (10.40% versus 6.79%[22]) than that in our previous study which may due to the use of different imputation software (Impute2 versus Beagle)[22]. About 2.70% (CSF $A\beta_{42}$) and 0.79% (CSF $ptau_{181}$) of the phenotypic variance is explained by variants in the *APOE* fine-mapping region, suggesting that genetic variants outside the *APOE* locus account for most of the phenotypic variance for both CSF $A\beta_{42}$ and $ptau_{181}$ levels.

3.4.3 *CELF1, EPHA1,* and *FERMT2* loci associated with CSF $A\beta_{42}$ levels

We examined the association between IGAP-identified top SNPs and CSF $A\beta_{42}$ levels[8]. Our CSF biomarker analyses show that the minor alleles of rs4147929 (*ABCA7*: $p=1.00\times10^{-2}$; **Table 2**), rs17125944 (*FERMT2*:

p=7.42×10^{-3}**; Table 2**) and rs35349669 (*INPP5D*: p=3.40×10^{-2}; **Table 2**) are associated with lower CSF Aß$_{42}$ levels, which is consistent with the reported direction of effect for each SNP in the IGAP paper (see **Table 2**)[8]. Regional plots for each locus can be found in **Figures S3-S25.**

We then tested for association between SNPs within each IGAP locus and CSF Aβ$_{42}$ levels to determine whether other SNPs within each locus showed stronger association with CSF Aβ$_{42}$ levels than those associated with AD risk. Several SNPs surpassed multiple test thresholds. Rs7937331, the strongest association within the *CELF1* fine-mapping region is approximately 57 kb away from the 5'end of *CELF1*, but located in an intron of another gene, solute carrier family 39 (zinc transporter), member 13 (*SLC39A13*) (see **Figure 1A**). The minor allele of rs7937331 is significantly associated with increased CSF Aβ$_{42}$ values (MAF=33.05%; p=2.30×10^{-4}; β=0.021 [0.01~0.032]; **Table 3**). Set-based analysis of 211 SNPs in the *CELF1* fine-mapping region supports the association with CSF Aβ$_{42}$ levels (p=6.30×10^{-3}; **Table 5)**. Rs7937331 is significantly associated with AD risk in the IGAP study and the direction of effects is in line with that in our CSF analysis (OR=0.93, p=5.11×10^{-7}, **Table 3**)[8]. Rs7937331 has a high D' but a low R^2 with the IGAP top SNP, rs10838725 (MAF=31.26%) in the *CELF1* region (D'=0.97; R^2=0.21; **Table S5**)[8]. When adjusting for rs10838725, the association of rs7937331 with CSF Aβ$_{42}$ levels only changes slightly (p=1.25×10^{-3}; β=0.02 [0.008~0.032]; **Table S6 and Figure S4**) but that of rs10838725 becomes insignificant (p=8.67×10^{-1}; β=-0.001 [-0.014~0.011]; **Table S6 and Figure S4**). We also analyzed the ADGC case-control series in order to confirm rs10838725 and rs7937331 are independent signals. Before adjusting for rs10838725, rs7937331 shows evidence of association with AD risk (OR=0.93, p=4.14×10^{-4}; **Table S7**). The association of rs7937331 does not change significantly (OR=0.93, p=3.98×10^{-3}; **Table S7**) after adjusting for rs10838725. However, there were significant differences for association of rs10838725 with AD risk before (OR=0.96, p=0.05; **Table S7**) and after adjusting for rs7937331 (OR=0.99, p=0.61; **Table S7**). Together, our analyses suggest that rs7937331 in *SLC39A13* is significantly associated with both AD risk and CSF Aβ$_{42}$ levels and that rs7937331 and rs10838725 in *CELF1* tag the same signal. Importantly, rs7937331 is more likely to be the causal variant than rs10838725.

Another significant association affecting CSF Aβ$_{42}$ levels was found for rs7802536 within the *EPHA1* region (MAF=22.36%; p=6.01×10^{-5}; β=-0.025 [-0.037~-0.013]; **Table 3**). Rs7802536, a directly-genotyped SNP, is located about 38 kb away from the 5' end of *EPHA1* and is in the intergenic region between the *Chloride Channel, Voltage-*

Sensitive 1 (*CLCN1*) and *Family with Sequence Similarity 131,* member B (*FAM131B*) genes (see **Figure 1B**), which suggests that the association with CSF $A\beta_{42}$ levels within the *EPHA1* region may not be driven by *EPHA1*. Moreover, rs7802536 and rs11771145 (MAF=34.36%; **Table 2**), the IGAP-top SNP in *EPHA1*, are not in LD (D'=0.01; R^2=0; **Table S5**)[8]. The association of rs7802536 with CSF $A\beta_{42}$ levels did not change after adjusting for rs11771145 ($p=5.97\times10^{-5}$; β=-0.025 [-0.037~-0.013]; **Table S6 and Figure S5**) and vice versa ($p=6.53\times10^{-1}$; β=0.002 [-0.008~0.013]; **Table S6 and Figure S5**). Set-based analysis of 127 SNPs in this locus suggests evidence of association with CSF $A\beta_{42}$ levels ($P = 1.63\times10^{-2}$; **Table 5**). However, we did not find rs7802536 in the IGAP dataset and therefore unable to determine whether rs7802536 is significantly associated with AD risk.

Rs62003531, located in an intron of *FERMT2* (see **Figure 1C**), also survived multiple test correction and was significantly associated with CSF $A\beta_{42}$ levels (MAF=11.54%; $p=8.50\times10^{-5}$; β=-0.033 [-0.049~-0.016]; **Table 3**). In the stage 1 analysis of the IGAP study, rs62003531 is suggestively associated with AD risk ($p=2.75\times10^{-4}$; **Table 3**). Rs62003531 is in significant LD (D'=0.95; R^2=0.70; **Table S5**) with rs17125944 (MAF=8.95%; **Table 2**), the IGAP top SNP in *FERMT2*. The association for rs62003531 with CSF $A\beta_{42}$ levels reduced and became non significant when conditioning on rs17125944 ($p=5.85\times10^{-3}$; β=-0.041 [-0.070~-0.012]; **Table S6 and Figure S6**) and vice versa, which suggests that rs62003531 and rs17125944 tag the same signal. Set-based analyses of 463 SNPs in *FERMT2* indicated that *FERMT2* is significantly associated with CSF $A\beta_{42}$ levels ($p=7.80\times10^{-3}$; **Table 5**).

Table 2. Summary statistics for most significant SNP in IGAP top loci

						CSF $A\beta_{42}$		CSF $ptau_{181}$	
SNP	Chr:bp	Nearest Gene	P[!]	OR[!]	MAF (%)[&]	Beta (95% CI)	P	Beta (95% CI)	P
rs4147929	19:1063443	*ABCA7*	1.06×10^{-15}	1.15	18.56	-0.017 (-0.03,-0.004)	1.00×10^{-2}	0.015 (0,0.03)	5.68×10^{-2}
rs6733839	2:127892810	*BIN1*	6.94×10^{-44}	1.22	33.07	-0.006 (-0.016,0.005)	2.97×10^{-1}	0.006 (-0.006,0.019)	3.40×10^{-1}
rs7274581	20:55018260	*CASS4*	2.46×10^{-8}	0.88	8.57	-0.007 (-0.025,0.011)	4.29×10^{-1}	-0.011 (-0.033,0.01)	3.03×10^{-1}
rs10948363	6:47487762	*CD2AP*	5.20×10^{-11}	1.1	28.31	0.007 (-0.005,0.018)	2.49×10^{-1}	0.004 (-0.01,0.017)	6.12×10^{-1}
rs3865444	19:51727962	*CD33*	2.97×10^{-6}	0.94	31.73	0.007 (-0.004,0.018)	2.37×10^{-1}	-0.007 (-0.02,0.006)	2.83×10^{-1}
rs10838725	11:47557871	*CELF1*	1.12×10^{-8}	1.08	31.26	-0.009 (-0.02,0.002)	1.09×10^{-1}	0.006 (-0.007,0.019)	3.55×10^{-1}
rs9331896	8:27467686	*CLU*	2.77×10^{-25}	0.86	39.51	0.01 (-0.001,0.021)	6.47×10^{-2}	-0.008 (-0.021,0.004)	1.86×10^{-1}
rs6656401	1:207692049	*CR1*	5.69×10^{-24}	1.18	19.50	-0.006 (-0.019,0.007)	3.47×10^{-1}	0.024 (0.009,0.039)	1.88×10^{-3}
rs11771145	7:143110762	*EPHA1*	1.12×10^{-13}	0.9	34.36	0.002 (-0.008,0.013)	6.70×10^{-1}	0.001 (-0.011,0.014)	8.49×10^{-1}
rs17125944	14:53400629	*FERMT2*	7.94×10^{-9}	1.14	8.95	-0.024 (-0.041,-0.006)	$7.42\times10^{-3*}$	0.026 (0.005,0.046)	1.57×10^{-2}
rs9271192	6:32578530	*HLA-DRB5/DRB1*	2.94×10^{-12}	1.11	27.08	-0.005 (-0.018,0.007)	3.79×10^{-1}	-0.001 (-0.016,0.013)	8.73×10^{-1}
rs35349669	2:234068476	*INPP5D*	3.17×10^{-8}	1.08	47.54	-0.011 (-0.022,-0.001)	$3.40\times10^{-2*}$	0.009 (-0.003,0.022)	1.44×10^{-1}
rs190982	5:88223420	*MEF2C*	3.23×10^{-8}	0.93	42.20	-0.002 (-0.013,0.008)	6.54×10^{-1}	-0.015 (-0.028,-0.003)	1.20×10^{-2}
rs983392	11:59923508	*MS4A6A*	6.14×10^{-16}	0.9	40.71	0.002 (-0.008,0.013)	6.63×10^{-1}	-0.011 (-0.023,0.001)	8.29×10^{-2}
rs10792832	11:85867875	*PICALM*	9.32×10^{-26}	0.87	35.00	0.01 (-0.001,0.021)	7.04×10^{-2}	-0.016 (-0.029,-0.003)	1.27×10^{-2}
rs28834970	8:27195121	*PTK2B*	7.37×10^{-14}	1.1	36.01	-0.01 (-0.021,0.001)	6.46×10^{-2}	0.016 (0.002,0.029)	1.97×10^{-2}
rs10498633	14:92926952	*SLC24A4*	5.54×10^{-9}	0.91	22.65	0.001 (-0.011,0.013)	8.78×10^{-1}	-0.003 (-0.017,0.012)	7.28×10^{-1}
rs11218343	11:121435587	*SORL1*	9.73×10^{-15}	0.7	11.56	0.013 (-0.004,0.029)	1.26×10^{-1}	-0.02 (-0.039,-0.001)	4.25×10^{-2}
rs2718058	7:37841534	*NME8*	4.8×10^{-9}	0.93	36.02	0.004 (-0.007,0.014)	5.20×10^{-1}	-0.009 (-0.022,0.003)	1.54×10^{-1}
rs12539172	7:100091795	*ZCWPW1*	6.02×10^{-10}	0.91	28.97	-0.004 (-0.016,0.007)	4.65×10^{-1}	-0.006 (-0.019,0.008)	4.13×10^{-1}

* These SNPs were not found in our combined dataset and the association with CSF $A\beta_{42}$ and $ptau_{181}$ levels was estimated by proxy SNPs: rs7561528 for rs6733839, rs11787077 for rs933189, rs3818361 for rs6656401, rs9271100 for rs9271192, rs28655385 for rs35349669, rs304132 for rs190982, rs10897009 for rs983392, and rs3781832 for rs11218343. Chr, chromosome; MAF, minor allele frequency. ! Statistics reported in overall-all analyses of the recent IGAP study[8]. & MAF in the combined CSF dataset.

Table 3. Most significant association with CSF Aβ42 level in each IGAP locus

Gene	SNP	Chr:bp	Function	MAF (%)	SNP Type	Case/control association&	Beta (95% CI)	P	GERP Score
ABCA7	rs76348507	19:1048116	intron	10.02	Imputed	1.80×10^{-8}	-0.026 (-0.042,-0.009)	2.74×10^{-3}	0.16
BIN1	rs1469979	2:127896232	intergenic	26.01	Genotyped	0.52	0.011 (-0.001,0.023)	7.89×10^{-2}	1.92
CD2AP	rs13190867	6:47314068	intergenic	8.59	Imputed	0.73	0.031 (0.013,0.05)	9.09×10^{-4}	-0.77
CASS4	rs17365060	20:55000949	intron	16.29	Imputed	0.96	0.017 (0.003,0.031)	1.76×10^{-2}	-3.79
CD33	rs10404590	19:51701749	intergenic	29.61	Imputed	3.41×10^{-3}	-0.012 (-0.023,-0.001)	3.81×10^{-2}	NA
CELF1	rs7937331	11:47430458	intron	33.05	Imputed	5.11×10^{-7}	0.021 (0.01,0.032)	2.30×10^{-4}	2.5
CLU	rs17383366	8:27504228	intron	7.83	Imputed	0.01	0.025 (0.006,0.044)	9.88×10^{-3}	2.81
CR1	rs79795098	1:207737870	intron	16.24	Imputed	0.1	0.017 (0.002,0.031)	2.61×10^{-2}	0.75
EPHA1	rs7802536	7:143049973	downstream	22.36	Genotyped	NA	-0.025 (-0.037,-0.013)	6.01×10^{-5}	-4.05
FERMT2	rs62003531	14:53357018	intron	11.54	Imputed	2.75×10^{-4}	-0.033 (-0.049,-0.016)	8.50×10^{-5}	0.59
HLA-DRB5/DRB1	rs115485493	6:32654807	intergenic	27.58	Imputed	0.66	0.013 (0.001,0.024)	2.90×10^{-2}	-2.61
IGHV1-67	rs75196489	14:107155455	intergenic	11.42	Imputed	7.49×10^{-7}	0.017 (0,0.034)	4.41×10^{-2}	NA
INPP5D	rs13385922	2:234081324	intron	39.86	Genotyped	0.02	0.014 (0.004,0.024)	7.98×10^{-3}	0.16
MEF2C	rs141729694	5:87999371	intergenic	7.75	Imputed	0.04	-0.023 (-0.042,-0.003)	2.07×10^{-2}	-0.47
MS4A6A	rs640219	11:59761679	intergenic	22.19	Imputed	0.13	-0.015 (-0.027,-0.003)	1.33×10^{-2}	1.37
NME8	rs9655029	18:37974718	intergenic	38.48	Imputed	0.64	-0.016 (-0.027,-0.006)	2.49×10^{-3}	1.62
PICALM	rs573167	11:85831246	intergenic	32.41	Imputed	8.30×10^{-23}	0.017 (0.006,0.028)	2.92×10^{-3}	2.43
PTK2B	rs71519637	8:27334098	intron	9.32	Imputed	7.58×10^{-8}	0.025 (0.007,0.044)	7.43×10^{-3}	2.11
SLC24A4	rs12887171	14:92824871	intron	31.63	Genotyped	0.31	-0.014 (-0.025,-0.003)	1.54×10^{-2}	1.64
SORL1	rs7103597	11:121418806	intron	39.19	Imputed	0.22	-0.013 (-0.024,-0.003)	1.46×10^{-2}	1.05
TP53INP1	rs12548367	8:95929202	intergenic	34.76	Imputed	0.16	0.018 (0.008,0.029)	8.64×10^{-4}	0.36
ZCWPW1	rs858503	7:99845866	intron	16.87	Imputed	0.11	0.021 (0.007,0.034)	3.11×10^{-3}	0.35

Association with CSF $A\beta_{42}$ levels was analyzed using multivariate linear regression adjusting for age, gender, first three principle components, and dummies for sites.* Association that passes the multiple-testing threshold calculated using simpleM algorithm34. Chr, chromosome; MAF, minor allele frequency; CI, confidence interval. & p-values in the IGAP study{Lambert, 2013 #643}.

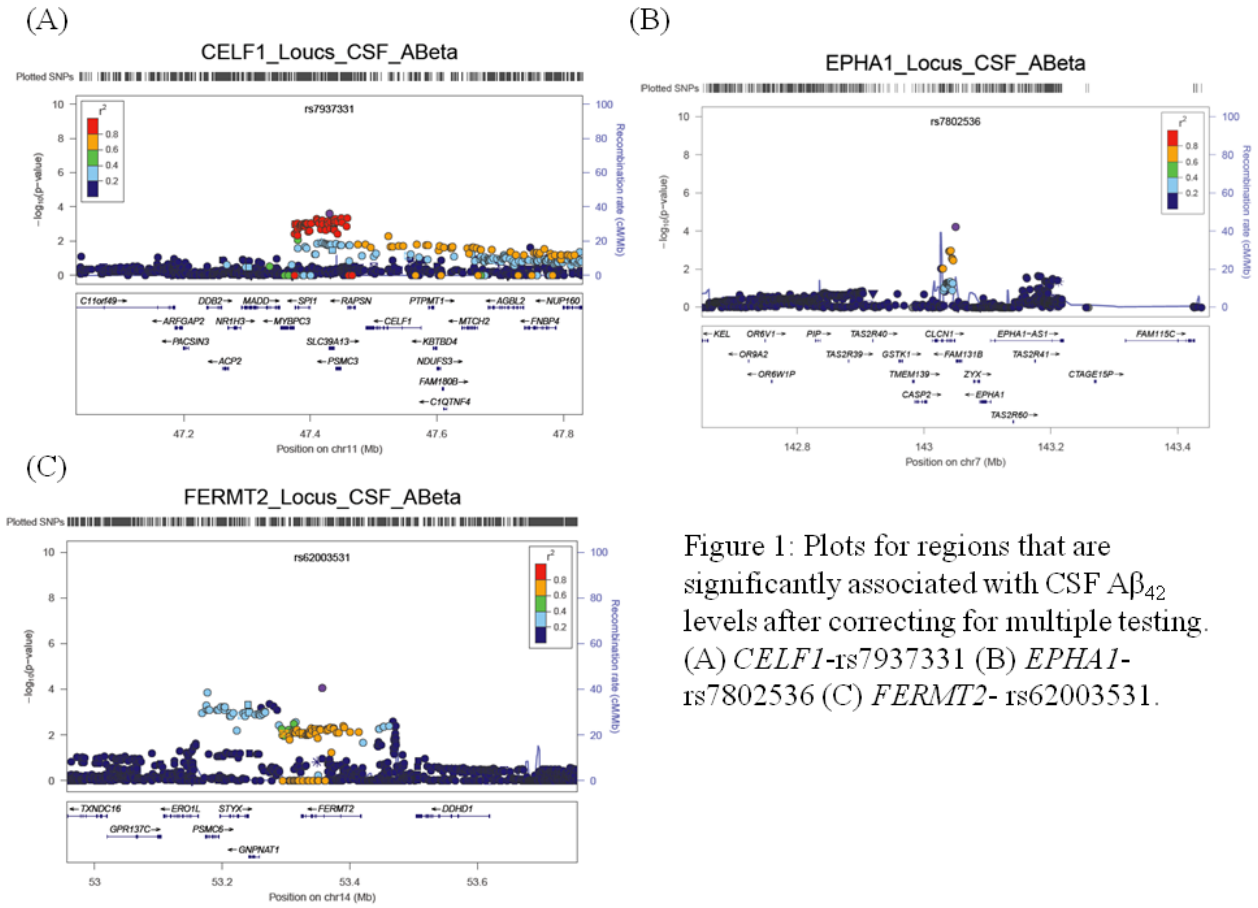


Figure 1: Plots for regions that are significantly associated with CSF $A\beta_{42}$ levels after correcting for multiple testing. (A) *CELF1*-rs7937331 (B) *EPHA1*-rs7802536 (C) *FERMT2*- rs62003531.

3.4.4 Several GWAS top hits show evidence of association with CSF $ptau_{181}$ levels

For the most significant associations identified in the IGAP paper [8], minor alleles of rs6656401 (*CR1*: $p=1.88\times10^{-3}$; **Table 2**), rs17125944 (*FERMT2*: $p=1.57\times10^{-2}$; **Table 2),** rs190982 (*MEF2C*: $p=1.20\times10^{-2}$; **Table 2**), rs10792832 (*PICALM*: $p=1.27\times10^{-2}$; **Table 2**), rs28834970 (*PTK2B*: $p=1.97\times10^{-2}$; **Table 2**), and rs11218343 (*SORL1*: $p=4.25\times10^{-2}$; **Table 2**) showed evidence of association ($p<0.05$) with CSF $ptau_{181}$ levels. Importantly, the directions of effects on CSF $ptau_{181}$ levels for these SNPs are in agreement with the results reported in the recent IGAP paper, i.e. AD risk alleles are associated with higher CSF $ptau_{181}$ levels (see **Table 2**). However, in the locus-wide analyses, we did not identify any SNPs that exceed multiple testing thresholds for association with CSF $ptau_{181}$ levels. Regional plots for each locus can be found in **Figures S26-S46.**

Table 4. Most significant association with CSF ptau$_{181}$ level in each IGAP locus

Gene	SNP	Chr:bp	Function	MAF (%)	SNP Type	Case/control association[&]	Beta (95% CI)	P	GERP Score
ABCA7	rs757232	19:1075979	intron	25.97	Genotyped	4.69×10^{-9}	0.015 (0.002,0.029)	2.40×10^{-2}	-3.43
BIN1	rs10166461	2:127859413	intron	19.36	Imputed	5.20×10^{-10}	-0.018 (-0.033,-0.003)	2.21×10^{-2}	-7.38
CD2AP	rs12201065	6:47626833	intron	41.06	Imputed	0.01	-0.016 (-0.029,-0.004)	9.02×10^{-3}	-2.1
CASS4	rs6127744	20:54986274	upstream	11.64	Imputed	8.79×10^{-7}	-0.019 (-0.039,0)	5.53×10^{-2}	0.97
CD33	rs11668174	19:51763498	intron	6.22	Imputed	1.58×10^{-3}	-0.024 (-0.049,0.001)	6.00×10^{-2}	-0.47
CELF1	rs7937331	11:47430458	intron	33.05	Imputed	5.11×10^{-7}	-0.015 (-0.028,-0.002)	2.38×10^{-2}	2.5
CLU	rs570164	8:27503653	intron	23.32	Imputed	0.28	0.013 (-0.002,0.027)	8.26×10^{-2}	-1.65
CR1	rs4844610	1:207802552	intron	18.46	Imputed	6.90×10^{-23}	0.026 (0.01,0.042)	1.14×10^{-3}	0.21
EPHA1	rs2272251	7:143042837	synonymous	45.72	Genotyped	0.52	-0.014 (-0.026,-0.002)	2.47×10^{-2}	-1.84
FERMT2	rs78459273	14:53274260	intergenic	8.37	Imputed	7.65×10^{-3}	0.035 (0.013,0.056)	1.54×10^{-3}	-4.07
HLA-DRB5/DRB1	rs114975350	6:32428062	intergenic	29.73	Imputed	1.98×10^{-4}	0.023 (0.009,0.036)	8.53×10^{-4}	1.45
IGHV1-67	rs11850600	14:107126781	intergenic	11.15	Imputed	1.85×10^{-3}	-0.028 (-0.048,-0.009)	4.53×10^{-3}	-0.60
INPP5D	rs10170794	2:233894598	upstream	17.71	Imputed	NA	0.023 (0.007,0.039)	5.69×10^{-3}	-0.48
MEF2C	rs304132	5:88215594	intergenic	42.2	Genotyped	1.21×10^{-7}	-0.015 (-0.028,-0.003)	1.20×10^{-2}	1.58
MS4A6A	rs663925	11:59815517	3' UTR	37.7	Genotyped	0.08	-0.021 (-0.033,-0.008)	9.79×10^{-4}	2.66
NME8	rs2598023	18:37915580	intergenic	37.76	Imputed	0.65	-0.019 (-0.031,-0.006)	3.98×10^{-3}	0.23
PICALM	rs10792828	11:85826797	intergenic	9.89	Imputed	0.20	0.033 (0.013,0.053)	1.51×10^{-3}	1.52
PTK2B	rs41276295	8:27289126	intron	45.03	Imputed	6.90×10^{-4}	0.017 (0.005,0.029)	6.14×10^{-3}	2.05
SLC24A4	rs12590749	14:92868468	intron	35.61	Genotyped	0.85	0.02 (0.007,0.032)	2.53×10^{-3}	0.63
SORL1	rs643010	11:121336471	intron	48.23	Imputed	0.15	-0.02 (-0.032,-0.007)	1.57×10^{-3}	-1.21
TP53INP1	rs61596977	8:95997165	intergenic	12.38	Imputed	5.17×10^{-3}	0.031 (0.012,0.049)	1.38×10^{-3}	-3.21
ZCWPW1	rs76281814	7:99846414	intron	22.04	Imputed	0.72	-0.016 (-0.031,-0.001)	3.57×10^{-2}	-0.22

Association with CSF ptau$_{181}$ levels was analyzed using multivariate linear regression adjusting for age, gender, first three principle components, and dummies for sites. Chr, chromosome; MAF, minor allele frequency; CI, confidence interval. [&] p-values in the IGAP study[8].

Table 5. Results of set-based analyses for each GWAS-identified locus

Gene	# of SNPs in the set	Empirical $P_{A\beta}$	Empirical P_{ptau}
ABCA7	73	1.23×10^{-1}	3.41×10^{-1}
BIN1	391	1	5.98×10^{-1}
CD2AP	804	1.41×10^{-1}	3.87×10^{-1}
CASS4	244	2.48×10^{-1}	1
CD33	118	4.49×10^{-1}	1
CELF1	211	$6.30\times10^{-3*}$	2.72×10^{-1}
CLU	149	2.55×10^{-1}	1
CR1	445	6.23×10^{-1}	2.25×10^{-1}
EPHA1	127	$1.63\times10^{-2*}$	3.52×10^{-1}
FERMT2	463	$7.80\times10^{-3*}$	9.62×10^{-2}
HLA-DRB5/DRB1	3350	6.01×10^{-1}	5.29×10^{-2}
IGHV1-67	126	1.96×10^{-1}	$2.26\times10^{-2*}$
INPP5D	517	6.06×10^{-1}	3.56×10^{-1}
MEF2C	195	4.24×10^{-1}	1.28×10^{-1}
MS4A6A	618	4.43×10^{-1}	1.95×10^{-1}
NME8	1071	$1.64\times10^{-2*}$	6.31×10^{-1}
PICALM	476	2.72×10^{-1}	5.56×10^{-2}
PTK2B	599	5.19×10^{-1}	1.90×10^{-1}
SLC24A4	588	7.96×10^{-1}	2.17×10^{-1}
SORL1	522	5.11×10^{-1}	2.95×10^{-1}
TP53INP1	677	6.80×10^{-2}	1.61×10^{-1}
ZCWPW1	244	1.25×10^{-1}	5.02×10^{-1}

Set-based analyses were implemented to evaluate association between all of the SNPs within the IGAP genes with CSF $A\beta_{42}$ and $ptau_{181}$ levels. Age, gender, population substructure, and dummies for sites were adjusted and 10,000 permutations were used to estimate the association using default parameters in PLINK. * denotes significant association.

3.4.5 CSF analyses for additional recently identified genetic markers

Recent studies have reported an association for *ABCA7* (rs3764650), *CR1* (rs6701713), and *CD2AP* (rs9349407) with neuritic plaque burden[163] and for *CLU* (rs11136000) and *MS4A6A* (rs2304933) with CSF $A\beta_{42}$ levels[164].Therefore, we analyzed the CSF dataset to determine whether these findings could be replicated. Rs3764650 in *ABCA7* is in borderline GWS association with AD risk ($p=3.22\times10^{-7}$, **Table S8**) in stage one analysis of the IGAP study[8]. In our combined analyses of the joint dataset, rs3764650 is associated with CSF $A\beta_{42}$ levels ($p=4.24\times10^{-3}$; β=-0.024 [-0.041~-0.008]; **Table S8**) but not with CSF $ptau_{181}$ levels (p=0.57; β=0.006 [-0.014~0.025]; **Table S8**). Rs3764650 is in modest LD with rs4147929, the IGAP top hit in *ABCA7* (D'=0.66;

R^2=0.22; **Table S5**) but in almost complete LD with rs76348507, the most significant SNP for the *ABCA7* locus in our CSF analyses (D'=1; R^2=0.99; **Table S5**). Conditional analyses suggest that rs3764650, rs76348507 and rs4147929 tag the same signal but rs3764650 and rs76348507 are in higher LD with the underlying functional variant. (**Table S6**). Our data supports the association of rs3764650 with Aβ metabolism and suggests that there is one signal within *ABCA7* associated with AD risk.

In stage 1 of the IGAP study, rs6701713 in *CR1* ($p=3.47\times10^{-14}$; **Table S8**) and rs11136000 in *CLU* ($p=1.72\times10^{-16}$; **Table S8**) are genome-wide significantly associated with AD risk but rs9349407 in *CD2AP* ($p=3.92\times10^{-7}$; **Table S8**) and rs2304933 in *MS4A6A* ($p=5.35\times10^{-5}$; **Table S8**) are only suggestively associated with AD risk. However, we did not observe evidence of association for rs6701713 in *CR1* ($p=3.47\times10^{-1}$; **Table S8**), rs9349407 in *CD2AP* ($p=2.66\times10^{-1}$; **Table S8**), rs11136000 in *CLU* ($p=5.75\times10^{-2}$; **Table S8**), and rs2304933 in *MS4A6A* ($p=1.64\times10^{-1}$; **Table S8**) with CSF $A\beta_{42}$ levels (**Table S8**). Surprisingly, rs6701713 in *CR1* is associated with CSF $ptau_{181}$ levels ($p=1.88\times10^{-3}$; β=0.024 [0.009~0.039]; **Table S8**). Rs6701713 is in significant LD with rs6656401 (D'=0.99; R^2=0.95; **Table S5**), the IGAP top SNP in *CR1*. Additionally, conditional analyses suggest that rs6701713, recently reported to be associated with neuritic plaque burden[163], tags the same signal as rs6656401 for association with CSF $ptau_{181}$ levels (**Table S6**). We did not find significant association for rs9349407 in *CD2AP*, rs11136000 in *CLU*, and rs2304933 with CSF $ptau_{181}$ levels (**Table S8**).

3.4.6 RegulomeDB prediction

To better understand potential function roles of the top SNPs within GWAS-identified loci, we utilized the RegulomeDB database[160] to predict their regulatory potential across several cell lines. With the exception of rs9271192 in *HLA-DRB5*/*DRB1*(score=1f) and rs1476679 in *ZCWPW1* (score=1f), IGAP top SNPs had a RegulomeDB score greater than 3, which indicates that the majority of GWAS top SNPs are unlikely to be the functional variants for AD risk. Since we did not observe significant regulatory function for variants which were significantly associated with CSF $A\beta_{42}$ levels (score=4 for rs7937331 in *SLC39A13*, score=4 for rs7802536 in *EPHA1* and score=7 for rs62003531 in *FERMT2*), we searched the SNAP server (based on the CEU populations in 1,000 Genomes Pilot 1) and our CSF dataset for proxy SNPs in LD (R^2>0.8) with these three top SNPs. In total, we found 77 proxy SNPs (38 for *SLC39A13*-rs7937331, 0 for *EPHA1*-rs7802536, and 39 for *FERMT2*-rs62003531 (see **Table S9**). Of these 77 proxy SNPs, 24 (24 for *CELF1*-rs7937331) had a RegulomeDB score <3 (see **Table 7**),

where scores 1 and 2 suggest a potential impact on binding change and targeted-gene expression differentiation (see **Table S10**). Proxy SNPs for rs7802536 in *SLC39A13* were predicted to be eQTLs for *AMBRA1*, *C1QTNF4*, *MYBPC3*, *NR1H3*, *SNRPG*, or *SPI1* and are located in the binding motifs for RREB1, Foxl1, E47, EWSR1-FLI1, SREBP1, SREBP2, CACCC-binding facot, ZIC1, PLAG1, RFX1, Zfx, or HOXA5. Additionally, these proxy SNPs may affect binding of FOXA1, POLR2A, TCF4, BLIMP1, HOXB3, SP1, NFYA, NFYB, FOS, STAT, CTCF, POU2F2, CEBPB, MAFK, CREBBP, HMGN3, Gfi1, or Lhx8 (see **Table 6**). Together, these predictions strongly suggest that some of the proxy variants in LD with rs7937331 in the *SLC39A13* locus may be functional variants affecting AD by regulating RNA transcripts levels and affecting transcription factor binding.

Table 6. Detailed descriptions for SNPs with putative regulatory function (RegulomeDB score<3)

rsid (coordinate)	RegulomeDB Score*	Function	Left Gene - Right Gene	eQTL	Motifs	Protein Binding
rs1057233 (11:47376448)	1f	3' UTR	*MYBPC3-SLC39A13*	*SNRPG*	-	FOXA1
rs10742803 (11:47439938)	1f	downstream	*SLC39A13-PSMC3*	*SNRPG*	-	POLR2A
rs10769258 (11:47391039)	1f	intron (*SPI1*)	*MYBPC3-SLC39A13*	*SNRPG, C1QTNF4, SPI1*	-	-
rs11604680 (11:47457539)	1b	downstream	*PSMC3-RAPSN*	*SNRPG, C1QTNF4, SPI1*	RREB1, Foxl1	TCF4
rs11605672 (11:47413733)	1f	*intergenic*	*SPI1-SLC39A13*	*SNRPG, C1QTNF4, SPI1, MYBPC3*	-	-
rs12364432 (11:47902883)	1f	*intergenic*	*NUP160-PTPRJ*	*SNRPG, C1QTNF4*	-	-
rs12418852 (11:47868853)	1f	intron (*NUP160*)	*LOC100132562-PTPRJ*	*SNRPG, C1QTNF4, SPI1*	-	BLIMP1, HOXB3, SP1
rs12419692 (11:47624714)	1f	*intergenic*	*C1QTNF4-MTCH2*	*SNRPG, C1QTNF4, SPI1, MYBPC3, NR1H3*	-	-
rs1317149 (11:47486885)	1f	*downstream*	*RAPSN-CUGBP1*	*SNRPG, C1QTNF4, SPI1*	-	-
rs2053979 (11:47439444)	1b	downstream	*SLC39A13-PSMC3*	*SNRPG, C1QTNF4, SPI1*	E47	POLR2A
rs2293578 (11:47437403)	2b	3' UTR (*SLC39A13*)	*SPI1-PSMC3*	-	EWSR1-FLI1	POLR2A
rs2856650 (11:47365199)	1f	intron (*MYBPC3*)	*MADD-SPI1*	*AMBRA1, SNRPG, C1QTNF4*	SREBP, SREBP1, SREBP2	-
rs35184771 (11:47475189)	2b	*upstream*	*RAPSN-CUGBP1*	-	CACCC-bindingfactor, ZIC1	SP1,NFYA, NFYB
rs3781625 (11:47443080)	1f	intron (*PSMC3*)	*SLC39A13-RAPSN*	*SNRPG, C1QTNF4, SPI1*	-	POLR2A
rs3781626 (11:47442893)	1f	intron (*PSMC3*)	*SLC39A13-RAPSN*	*SNRPG, C1QTNF4, SPI1*	-	POLR2A
rs4752797 (11:47874364)	1f	*upstream*	*NUP160-PTPRJ*	*MYBPC3, NR1H3, SPI1*	-	-
rs4752801 (11:47907641)	1f	*intergenic*	*NUP160-PTPRJ*	*SNRPG, C1QTNF4*	-	TCF4, FOS
rs4752990 (11:47410393)	1b	*intergenic*	*SPI1-SLC39A13*	*SNRPG, C1QTNF4, SPI1*	PLAG1	STAT3
rs4752993 (11:47410951)	1f	*intergenic*	*SPI1-SLC39A13*	*SNRPG, C1QTNF4, SPI1*	-	-
rs4752999 (11:47428565)	1b	upstream	*SPI1-SLC39A13*	*SNRPG, C1QTNF4, SPI1, MYBPC3*	RFX1	CTCF, POLR2A, POU2F2, CEBPB, MAFK
rs55876153 (11:47416636)	2b	*intergenic*	*SPI1-SLC39A13*	-	Zfx	POLR2A, CREBBP, HMGN3
rs755553 (11:47432303)	1f	intron (*SLC39A13*)	*SPI1-PSMC3*	*SNRPG, C1QTNF4, SPI1*	-	-
rs7940536 (11:47395240)	1f	intron (*SPI1*)	*MYBPC3-SLC39A13*	*SNRPG, C1QTNF4, SPI1*	-	Gfi1 , Lhx8
rs896816 (11:47394338)	1f	intron (*SPI1*)	*MYBPC3-SLC39A13*	*SNRPG, C1QTNF4, SPI1*	HOXA5	-

DegulomeDB scores and detailed annotations for putative regulatory SNPs in high LD ($R^2 \geq 0.8$) with rs7937331 were shown. * For categories of RegulomeDB scores, please check Table S9.

3.4.7 *TREM2* and *TREML2* distinct effects on AD risk

The IGAP study also identified a suggestive association (rs9381040, $p=6.3\times10^{-7}$) located in the intergenic region between *TREM2* and *TREML2*; however, it was not possible to determine whether this intergenic signal was independent of p.*R47H* in *TREM2* due to the study design (a meta-analysis). Moreover, LD prevents the identification of functional variants and genes within this region. To identify potential functional coding variants within this region, we first analyzed the whole exome sequencing data from the Knight-ADRC study (46 AD cases and 39 controls) and UK-NIA study (143 AD cases and 186 controls). Eight coding variants were validated in the *TREML2* gene (see **Table 7**), which constitute 53% (8/15) of the missense variants reported for *TREML2* gene in the Exome Variant Server (release ESP6500SI-V2) for European Americans. Only 3 variants exhibit a MAF higher than 1%: p.V25A (MAF=5%), p.T129S (MAF=4.5%), and p.S144G (MAF=30%). Interestingly, according to our exome sequencing results, all of these variants are more common in controls than in cases; however they did not reach statistical significance because of the small sample size (see **Table 7**). Interestingly, the missense variant p.S144G (rs3747742) exhibited the highest LD (D'=0.86; R^2=0.73) with the GWAS SNP, rs9381040 (see **Table 7**), and exhibited the highest MAF among the validated missense variants in *TREML2*, which made it suitable for further analysis. Next, we performed a meta-analysis of the data from the ADGC, GERAD, EADI, and the Alzheimer's Research UK studies (16,254 cases and 20,052 control subjects); we found that the minor alleles of both rs9381040 ($p=1.21\times10^{-5}$; OR=0.92 [0.88~0.95], and rs3747742 ($p=8.66\times10^{-5}$; OR=0.93[0.89~0.96]) reduce risk for AD (see **Figures 2A and 2B**). When rs3747742 is included in a logistic regression model as a covariate, rs9381040 is no longer significant (p=0.43), and vice-versa, indicating that these SNPs tag the same signal. In addition, *TREM2*-p.R47H (rs75932628) was successfully imputed (imputation quality score information=0.84 and 0.79) in the GERAD and EADI studies, and displays a strong association with AD risk ($p=1.3\times10^{-3}$; OR=1.92 [1.29~2.85]) (see **Figure 2C**). When rs3747742 or rs9381040 are included as covariates in a conditional analysis, rs75932628 remains highly significant ($p=1.27\times10^{-4}$ and $p=1.19\times10^{-4}$ respectively) (see **Figure 2D**), which suggests that the *TREML2* and *TREM2* signals are independent from each other.

We then performed a linear regression analysis for rs9381040 and rs3747742 with CSF tau and $ptau_{181}$ levels (N = 1,269: 501 from Knight-ADRC, 394 from ADNI, 323 from UW, and 51 from Penn)[22]. Rs9381040 ($p=4.11\times10^{-4}$, β=-0.02) and rs3747742 ($p=1.4\times10^{-4}$, β=-0.02) both exhibit a strong association with CSF $ptau_{181}$

levels. The respective associations with CSF $ptau_{181}$ levels are no longer significant when either SNP is included as a covariate in the conditional analysis. These results confirm via 2 independent datasets that the associations of rs9381040 and rs3747742 with CSF biomarker levels and with AD risk represent the same signal. The *TREM2*-p.R47H variant was also genotyped in a subset of the CSF samples (N=835). In these samples, 3 variants, rs9381040 (p=0.04, β=-0.02) (see **Figure 3A**), rs3747742 (p=0.02, β=-0.02) (see **Figure 3B**), and rs75932628 (p=0.0016, β=0.2) (see **Figure 3C**) demonstrate a nominally significant association with CSF $ptau_{181}$ levels. To determine whether the *TREML2* signal (rs3747742) is independent of *TREM2*-p.R47H, we removed all of the p.R47H carriers from the analysis. Rs3747742 remained significantly associated with CSF $ptau_{181}$ levels (p=0.03) (see **Figure 3D**). Furthermore, when *TREM2*-p.R47H was included in the model as a covariate for rs3747742 analysis, the association remained significant (p=0.02), which suggests that the *TREM2* and *TREML2* signals are independent. Importantly, these associations confirmed the direction of the effect on CSF $ptau_{181}$ levels: the minor allele of rs3747742 is associated with lower $ptau_{181}$ levels (β=-0.02) and is predicted to be protective for AD risk (OR=0.91 [0.86~0.97]), while the minor allele of *TREM2*-p.R47H is associated with an increased risk for AD (OR=1.91 [1.85~1.97]) and higher levels of CSF $ptau_{181}$ levels (β=0.2). Together, these results demonstrate that the associations of missense variants in *TREM2* and *TREML2* with AD risk are independent. Our analyses also suggest that the AD-associated GWAS signal is more likely driven by the *TREML2* coding missense variant p.S144G (rs3747742) than by rs9381040, the most significant GWAS variant in the *TREM2*-*TREML2* region.

Table 7. *TREML2* variants identified by exome-sequencing

Location in chr6	rs#	AA change	EVS, MAF	AD (n=189)		Controls (n=225)		OR (95% CI)	p	LD with rs9381040		Condel	Sift	Polyphen
				Hets	MAF	Hets	MAF			R^2	D'			
41166154	rs77704965	D23G	0.22	0	0%	4	2%	NA	0.17	0.018	1	Neutral	Tolerated	Benign
41166149	rs62396355	V25A	5.05	6	3%	15	7%	0.45 (0.17-1.2)	0.11	0.018	1	Neutral	Tolerated	Benign
41166075	rs35512890	M50V	NA	16	8%	27	12%	0.67 (0.35-1.3)	0.24	NA	NA	Neutral	Tolerated	Benign
41162562	rs61734887	S129T	4.52	12	6%	22	10%	0.62 (0.30--1.3)	0.2	0.051	1	Neutral	Tolerated	Benign
41162538	NA	L137H	NA	0	0%	1	0%	NA	0.35	NA	NA	Neutral	Tolerated	Benign
41162518	rs3747742	S144G	30.44	82	43%	104	47%	0.89 (0.6-1.31)	0.56	0.67	0.86	Neutral	Tolerated	Benign
41162371	rs145455750	T193A	0.27	0	0%	1	0%	NA	0.35	NA	NA	Neutral	Tolerated	Benign
41162204	rs115991880	S248A	0.34	2	1%	5	2%	0.47 (0.09-2.45)	0.36	0	0	Deleterious	Deleterious	Benign

Coding variants in *TREML2* were extracted from 46 unrelated AD cases and 39 unrelated controls from the Knight-ADRC study and from 143 unrelated AD cases and 183 unrelated controls from the NIA-UK exome-sequencing study. R^2 and D' values reported here are coming from the Pilot 1 of the 1000K genome project. Key: AD, Alzheimer's disease; CI, confidence interval; LD, linkage disequilibrium; OR, odds ratio.

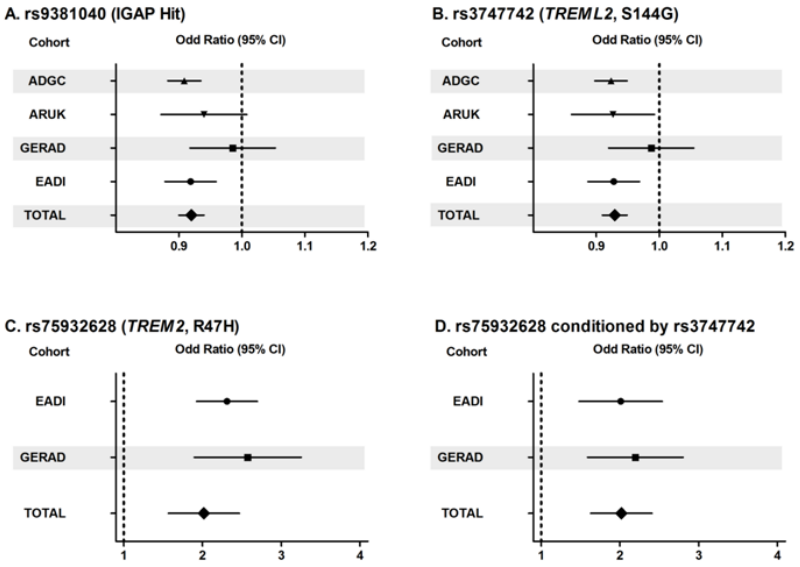


Figure 2: Odds ratio for rs9381040 (IGAP hit), rs3747742 (*TREML2*, p.S144G), and rs75932628 (*TREM2*, p.R47H) among AD patient , as compared with control subjects, at each study center and overall. Shown are the combined estimates of the AD risk of possessing rs9381040 (IGAP hit), combined odds ratios analyses were homogeneous (p = 0.69, by Woolf test for heterogeneity). Panel (A), the rs3747742 (*TREML2*, p.S144G) (p = 0.81, by Woolf test for heterogeneity), panel (B), the rs75932628 (*TREM2*, p.R47H) (p = 0.97, by Woolf test for heterogeneity), panel (C), rs75932628 (*TREM2*, p.R47H) after conditioning for rs3747742 (*TREML2*, p.S144G) panel (D). The triangles represent ADGC study, the inverted triangles represent ARUK study, squares represent GERAD study, circles represent EADI study and the diamonds represent the summary odds ratio. The horizontal lines indicate the 95% confidence intervals of the estimates. Abbreviations: ADGC, Alzheimer's disease genetic consortium; ARUK, Alzheimer's Research UK; EADI, European Alzheimer's disease initiative; IGAP, international genomics of Alzheimer's project; GERAD, genetic and environmental risk for Alzheimer's disease.

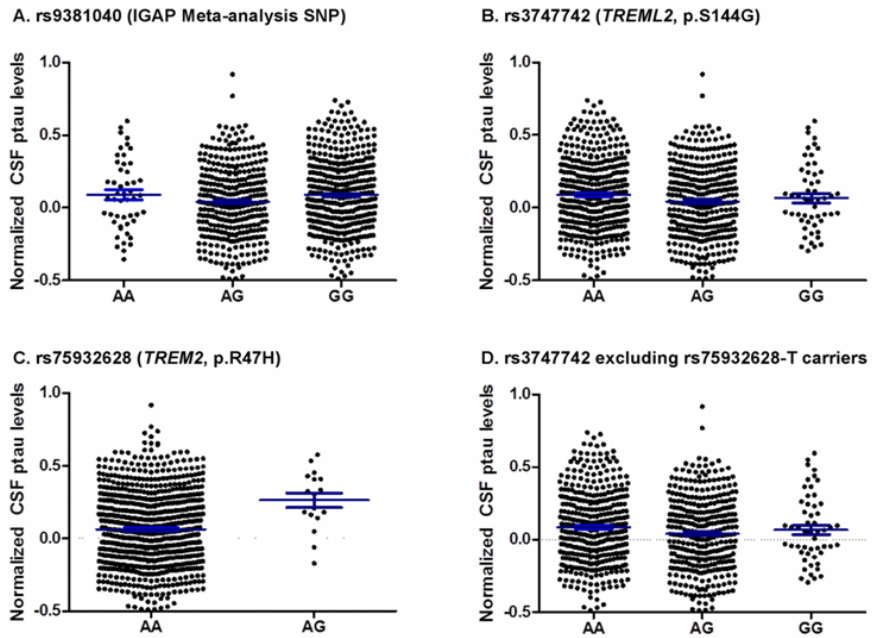


Figure 3: Association of *TREM2* and *TREML2* variants with CSF ptau$_{181}$ levels. Panel (A) CSF ptau$_{181}$ levels by rs9381040 genotype (IGAP meta-analysis most significant SNP). AG + GG versus AA p = 0.04. (Panel B) CSF ptau$_{181}$ levels by rs3747742 genotype (*TREML2*, missense variant p.S144G). AG + GG versus AA p = 0.02. Panel (C) CSF ptau$_{181}$ levels by
rs75932628 genotype (*TREM2*, missense variant p.R47H). AG versus AA p = 0.0016. Panel (D) CSF ptau$_{181}$ levels by rs3747742 genotype (*TREML2*, missense variant p.S144G). AG +GG versus AA excluding the variant p.R47H carriers p = 0.03. The mean and the standard error of the mean (SEM) for the normalized residuals CSF ptau$_{181}$ levels are shown in blue. Abbreviations: CSF, cerebrospinal fluid; IGAP, international genomics of Alzheimer's project; SNP, single nucleotide polymorphisms.

3.5 DISCUSSION

Using CSF biomarkers as endophenotypes in genetic studies has been shown to be a useful method to generate testable hypotheses regarding biological mechanisms attributable to identified genetic variants and to identify novel variants associated with different aspects of disease[17-23]. To test the hypothesis that GWAS-identified AD risk variants are also associated with CSF $A\beta_{42}$ and $ptau_{181}$ biomarkers at SNP- and SNP set- levels. Power analyses (see **Table S2**) suggest that we have >80% power to find a statistically significant, additive effect of a 0.21 fold-difference between the means when the alpha level is 0.05 for all SNPs in this study. Assuming an extremely conservative alpha level of 10^{-5}, all SNPs in this study still achieve power of at least 80% for a 0.365 delta (differences in standard deviation). In order to distinguish the effects of GWAS-identified genes from flanking genomic regions, we carefully selected fine mapping regions taking into account the recombination rates. To avoid using the conservative and stringent Bonferroni correction which can lead to false negative association, we utilized the simpleM algorithm [157] to calculate the number of independent association tests within the fine mapping regions.

For the top SNPs identified in the IGAP study, we found evidence of association for *ABCA7* (rs4147929), *FERMT2* (rs17125944), and *INPP5D* (rs28655385) with CSF $A\beta_{42}$ levels and evidence of association for *CR1* (rs6656401), *FERMT2* (rs17125944), *MEF2C* (rs190982), *PICALM* (rs10792832), *PTK2B* (rs28834970), and SORL1 (rs11218343) with CSF $ptau_{181}$ levels. However, the top SNPs within each locus in the IGAP study are not the most significant SNPs within each locus in our CSF biomarker analyses (see **Tables 2** and **3**)[8]. After correcting for multiple testing, the most significant SNPs within the *CELF1*, *EPHA1*, and *FERMT2* fine-mapping regions are significantly associated with CSF $A\beta_{42}$ levels. The first association appears to be through rs7937331, located in an intron of *SLC39A13*. We showed that *SLC39A13*-rs7937331, a SNP significantly associated with CSF $A\beta_{42}$ levels, and *CELF1*-rs10838725, the IGAP top SNP significantly associated with AD risk, tag the same association. Additionally, although the association between SNPs in *SLC39A13* and AD risk was not genome-wide significant ($p= 5.11\times10^{-7}$) it was still strong in the IGAP analysis[8]. Set-based analysis for the *CELF1* fine-mapping supports the association with CSF $A\beta_{42}$ levels. Together, our result suggests that *SLC39A13* may be a novel candidate gene affecting AD through an $A\beta$ dependent mechanism. The second association is through rs7802536, located in the intergenic region between *CLCN1* and *FAM131B*. The association for this intergenic signal is independent of

EPHA1. Set-based analysis for the *EPHA1* fine-mapping region provides extra evidence of association with CSF $A\beta_{42}$ levels. However, this intergenic SNP is not associated with AD risk in the IGAP study, which suggests that the signal within this intergenic region may be a false positive finding. The last significant association is through rs62003531, located in the intronic region of *FERMT2*. Our analyses further confirm that rs62003531 tags the same signal as the IGAP top SNP in *FERMT2*. Set-based analysis for the *FERMT2* fine-mapping region supports the association with CSF $A\beta_{42}$ levels. Our data clearly suggest that there is only one signal within the *FERMT2* region and that the functional allele affects AD risk through an Aβ-dependent mechanism.

We also confirmed that the effect size and direction of effect for SNPs that surpassed multiple test correction thresholds are similar and consistent in each locus by conducting stratified, case-only, and control-only analyses (see **Tables S11** and **S12**). Genome partition analyses suggest that *CELF1*-rs7937331, *EPHA1*-rs7802536, and *FERMT2*-rs62003531 only explain 0.34% of the phenotypic variability for CSF $A\beta_{42}$ levels (31.64% explained by the GWAS chip as a whole), suggesting that additional genetic variants, which may include a combination of common, low frequency, rare, or structural variants remain unidentified.

BIN1 was previously reported to be associated with AD risk via a tau dependent mechanism [165]. Moreover, a recent study performed functional screening in drosophila and found that the fly orthologs of *CD2AP*, *CELF1*, and *FERMT2* modulate tau metabolism [166]. In our analyses, neither the most significant variants in our analyses nor the IGAP top SNPs in these genes surpassed multiple test correction thresholds for association with CSF $ptau_{181}$ and tau levels (see **Tables 2** and **4**) which may be due to lack of statistical power.

Several recent studies have shown that the RegulomeDB database[160] is a comprehensive and fast tool to annotate noncoding variants and predict the potential regulatory function associated with SNPs of interest[142,160]. In this study, we used RegulomeDB to examine whether IGAP-identified top SNPs and our CSF-identified top SNPs may have any potential regulatory functional impact. We found that several proxy SNPs in LD with rs7937331 in the *SLC39A13* region are implicated as cis-eQTLs for nearby genes and may affect transcription factor binding. However, these regulatory function predictions are primarily based on data in blood and cancer cells. Further functional studies using brain tissues are essential to elucidate whether these non-coding variants have any effects in human brain and thus contribute to AD.

Our power calculation predicts a high likelihood of detecting a significant association in the single-variant analyses given our sample size. Plausible explanations for the lack of significant association with CSF $ptau_{181}$ levels include: (1) Current samples size is still not large enough to identify a signal with CSF $ptau_{181}$ levels given the differences in measurement of $ptau_{181}$ across studies; (2) variants that affect AD risk by modulating tau pathology may be involved in other processes rather than affecting CSF $ptau_{181}$ levels; (3) underlying functional variants affecting CSF $ptau_{181}$ levels are low frequency, rare, or common variants with very small effects; (4) assuming an additive genetic effects model did not detect the association with CSF $ptau_{181}$ levels well; (5) the association for CSF $ptau_{181}$ levels involves gene-gene or gene-environment interaction that cannot be captured using our study design.

For the *TREM2* and *TREML2* regions, our results demonstrate that the associations of missense variants in *TREM2* and *TREML2* with AD risk are independent. Moreover, our analyses suggest that the AD-associated GWAS signal is likely driven by the *TREML2* missense variant p.S144G (rs3747742); it results in a similar OR to rs9381040. We also validated 2 other coding variants p.V25A and p.S129T in *TREML2* gene in moderate LD (R^2=0.05 and D'=1) with the GWAS SNP, which both exhibited a higher frequency among control subjects than in AD cases (see **Table 7**). However, for both variants we only obtained data by whole-exome sequencing which limited our analysis about the role that these variants may play in the association of *TREML2* with AD risk. To prove that these additional variants are associated with AD risk we will need a larger sample size. Additionally, the purpose of this study was to find a functional coding variant in the *TREML2* gene that could explain the association for *TREML2* which was found in the recent IGAP meta-analysis. Our data suggest that there is a coding variant in *TREML2* that could explain the GWAS signal, but our data cannot rule out the presence of functional variants outside of the coding region.

In summary, our fine-mapping analyses in the *APOE* locus identified a genome-wide significant association for rs769449 in *APOE* locus with both CSF $A\beta_{42}$ and $ptau_{181}$ independently of *APOE*-ε2 and *APOE*-ε4 SNPs. We also showed that at least part of the association for rs769449 with CSF $ptau_{181}$ levels is Aβ-independent. Our locus-specific analyses identified signals within the *CELF1*, *EPHA1*, and *FERMT2* fine-mapping regions significantly associated with CSF $A\beta_{42}$ levels after correction for multiple testing. Set-based analyses within the *CELF1*, *EPHA1*, and *FERMT2* fine-mapping regions also supported the significant associations. Additionally, the RegulomeDB database showed that several proxy SNPs for rs7937331 in *SLC39A13* may have potential regulatory effects. However, our data lack statistical power to detect any significant association with CSF $ptau_{181}$ levels. We

also showed that *TREM2*- p.R47H is associated with increased risk for AD (OR=1.91 [1.85~1.97]) and *TREML2*-p.S144G is associated with reduced risk for AD (OR=0.91 [0.86~0.97]). The mechanisms by which these variants influence AD risk are not currently understood, but it has been suggested that modulation of microglial activation might influence clearance of Aβ. These results underline the importance of the inflammatory response in modulating risk for AD and suggest that other genes in this gene family may also harbor risk alleles for AD.

SUPPLEMENTAL TABLES

Table S1. Fine-mapping region and multiple test threshold for each gene

Gene	Chr	Left Boundary	Right Boundary	Fine-mapping region (kb)	# SNPs	# independent SNPs	Multiple testing threshold
ABCA7	19	1038925	1093453	54.528	73	48	1.04×10^{-3}
APOE	19	45357291	45447145	89.854	118	57	8.77×10^{-4}
BIN1	2	127750462	127896232	145.77	391	125	4.00×10^{-4}
CD2AP	6	47310938	47709696	398.758	803	178	2.81×10^{-4}
CD33	19	51676649	51780994	104.345	113	44	1.14×10^{-3}
CLU	8	27404137	27504956	100.819	149	64	7.81×10^{-4}
CR1	1	207568023	207878272	310.249	444	95	5.26×10^{-4}
CELF1	11	47412614	47639525	226.911	210	47	1.06×10^{-3}
CASS4	20	54977598	55148084	170.486	243	91	5.49×10^{-4}
EPHA1	7	143025985	143132563	106.578	127	65	7.69×10^{-4}
FERMT2	14	53071715	53474112	402.397	461	157	3.18×10^{-4}
HLA_DRB5/DRB1	6	32427789	32682019	254.23	3350	1053	4.75×10^{-5}
IGHV1-67	14	107122925	107169570	46.645	126	25	2.00×10^{-3}
INPP5D	2	233805812	234121692	315.88	515	213	2.35×10^{-4}
MS4A6A	11	59744901	60107747	362.846	618	138	3.62×10^{-4}
MEF2C	5	87978904	88223420	244.516	195	57	8.77×10^{-4}
NME8	7	37405370	37983235	577.865	1070	261	1.92×10^{-4}
PICALM	11	85601792	85908139	306.347	476	170	2.94×10^{-4}
PTK2B	8	27067745	27404137	336.392	599	152	3.29×10^{-4}
SORL1	11	121204601	121661507	456.906	521	165	3.03×10^{-4}
SLC24A4	14	92758445	93010112	251.667	588	196	2.55×10^{-4}
TP53INP1	8	95793022	96204513	411.491	677	184	2.72×10^{-4}
ZCWPW1	7	99816488	100159567	343.079	242	73	6.85×10^{-4}

The fine-mapping region was defined as the region between the closest recombination hot spots in either direction based on the estimated recombination rate from HapMap samples (May 2008). Left and right boundaries were annotated based on hg 19 build. The simpleM method was used to calculate the number of independent signals within the fine-mapping region[157]. The multiple testing threshold was defined as 0.05 divided by the number of independent SNPs. CHR, chromosome.

Table S2. Power calculation

Minor allele frequency	Effect size when power = 0.8	
	Alpha = 0.05	Alpha = 10^{-5}
0.10	0.21	0.365
0.15	0.175	0.308
0.20	0.157	0.275
0.25	0.146	0.255
0.30	0.137	0.24
0.35	0.132	0.23
0.40	0.129	0.224
0.45	0.126	0.222
0.50	0.125	0.22

Power estimation for genetic association detection. Power was calculated using an overall F test in a one-way, three-group analysis of variance. The effect size which was measured in fold-difference between the means by assigning statistical power to 80% for minor allele frequencies from 0.1 to 0.5 and Type I error equal to 0.05 and 0.00001.

Table S3. Top SNP in the *APOE* locus in each study

Study	Top SNP for CSF $A\beta_{42}$	$P_{A\beta}$	Top SNP for CSF $ptau_{181}$	P_{ptau}
ADNI-1	rs769449	1.42×10^{-15}	rs2075650	2.90×10^{-6}
ADNI-2	rs429358	4.81×10^{-28}	rs429358	1.67×10^{-12}
Mayo	rs429358	2.73×10^{-16}	rs429358	1.30×10^{-5}
UW	rs769449	6.92×10^{-17}	rs769449	6.48×10^{-9}
Knight-ADRC	rs769449	2.96×10^{-23}	rs769449	1.10×10^{-8}

Association for *APOE* with CSF $A\beta_{42}$ and $ptau_{181}$ levels was examined using a multivariate linear regression adjusting for age, gender, and PCs in each study.

Table S4. Top SNP in the *APOE* locus in the combined analysis

SNP	MAF (%)	$P_{A\beta}$	P_{ptau}
rs769449	17.88	7.60×10^{-79}	3.31×10^{-32}

Top SNP in APOE locus, minor allele frequency (MAF), and association with CSF $A\beta_{42}$ and $ptau_{181}$ levels were shown. Multivariate linear regression adjusting for age, gender, PCs and dummies for sites were performed to evaluate the association.

Table S5. Linkage disequilibrium between selected SNPs

Gene	SNP1	SNP2	D'	R^2
ABCA7	rs4147929	rs3764650	0.66	0.22
ABCA7	rs4147929	rs76348507	0.69	0.23
ABCA7	rs3764650	rs76348507	1	0.99
CR1	rs6656401	rs6701713	1	1
CR1	rs6656401	rs4844610	1	0.95
CELF1	rs10838725	rs7937331	0.97	0.21
EPHA1	rs11771145	rs7802536	0.01	0
FERMT2	rs17125944	rs62003531	0.95	0.7

Table S6. Conditional analyses for selected SNPs

ABCA7	Separate Analysis		Conditioning on rs76348507		Conditioning on rs4147929	
rsid	$Beta_{A\beta}$	$P_{A\beta}$	$Beta_{A\beta}$	$P_{A\beta}$	$Beta_{A\beta}$	$P_{A\beta}$
rs76348507	-0.026 (-0.042,-0.009)	2.74×10^{-3}	NA	NA	-0.019 (-0.037,0)	5.52×10^{-2}
rs4147929	-0.017 (-0.030,-0.004)	1.00×10^{-2}	-0.010 (-0.025,0.004)	1.66×10^{-1}	NA	NA

ABCA7	Separate Analysis		Conditioning on rs4147929		Conditioning on rs3764650	
rsid	$Beta_{A\beta}$	$P_{A\beta}$	$Beta_{A\beta}$	$P_{A\beta}$	$Beta_{A\beta}$	$P_{A\beta}$
rs4147929	-0.017 (-0.03,-0.004)	1.00×10^{-2}	NA	NA	-0.011 (-0.026,0.004)	1.41×10^{-1}
rs3764650	-0.024 (-0.041,-0.008)	4.24×10^{-3}	-0.017 (-0.036,0)	7.36×10^{-2}	NA	NA

ABCA7	Separate Analysis		Conditioning on rs3764650		Conditioning on rs76348507	
rsid	$Beta_{A\beta}$	$P_{A\beta}$	$Beta_{A\beta}$	$P_{A\beta}$	$Beta_{A\beta}$	$P_{A\beta}$
rs3764650	-0.024 (-0.041,-0.008)	1.00×10^{-2}	NA	NA	NA	NA
rs76348507	-0.026 (-0.042,-0.009)	4.24×10^{-3}	NA	NA	NA	NA

CR1	Separate Analysis		Conditioning on rs6701713		Conditioning on rs6656401	
rsid	$Beta_{ptau}$	P_{ptau}	$Beta_{ptau}$	P_{ptau}	$Beta_{ptau}$	P_{ptau}
rs6701713	0.024 (0.009,0.039)	1.88×10^{-3}	NA	NA	NA	NA
rs6656401	-0.006 (-0.019,0.006)	3.47×10^{-1}	NA	NA	NA	NA

CELF1	Separate Analysis		Conditioning on rs7937331		Conditioning on rs10838725	
rsid	$Beta_{A\beta}$	$P_{A\beta}$	$Beta_{A\beta}$	$P_{A\beta}$	$Beta_{A\beta}$	$P_{A\beta}$
rs7937331	0.021 (0.01,0.032)	2.30×10^{-4}	NA	NA	0.02 (0.008,0.032)	1.25×10^{-3}
rs10838725	-0.009 (-0.02,0.002)	1.09×10^{-1}	-0.001 (-0.014,0.011)	8.67×10^{-1}	NA	NA

EPHA1	Separate Analysis		Conditioning on rs7802536		Conditioning on rs11771145	
rsid	$Beta_{A\beta}$	$P_{A\beta}$	$Beta_{A\beta}$	$P_{A\beta}$	$Beta_{A\beta}$	$P_{A\beta}$
rs7802536	-0.025 (-0.037,-0.013)	6.01×10^{-5}	NA	NA	-0.025 (-0.037,-0.013)	5.97×10^{-5}
rs11771145	0.002 (-0.008,0.013)	6.70×10^{-1}	0.002 (-0.008,0.013)	6.53×10^{-1}	NA	NA

FERMT2	Separate Analysis		Conditioning on rs62003531		Conditioning on rs17125944	
rsid	$Beta_{A\beta}$	P	$Beta_{A\beta}$	P	$Beta_{A\beta}$	P
rs62003531	-0.033 (-0.049,-0.016)	8.50×10^{-5}	NA	NA	-0.041 (-0.070,-0.012)	5.85×10^{-3}
rs17125944	0.026 (0.005,0.046)	7.42×10^{-3}	0.011 (-0.021,0.043)	5.00×10^{-1}	NA	NA

Table S7. Conditional analyses for rs7937331 and rs10838725 in the *CELF1* fine-mapping region in the ADGC datasets

	Separate Analysis		Conditioning on rs7937331		Conditioning on rs10838725	
rsid	OR	P	OR	P	OR	P
rs7937331	0.93	4.14×10^{-4}	NA	NA	0.93	3.98×10^{-3}
rs10838725	0.96	4.63×10^{-2}	0.99	6.61×10^{-1}	NA	NA

Newly imputed ADGC datasets (1,000 Genomes Phase I March 2012 v3) were used for the conditional analyses. Effect sizes and p values were reported.

Table S8. Summary statistics for recently identified candidate variants

SNP	Chr:bp	Nearest gene	MAF (%)	Case-control association[&]	$Beta_{Aß}$ (95% CI)	$P_{Aß}$	$Beta_{Ptau}$ (95% CI)	P_{Ptau}
rs3764650	19:1046520	*ABCA7*	10.28	3.22×10^{-7}	-0.024 (-0.041,-0.008)	4.24×10^{-3}	0.006 (-0.014,0.025)	5.71×10^{-1}
rs6701713	1:207786289	*CR1*	19.50	3.48×10^{-14}	-0.006 (-0.019,0.007)	3.47×10^{-1}	0.024 (0.009,0.039)	1.88×10^{-3}
rs9349407	6:47453378	*CD2AP*	28.31	3.92×10^{-7}	0.007 (-0.005,0.018)	2.66×10^{-1}	0.003 (-0.01,0.017)	6.16×10^{-1}
rs11136000	8:27464519	*CLU*	39.57	1.72×10^{-16}	0.01 (0,0.021)	5.75×10^{-2}	-0.008 (-0.02,0.005)	2.21×10^{-1}
rs2304933	11:60102507	*MS4A6A*	33.30	5.35×10^{-5}	-0.008 (-0.018,0.003)	1.64×10^{-1}	0.012 (-0.001,0.025)	6.30×10^{-2}

These candidate variants were recently found in Shulman et al [163] and Lyzel et al [164]. Association with CSF $A\beta_{42}$ and $ptau_{181}$ levels was analyzed using multivariate linear regression adjusting for age, gender, first three principle components, and dummies for sites. Chr, chromosome; MAF, minor allele frequency; CI, confidence interval.*Proxy SNPs: rs9296559 for rs9349407 and rs7935829 for rs7232. [&]These p values were from stage 1 analyses of the recent IGAP study.

Table S9. Number of proxy SNPs in linkage disequilibrium with significant SNPs

Top SNP	Find-mapping region	Number of proxy SNPs (from SNAP)	Number of proxy SNPs (from CSF)
rs7937331	*CELF1*	38	29
rs7802536	*EPHA1*	0	0
rs62003531	*FERMT2*	0	39

Proxy SNPs were first searched using the SNAP database (https://www.broadinstitute.org/mpg/snap/ldsearch.php) or our CSF GWAS dataset based on linkage disequilibrium with significant or suggestive SNPs using the CEU populations from the 1000 Genome Pilot 1. Proxy SNPs were defined as SNPs with a R^2 greater or equal to 0.8 relative to selected SNPs.

Table S10. RegulomeDB Score Category

Score	Supporting data
1a	eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding / DNase peak
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak
2b	TF binding + any motif + DNase Footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
4	TF binding + DNase peak
5	TF binding or DNase peak
6	other

Score categories were obtained from http://regulome.stanford.edu/help.

Table S11. Stratified analyses for statistically significant SNPs with CSF $A\beta_{42}$ levels

	ADNI1 (N=357)		ADNI2 (N=349)		Mayo (N=426)		UW (N=290)		Knight-ADRC (N=614)	
Gene: SNP	Beta	$P_{A\beta}$	Beta	$P_{A\beta}$	Beta	$P_{A\beta}$	Beta	$P_{A\beta}$	Beta	$P_{A\beta}$
CELF1: rs7937331	0.029	6.90×10^{-3}	0.017	1.87×10^{-1}	0.027	3.05×10^{-2}	0.017	1.47×10^{-1}	0.011	3.89×10^{-1}
EPHA1:rs7802536	-0.028	2.14×10^{-2}	-0.024	1.04×10^{-1}	-0.003	8.46×10^{-1}	-0.008	5.51×10^{-1}	-0.042	8.99×10^{-4}
FERMT2:rs62003531	-0.041	1.75×10^{-2}	-0.008	6.84×10^{-1}	-0.042	1.17×10^{-2}	-0.007	7.00×10^{-1}	-0.045	2.10×10^{-2}

Association with CSF Aβ42 levels was analyzed using multivariate linear regression adjusing for age, gender, and first 3 PCs in each site. The effect size and p-values were given in the table.

Table S12. Case-only and control-only analyses for statistically significant SNPs with CSF $A\beta_{42}$ levels

	Default model		Case-only analysis		Control-only analysis	
Gene: SNP	Beta (95% CI)	$P_{A\beta}$	Beta (95% CI)	$P_{A\beta}$	Beta (95% CI)	$P_{A\beta}$
CELF1: rs7937331	0.021 (0.01,0.032)	2.30×10^{-4}	0.019 (0.003,0.034)	1.78×10^{-2}	0.022 (0.008,0.036)	1.77×10^{-3}
EPHA1: rs7802536	-0.025 (-0.037,-0.013)	6.01×10^{-5}	-0.012 (-0.029,0.005)	1.82×10^{-1}	-0.032 (-0.047,-0.016)	5.30×10^{-5}
FERMT2: rs62003531	-0.033 (-0.049,-0.016)	8.50×10^{-5}	-0.028 (-0.05,-0.005)	1.57×10^{-2}	-0.028 (-0.049,-0.006)	1.12×10^{-2}

Cases and controls were classified based on the Clinical Dementia Rating (CDR). Cases were defined as CDR>0 and controls were defined as CDR=0. Association with CSF Aβ42 levels was analysed using multivariate linear regression adjusting for age, gender, and first 3 PCs. The effect size and p-values were given in the table.

SUPPLEMENTAL FIGURES

Figure S1. Distributions of transformed CSF $A\beta_{42}$ and $ptau_{181}$ measurements.

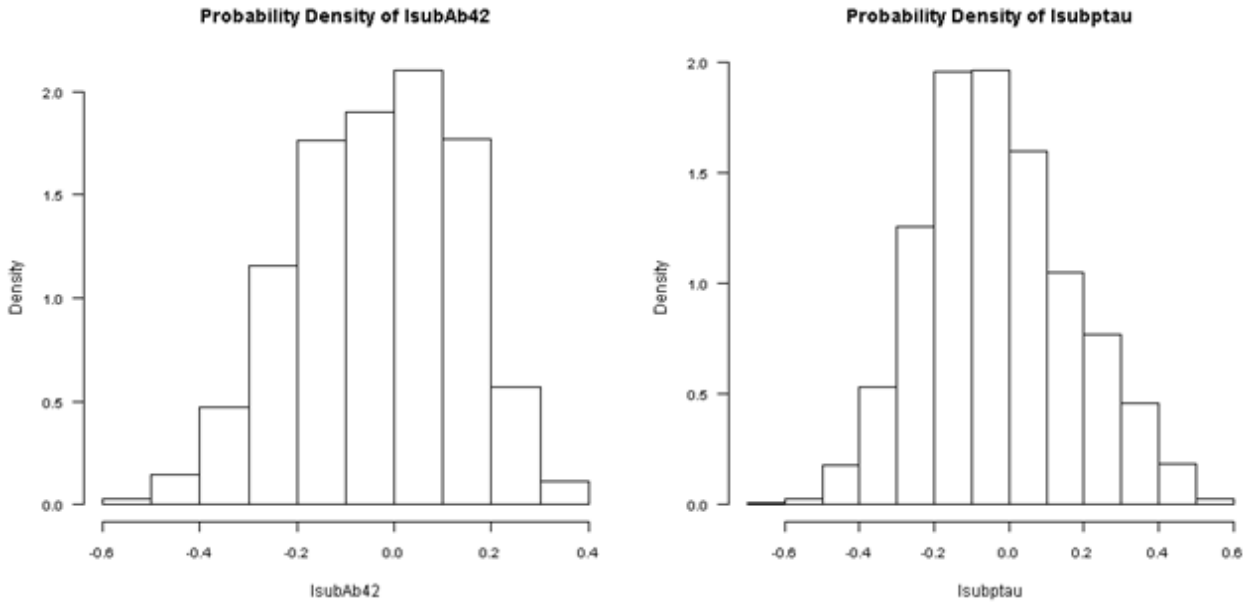
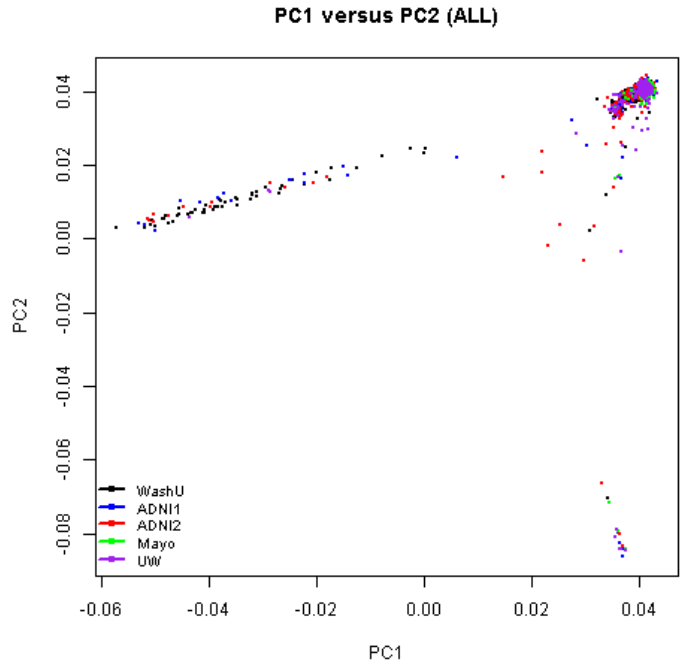
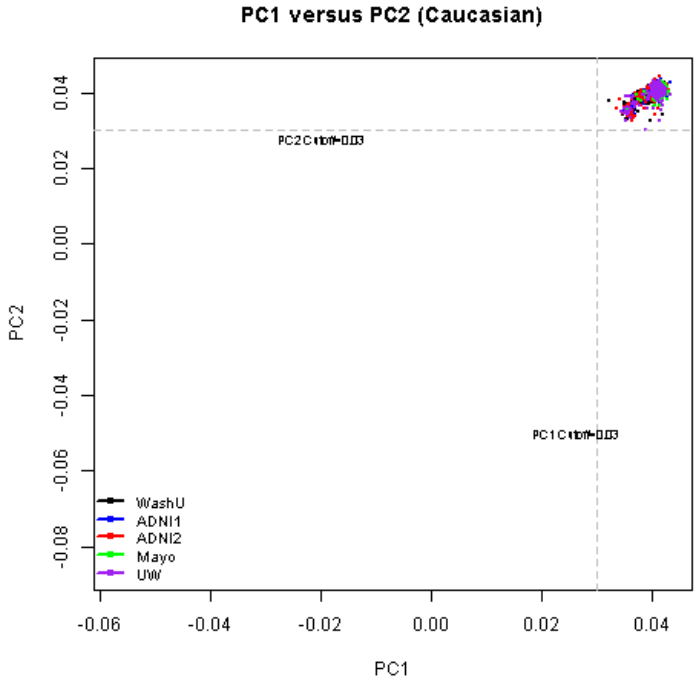


Figure S2. . Principal component analyses of the studied individuals in the CSF biomarker dataset (A-B) and ADGC case-control dataset (C-D)

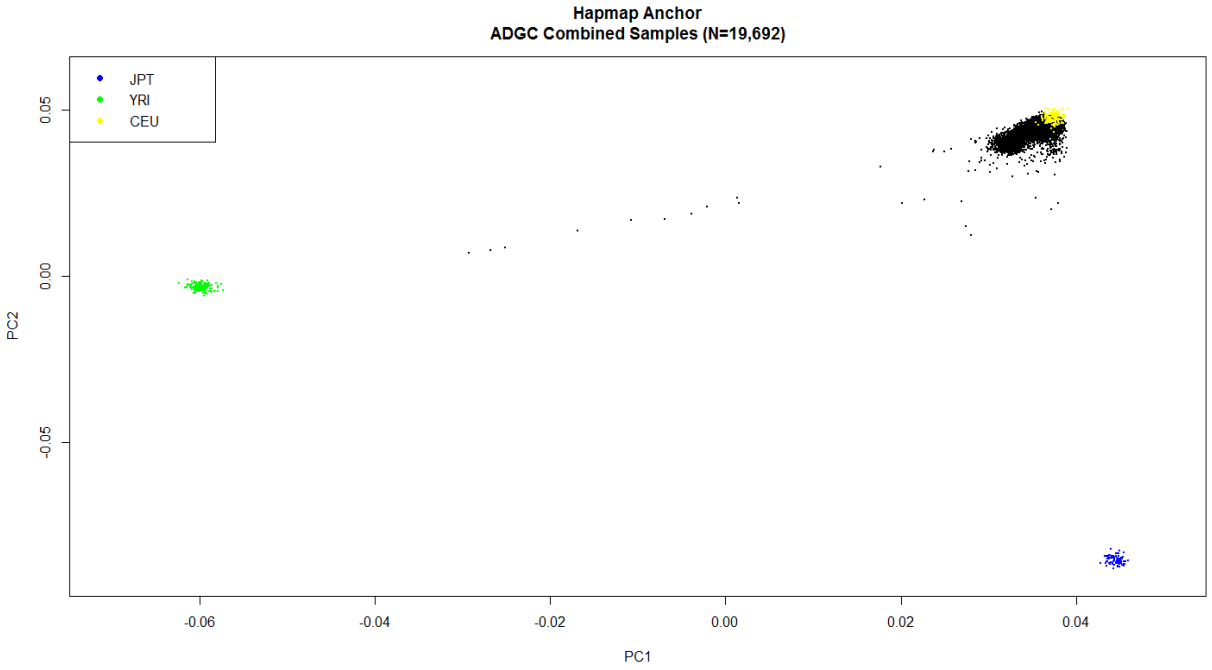
(A) All individuals



(B) Individuals of European American descent who were finally analyzed (N=2,036)



(C) All ADGC individuals



(D) ADGC Individuals of European American descent who were finally analyzed (N=19,673)

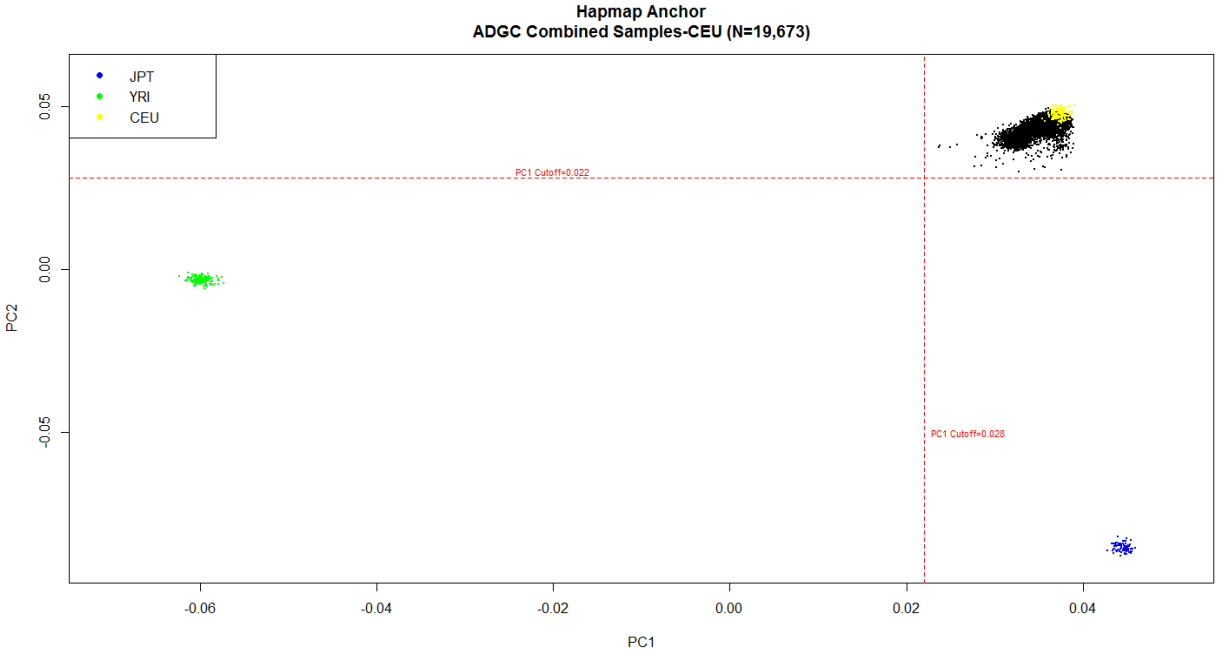
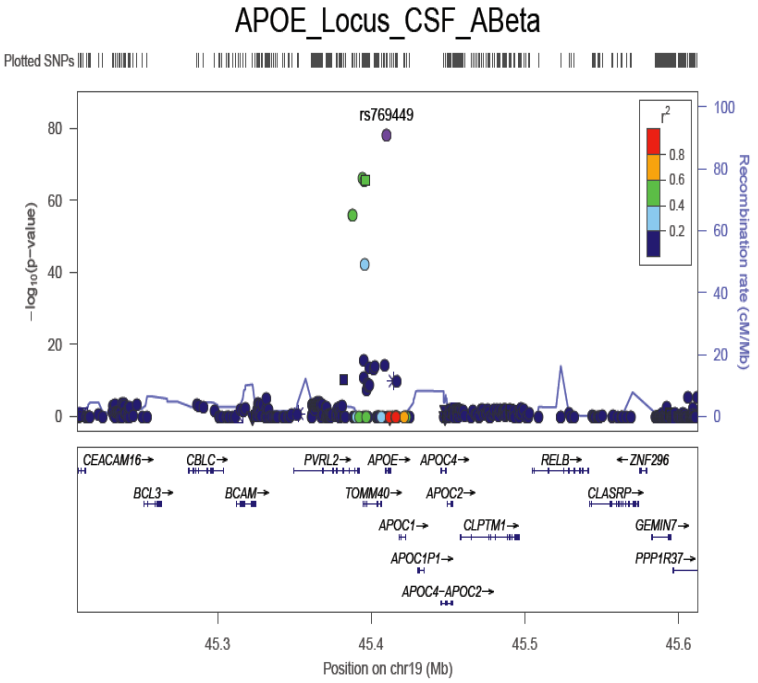
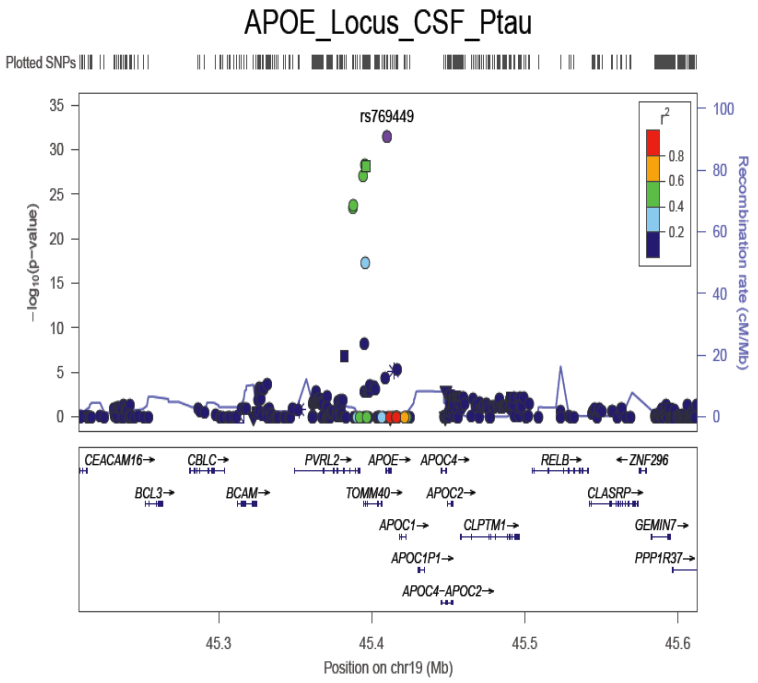


Figure S3. Regional plots for the *APOE* locus with CSF $A\beta_{42}$ and $ptau_{181}$ levels. (A) and (B) CSF $A\beta_{42}$ and $ptau_{181}$ associations for rs769449; (C) and (D) CSF $A\beta_{42}$ and $ptau_{181}$ association after conditioning on APOE-ε2 and *APOE*- ε4 SNPs.

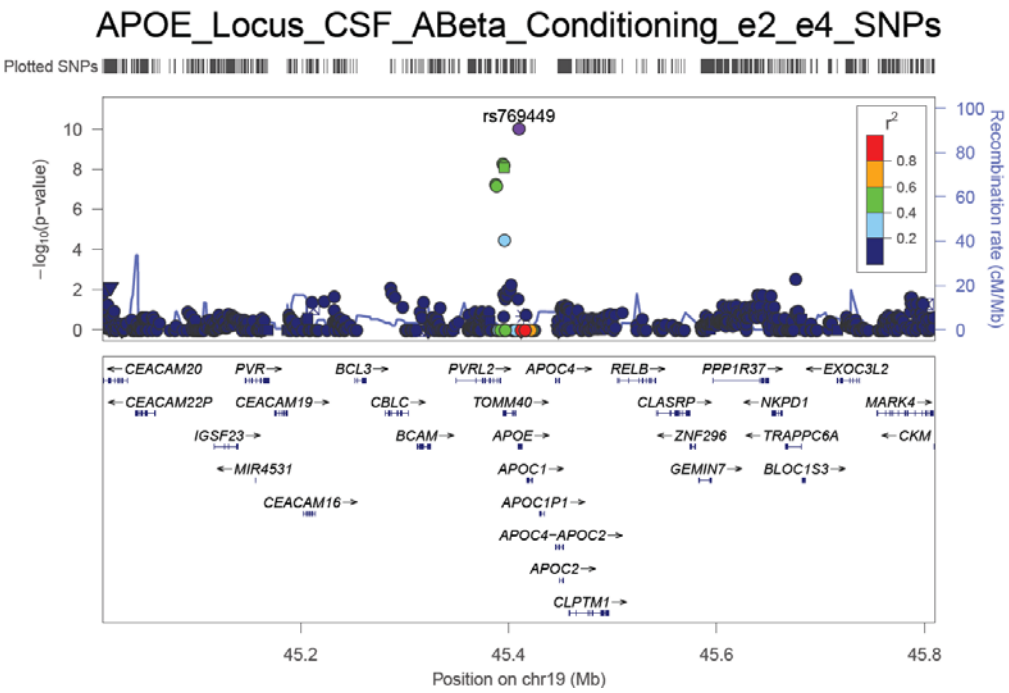
(A)



(B)



(C)


APOE_Locus_CSF_ABeta_Conditioning_e2_e4_SNPs
Plotted SNPs
rs769449
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
CEACAM20
PVR
BCL3
PVRL2
APOC4
RELB
PPP1R37
EXOC3L2
CEACAM22P
CEACAM19
CBLC
TOMM40
CLASRP
NKPD1
MARK4
IGSF23
BCAM
APOE
ZNF296
TRAPPC6A
CKM
MIR4531
APOC1
GEMIN7
BLOC1S3
CEACAM16
APOC1P1
APOC4−APOC2
APOC2
CLPTM1
45.2
45.4
45.6
45.8
Position on chr19 (Mb)

(D)

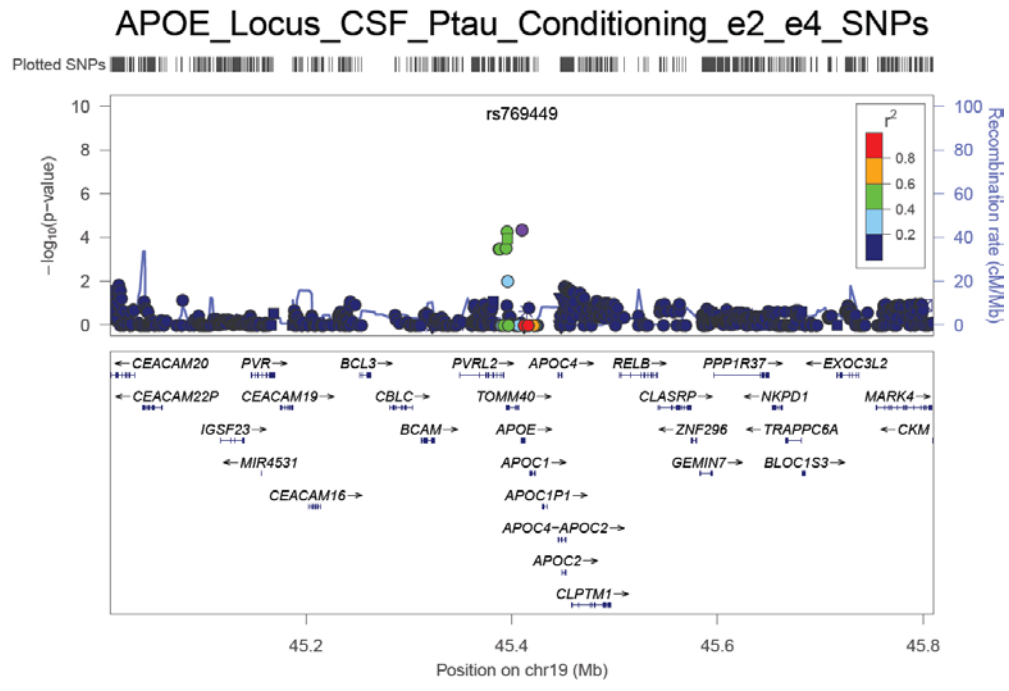
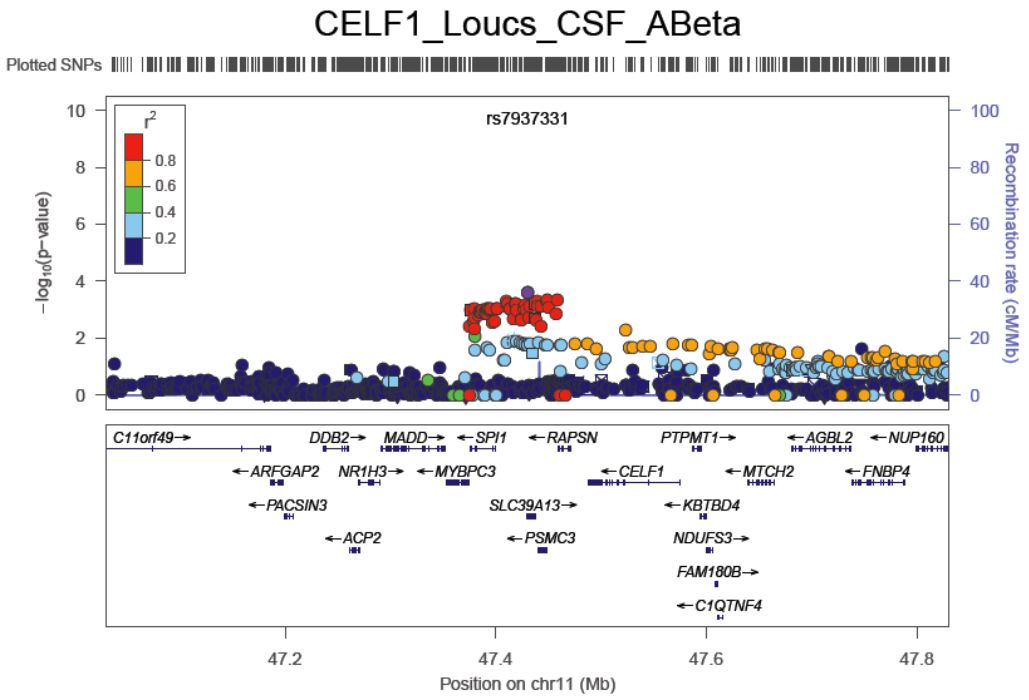
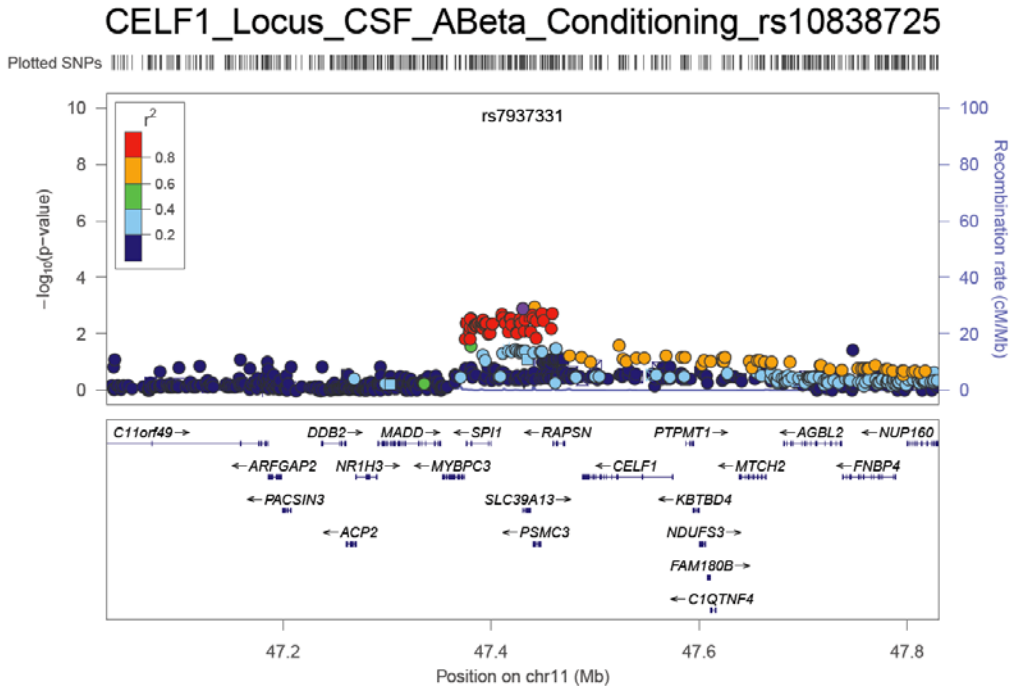

APOE_Locus_CSF_Ptau_Conditioning_e2_e4_SNPs
Plotted SNPs
rs769449
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
CEACAM20
PVR
BCL3
PVRL2
APOC4
RELB
PPP1R37
EXOC3L2
CEACAM22P
CEACAM19
CBLC
TOMM40
CLASRP
NKPD1
MARK4
IGSF23
BCAM
APOE
ZNF296
TRAPPC6A
CKM
MIR4531
APOC1
GEMIN7
BLOC1S3
CEACAM16
APOC1P1
APOC4−APOC2
APOC2
CLPTM1
45.2
45.4
45.6
45.8
Position on chr19 (Mb)

Figure S4. Regional plots for the *CELF1* locus with CSF $A\beta_{42}$ levels before (A) and after conditioning on (B) rs10838725 and (C) rs7937331. (D) IGAP regional plot for the *EPHA1* (rs10838725).

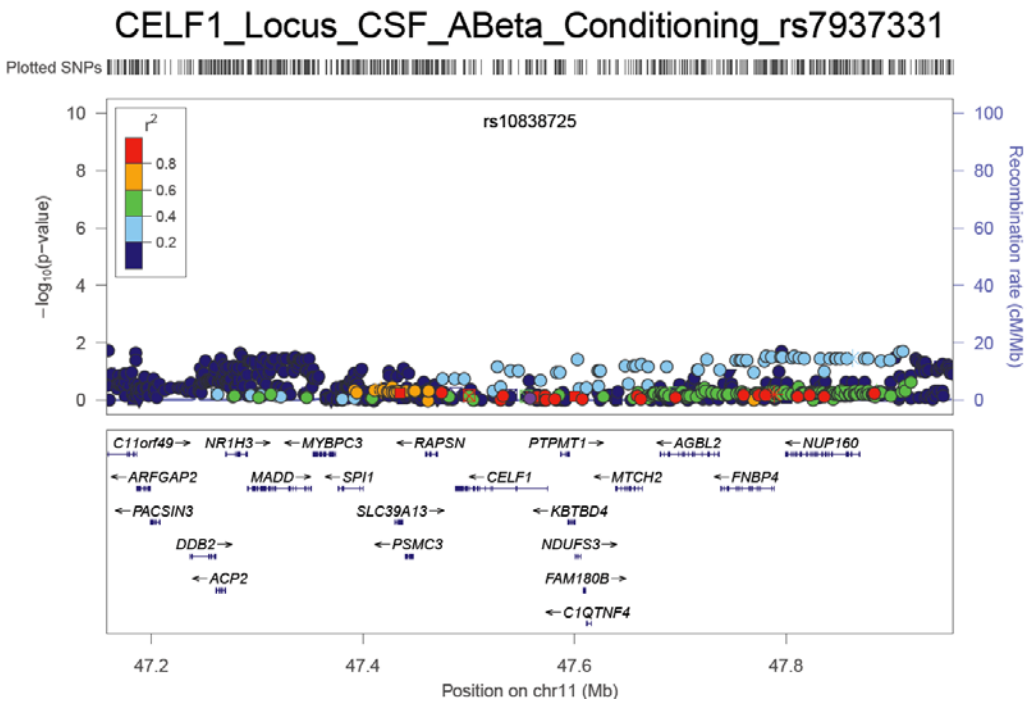
(A)



(B)



(C)


CELF1_Locus_CSF_ABeta_Conditioning_rs7937331
Plotted SNPs
rs10838725
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
C11orf49→ NR1H3→ ←MYBPC3 ←RAPSN PTPMT1→ ←AGBL2 ←NUP160
←ARFGAP2 MADD→ ←SPI1 ←CELF1 ←MTCH2 ←FNBP4
←PACSIN3 SLC39A13→ ←KBTBD4
DDB2→ ←PSMC3 NDUFS3→
←ACP2 FAM180B→
←C1QTNF4
47.2
47.4
47.6
47.8
Position on chr11 (Mb)

(D)

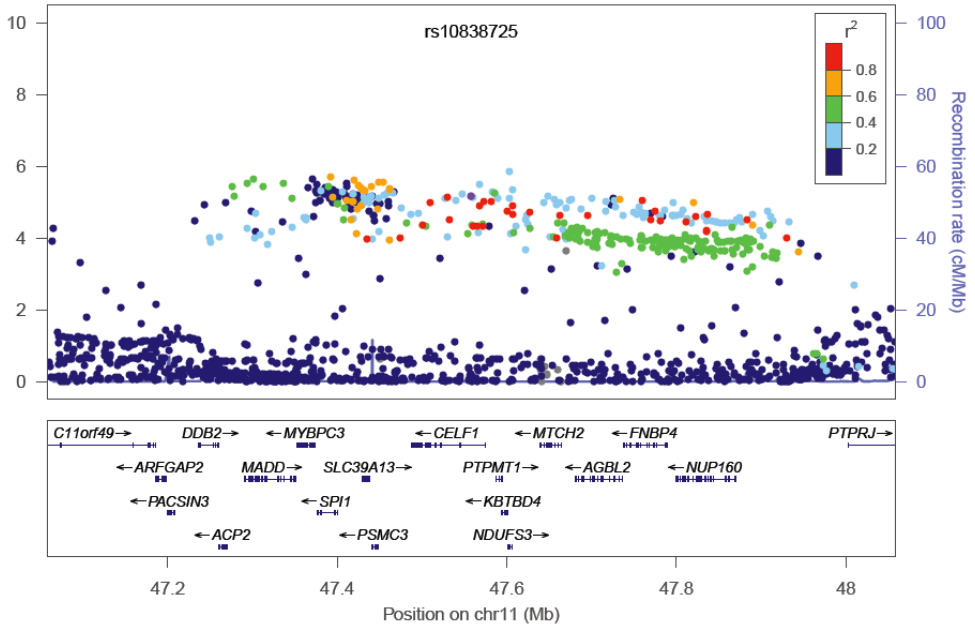
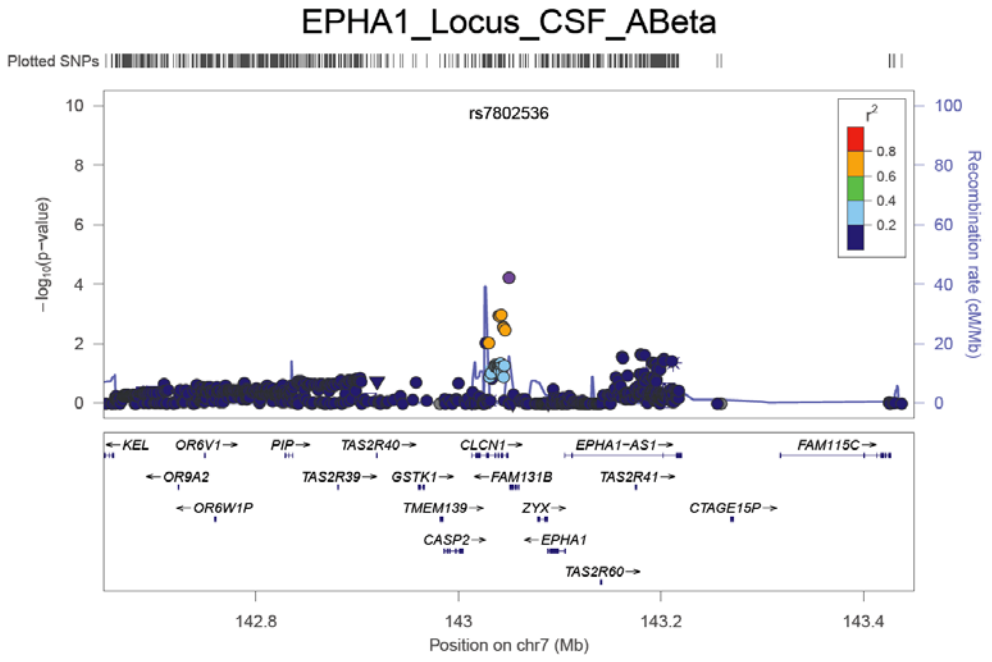
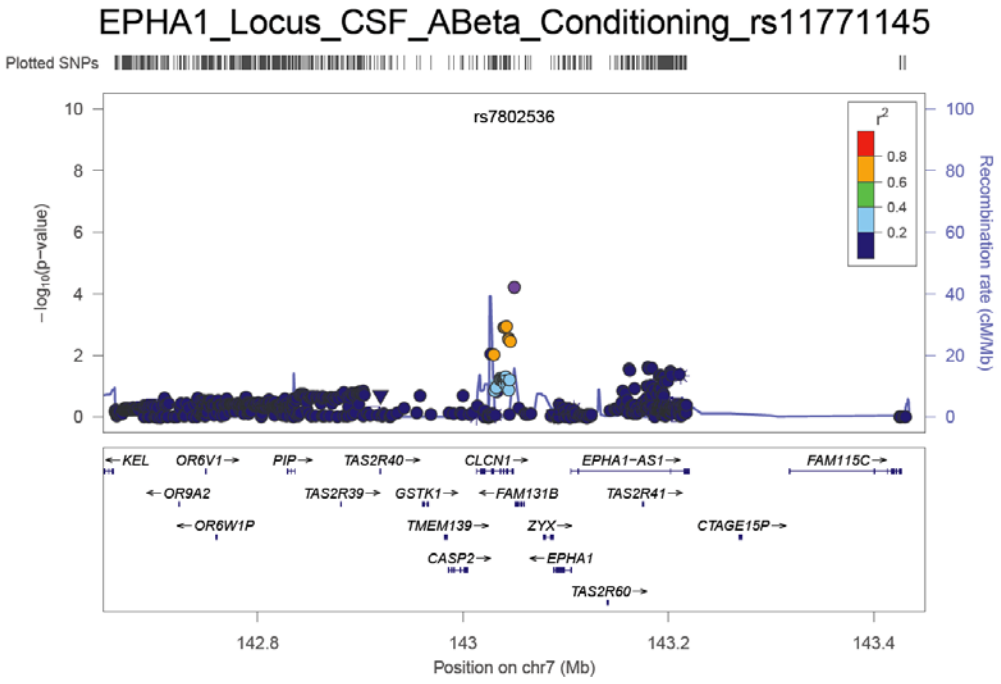

rs10838725
r2
0.8
0.6
0.4
0.2
Recombination rate (cM/Mb)
C11orf49→ DDB2→ ←MYBPC3 ←CELF1 ←MTCH2 ←FNBP4 PTPRJ→
←ARFGAP2 MADD→ SLC39A13→ PTPMT1→ ←AGBL2 ←NUP160
←PACSIN3 ←SPI1 ←KBTBD4
←ACP2 ←PSMC3 NDUFS3→
47.2
47.4
47.6
47.8
48
Position on chr11 (Mb)

Figure S5. Regional plots for the *EPHA1* locus with CSF $A\beta_{42}$ levels before (A) and after conditioning on (B) rs11771145 and (C) rs7802536. (D) IGAP regional plot for the *EPHA1* (rs11771145).

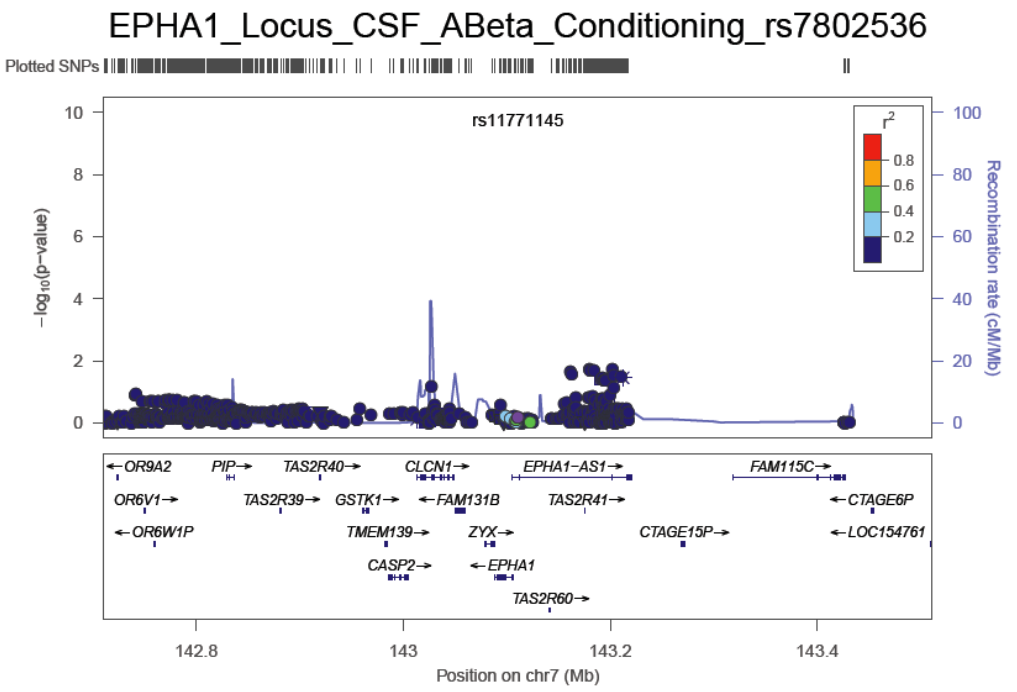
(A)



(B)



(C)


EPHA1_Locus_CSF_ABeta_Conditioning_rs7802536
Plotted SNPs
rs11771145
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
OR9A2→
PIP→
TAS2R40→
CLCN1→
EPHA1−AS1→
FAM115C→
OR6V1→
TAS2R39→
GSTK1→
←FAM131B
TAS2R41→
←CTAGE6P
←OR6W1P
TMEM139→
ZYX→
CTAGE15P→
←LOC154761
CASP2→
←EPHA1
TAS2R60→
142.8
143
143.2
143.4
Position on chr7 (Mb)

(D)

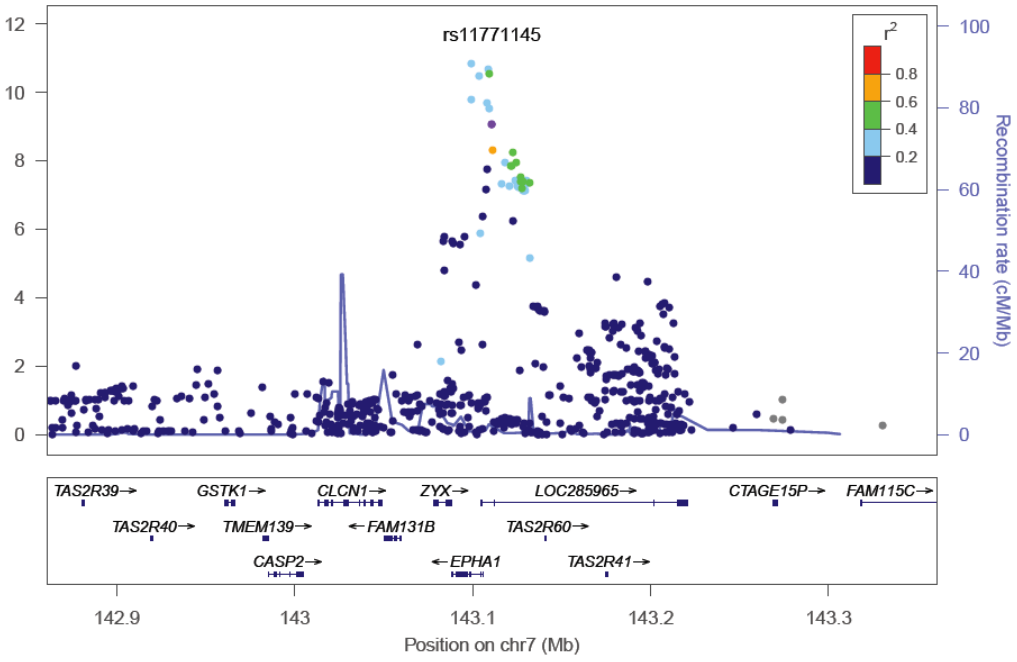
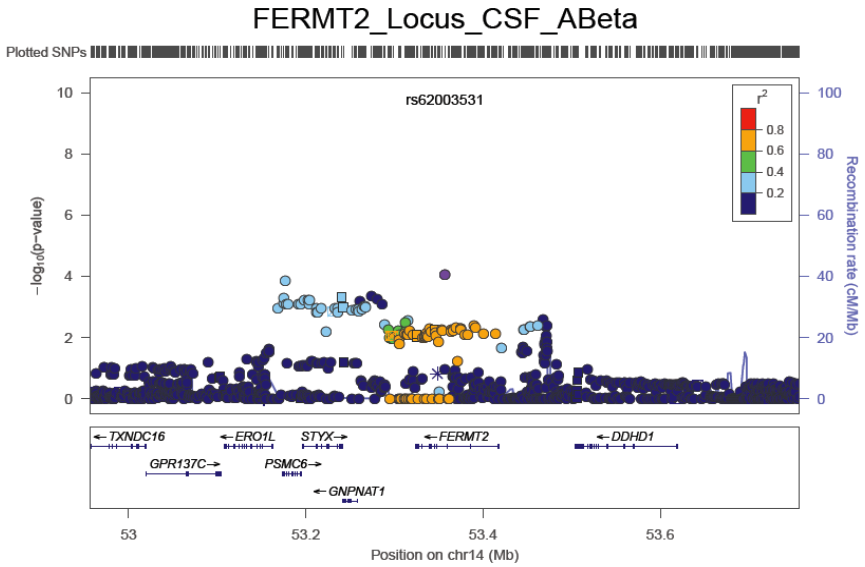
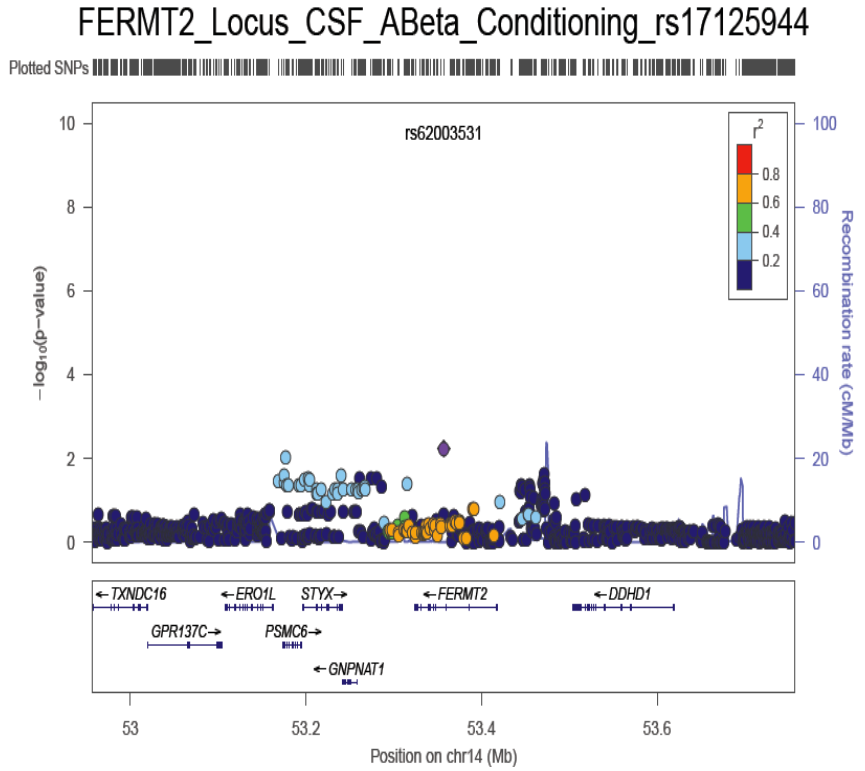

rs11771145
r2
0.8
0.6
0.4
0.2
Recombination rate (cM/Mb)
TAS2R39→
GSTK1→
CLCN1→
ZYX→
LOC285965→
CTAGE15P→
FAM115C→
TAS2R40→
TMEM139→
←FAM131B
TAS2R60→
CASP2→
←EPHA1
TAS2R41→
142.9
143
143.1
143.2
143.3
Position on chr7 (Mb)

Figure S6. Regional plots for the *FERMT2* locus with CSF $A\beta_{42}$ levels before (A) and after conditioning on (B) rs17125944 and (C) rs62003531. (D) IGAP regional plot for the *FERMT2* (rs17125944).

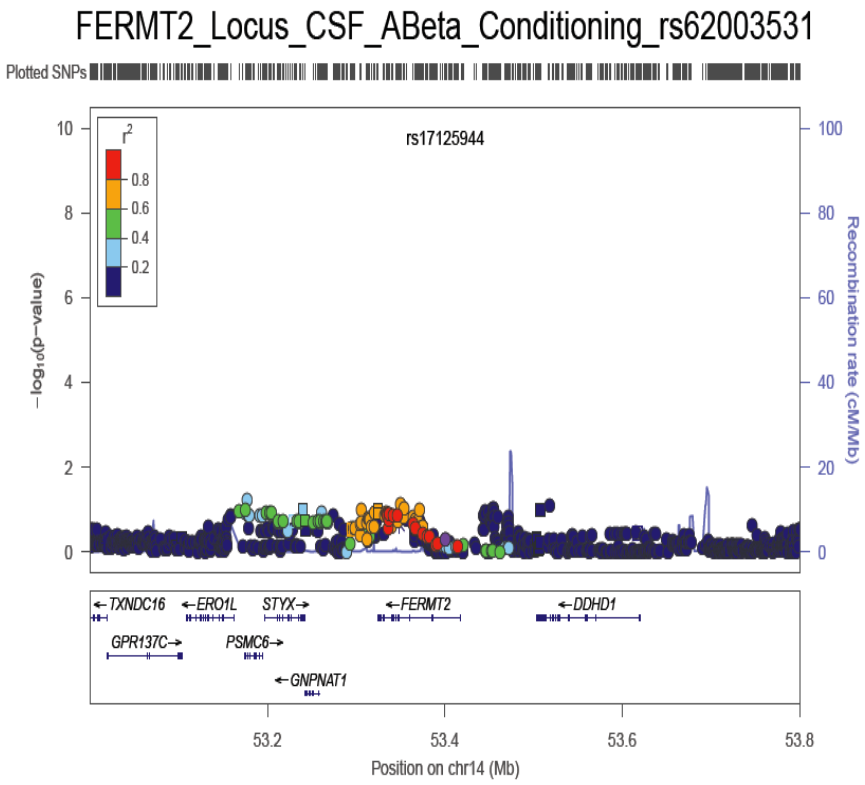
(A)



(B)



(C)


FERMT2_Locus_CSF_ABeta_Conditioning_rs62003531
Plotted SNPs
rs17125944
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
TXNDC16
ERO1L
STYX
FERMT2
DDHD1
GPR137C
PSMC6
GNPNAT1
Position on chr14 (Mb)

(D)

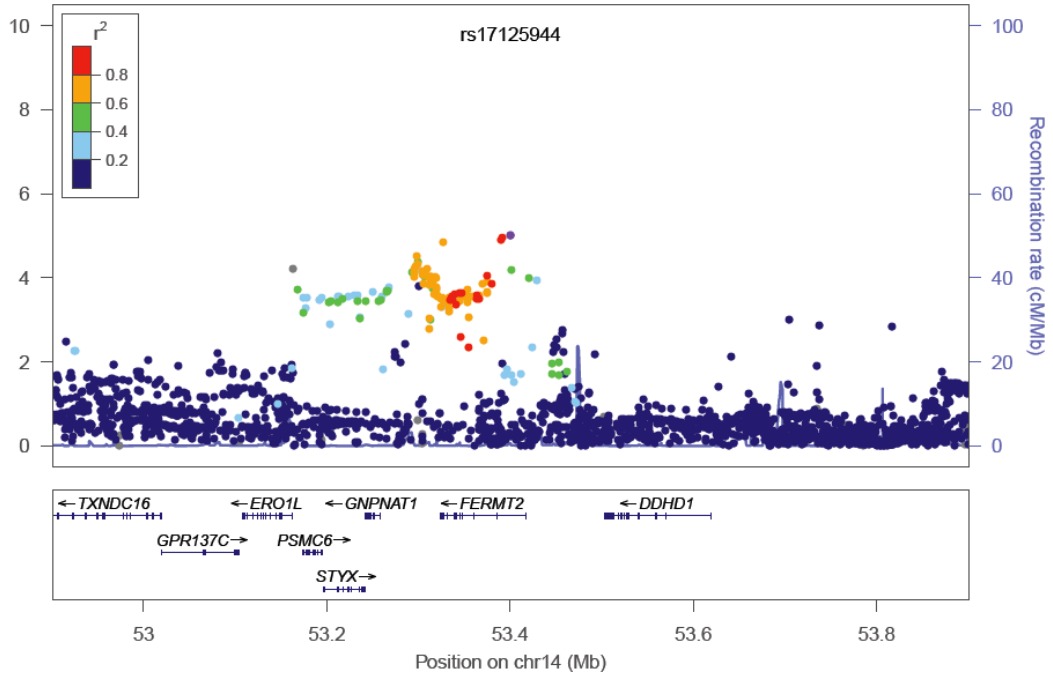
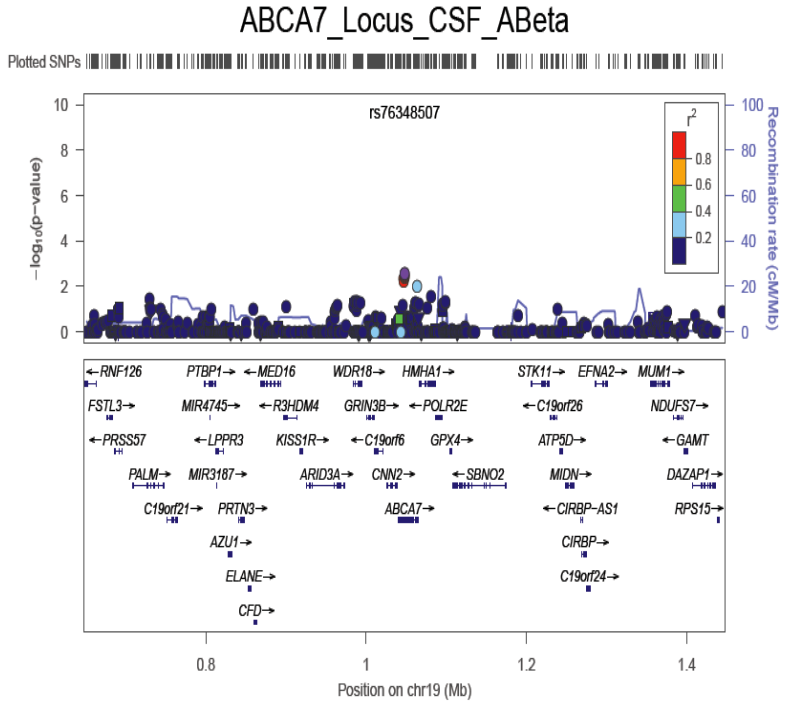
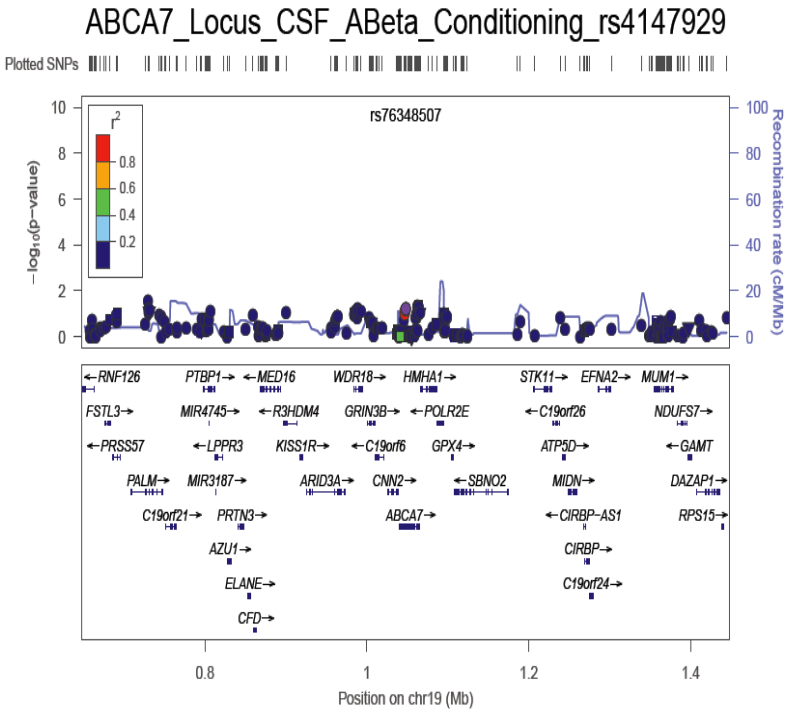

rs17125944
r2
0.8
0.6
0.4
0.2
Recombination rate (cM/Mb)
TXNDC16
ERO1L
GNPNAT1
FERMT2
DDHD1
GPR137C
PSMC6
STYX
Position on chr14 (Mb)

Figure S7. Regional plots for the *ABCA7* locus with CSF $A\beta_{42}$ levels before (A) and after conditioning on (B) rs4147929 and (C) rs76348507. (D) IGAP regional plot for the *ABCA7* (rs4147929).

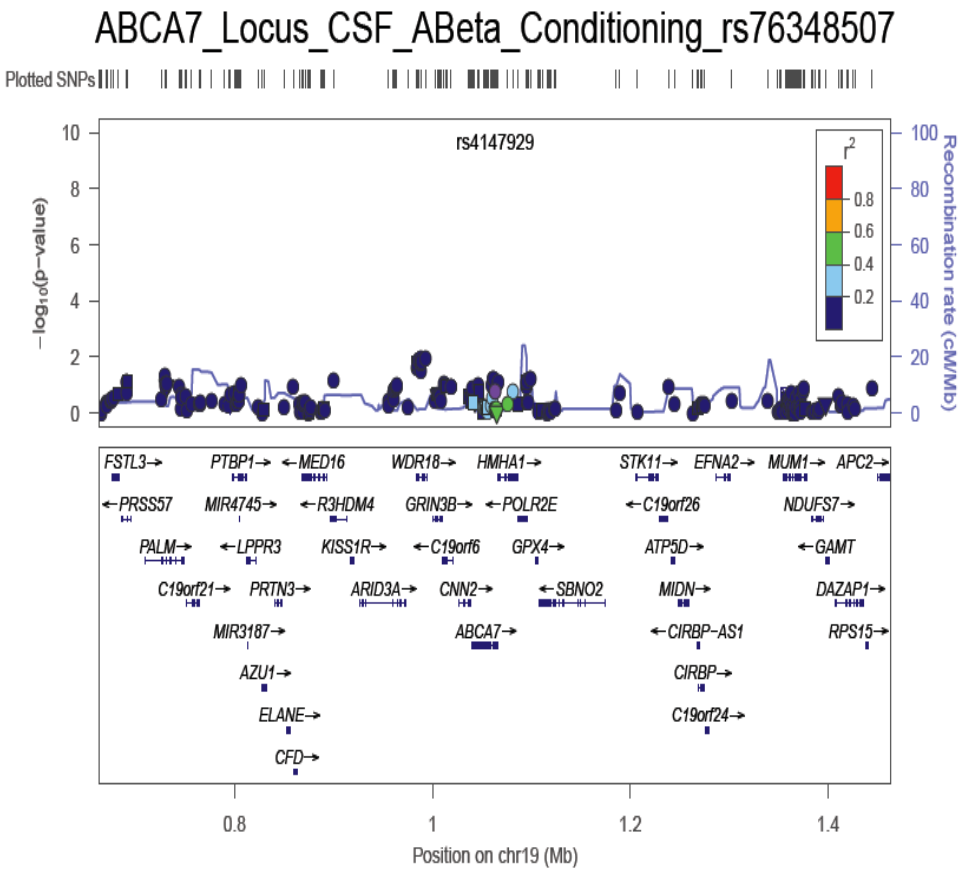
(A)



(B)



(C)


ABCA7_Locus_CSF_ABeta_Conditioning_rs76348507
Plotted SNPs
rs4147929
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
FSTL3→ PTBP1→ ←MED16 WDR18→ HMHA1→ STK11→ EFNA2→ MUM1→ APC2→
←PRSS57 MIR4745→ ←R3HDM4 GRIN3B→ ←POLR2E ←C19orf26 NDUFS7→
PALM→ ←LPPR3 KISS1R→ ←C19orf6 GPX4→ ATP5D→ ←GAMT
C19orf21→ PRTN3→ ARID3A→ CNN2→ ←SBNO2 MIDN→ DAZAP1→
MIR3187→ ABCA7→ ←CIRBP-AS1 RPS15→
AZU1→ CIRBP→
ELANE→ C19orf24→
CFD→
Position on chr19 (Mb)

(D)

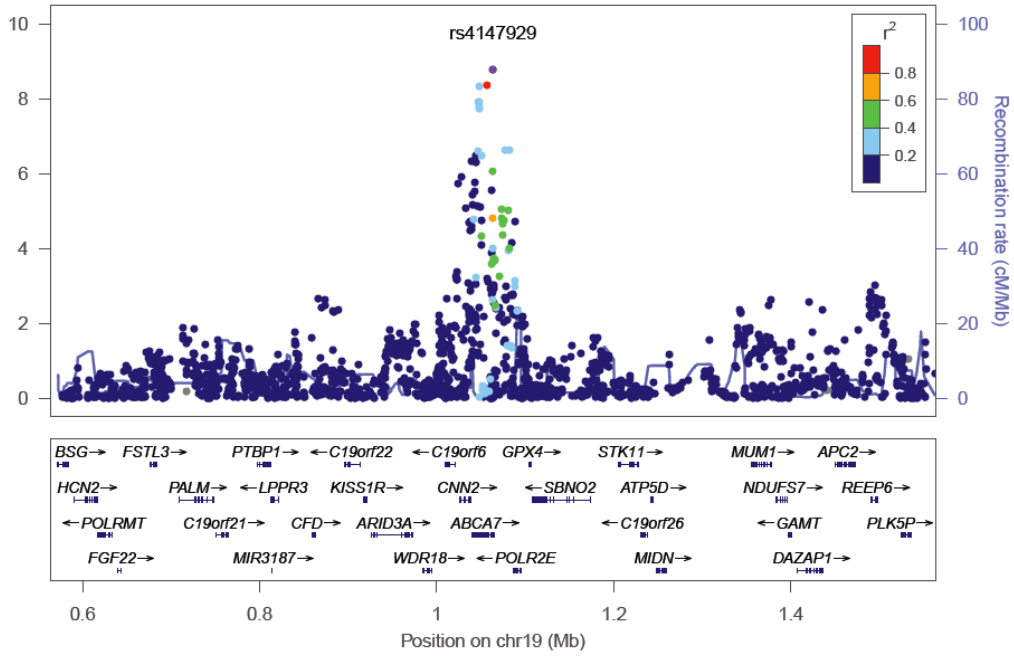
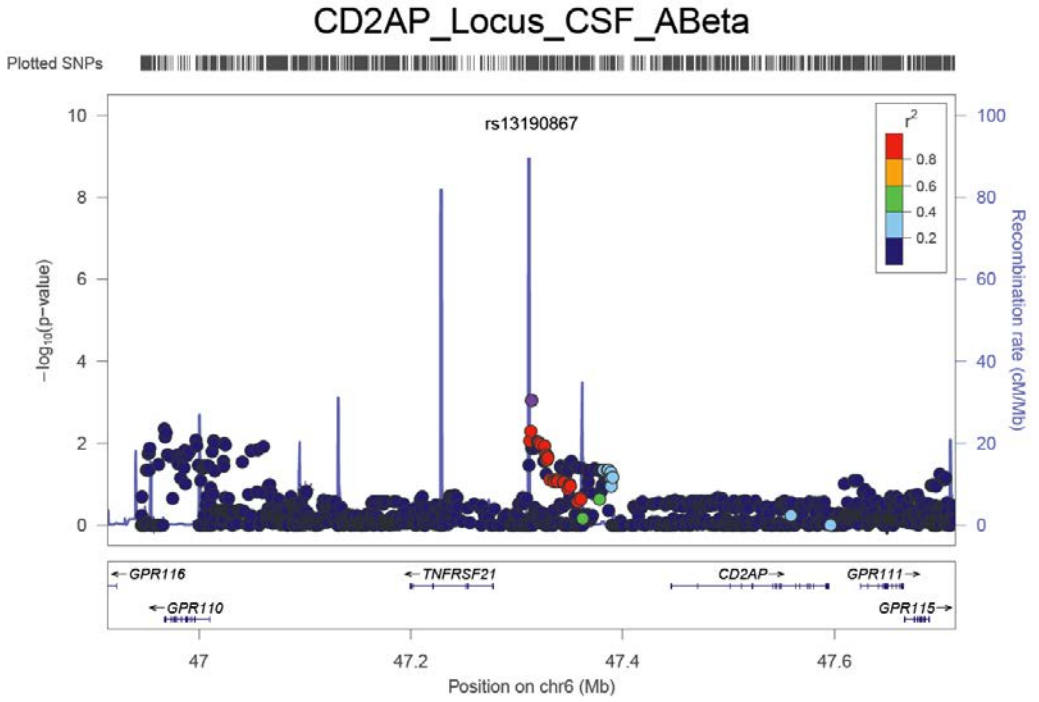
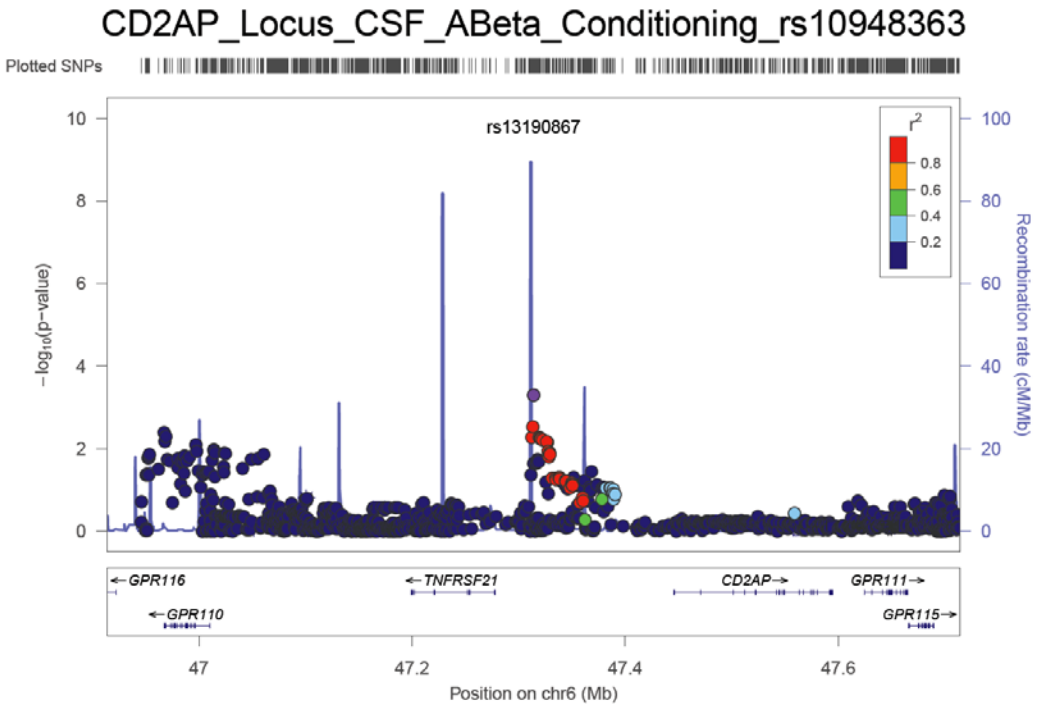

rs4147929
r²
0.8
0.6
0.4
0.2
Recombination rate (cM/Mb)
BSG→ FSTL3→ PTBP1→ ←C19orf22 ←C19orf6 GPX4→ STK11→ MUM1→ APC2→
HCN2→ PALM→ ←LPPR3 KISS1R→ CNN2→ ←SBNO2 ATP5D→ NDUFS7→ REEP6→
←POLRMT C19orf21→ CFD→ ARID3A→ ABCA7→ ←C19orf26 ←GAMT PLK5P→
FGF22→ MIR3187→ WDR18→ ←POLR2E MIDN→ DAZAP1→
Position on chr19 (Mb)

Figure S8. Regional plots for the *CD2AP* locus with CSF $A\beta_{42}$ levels before (A) and after conditioning on (B) rs10948363 and (C) rs13190867. (D) IGAP regional plot for the *CD2AP* (rs10948363).

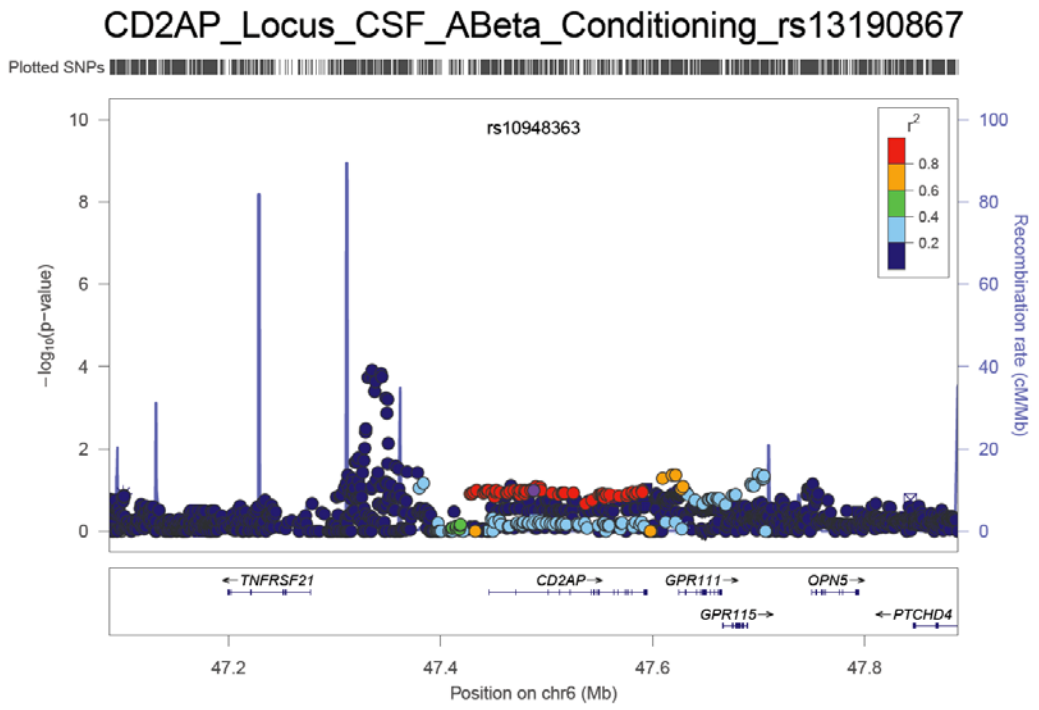
(A)



(B)



(C)


CD2AP_Locus_CSF_ABeta_Conditioning_rs13190867
Plotted SNPs
rs10948363
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
TNFRSF21
CD2AP
GPR111
OPN5
GPR115
PTCHD4
Position on chr6 (Mb)

(D)

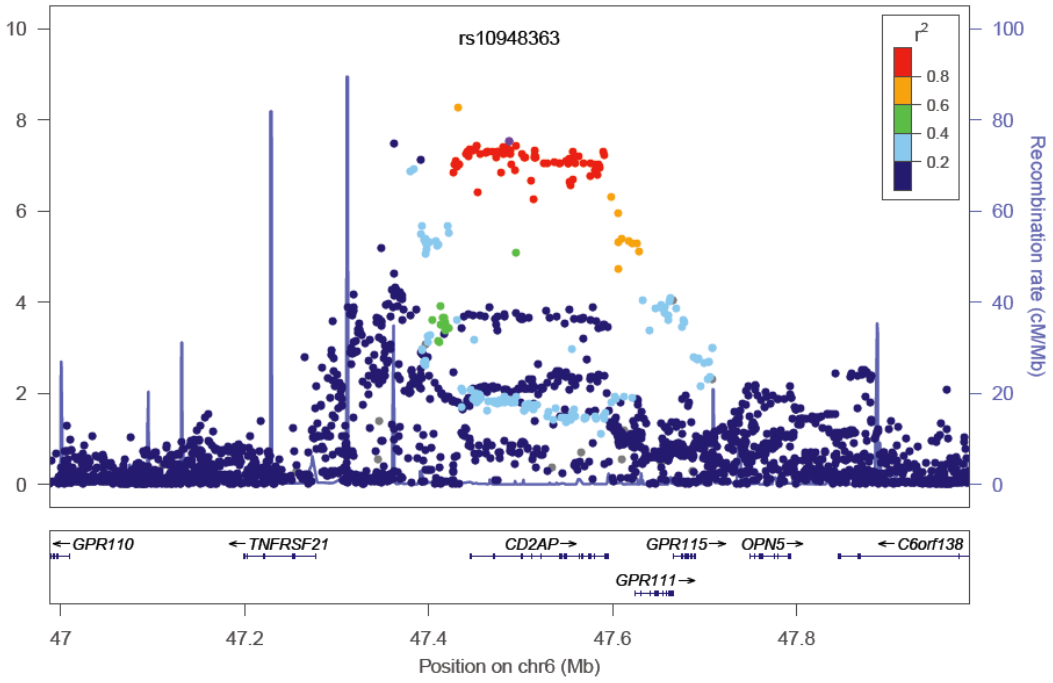

rs10948363
r2
0.8
0.6
0.4
0.2
Recombination rate (cM/Mb)
GPR110
TNFRSF21
CD2AP
GPR115
OPN5
C6orf138
GPR111
Position on chr6 (Mb)

Figure S9. Regional plot for the *BIN1* locus with CSF $A\beta_{42}$ levels.

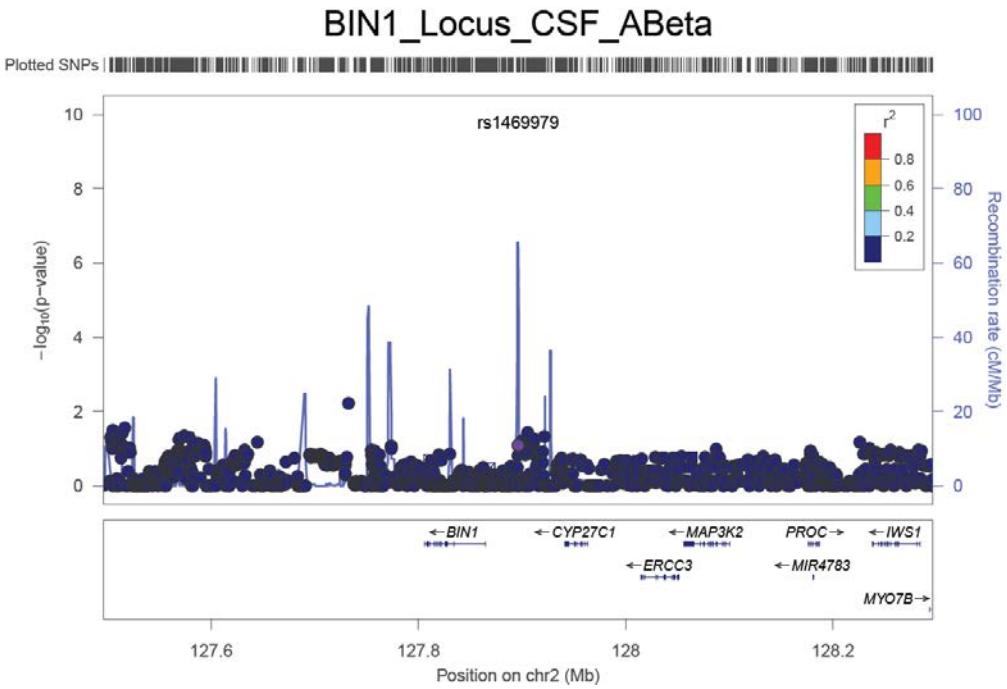


Figure S10. Regional plot for the *CASS4* locus with CSF $A\beta_{42}$ levels.

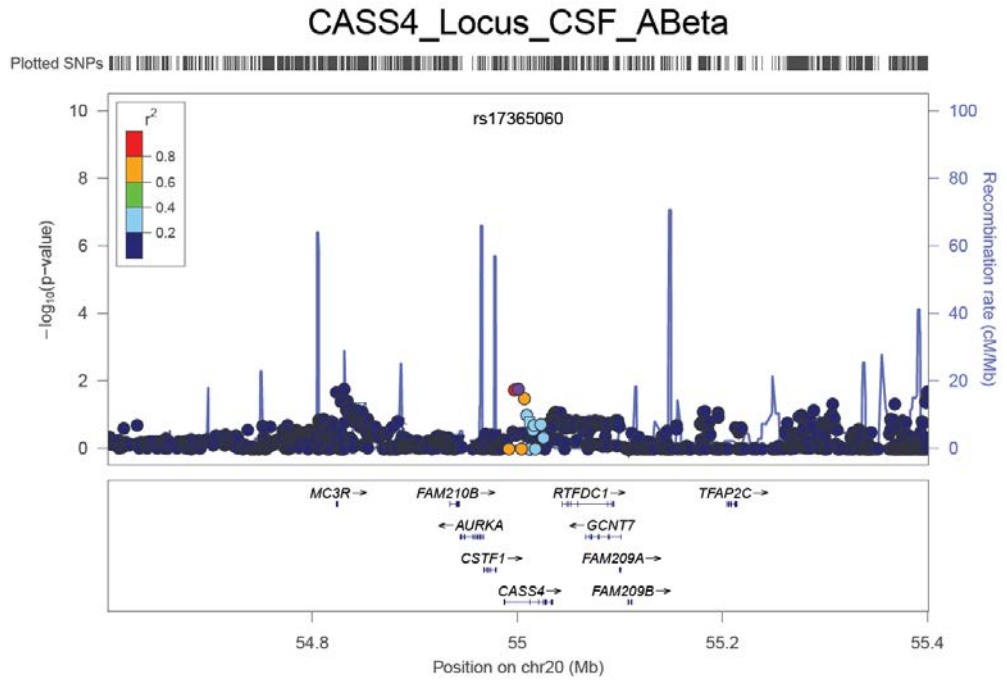


Figure S11. Regional plot for the *CD33* locus with CSF $A\beta_{42}$ levels.

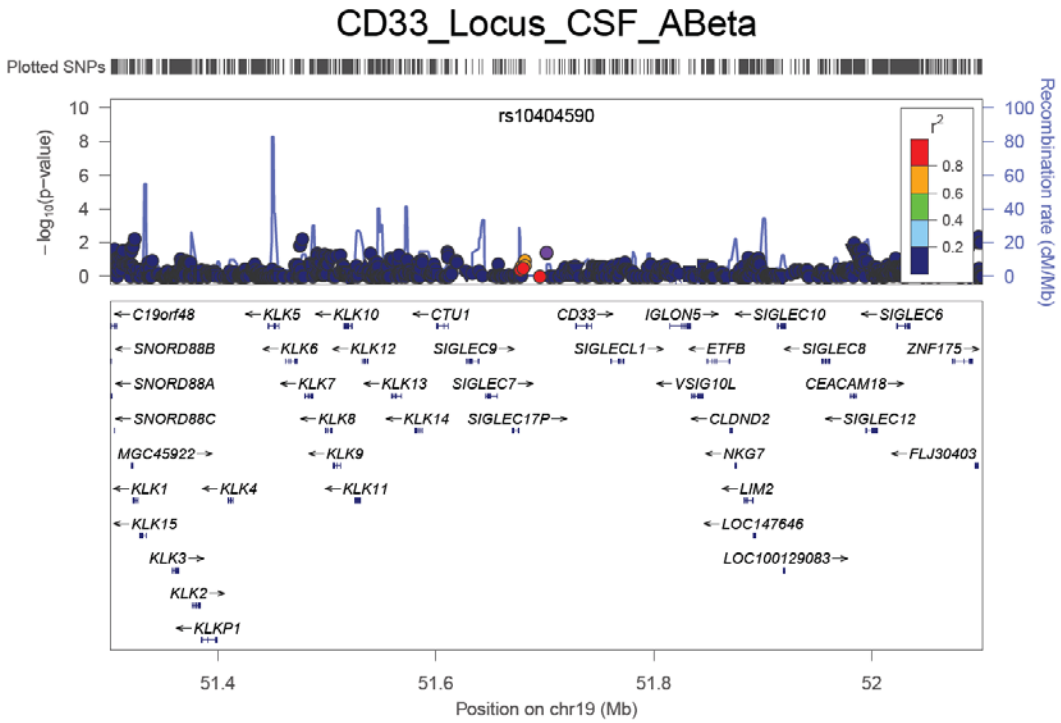


Figure S12. Regional plot for the *CLU* locus with CSF $A\beta_{42}$ levels.

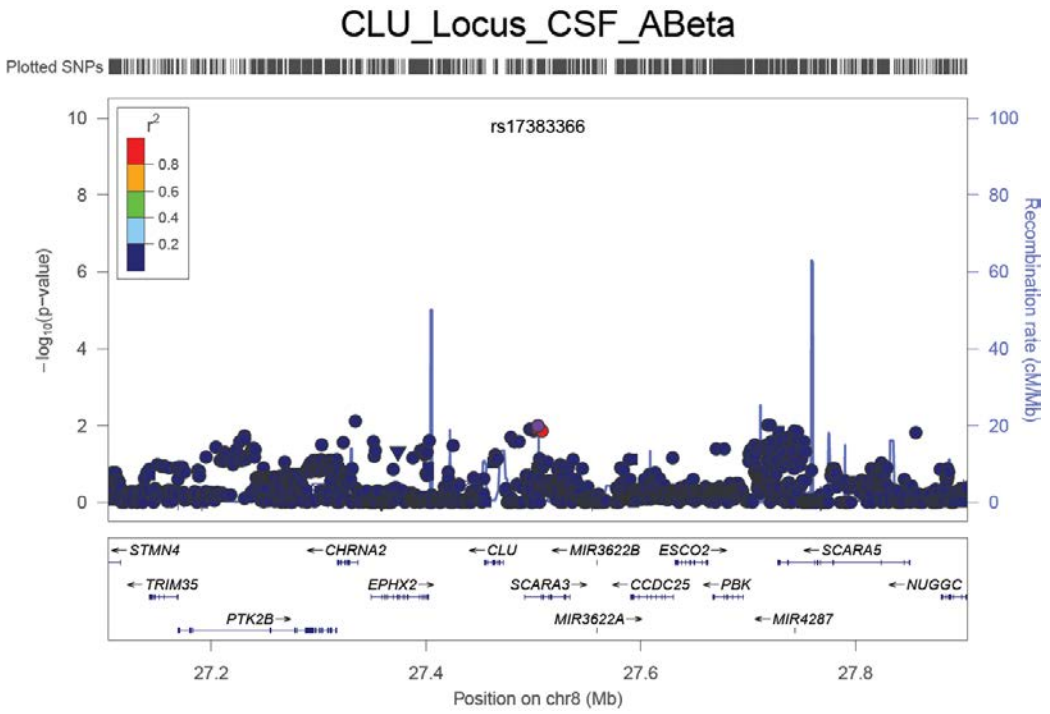


Figure S13. Regional plot for the *CR1* locus with CSF $A\beta_{42}$ levels.

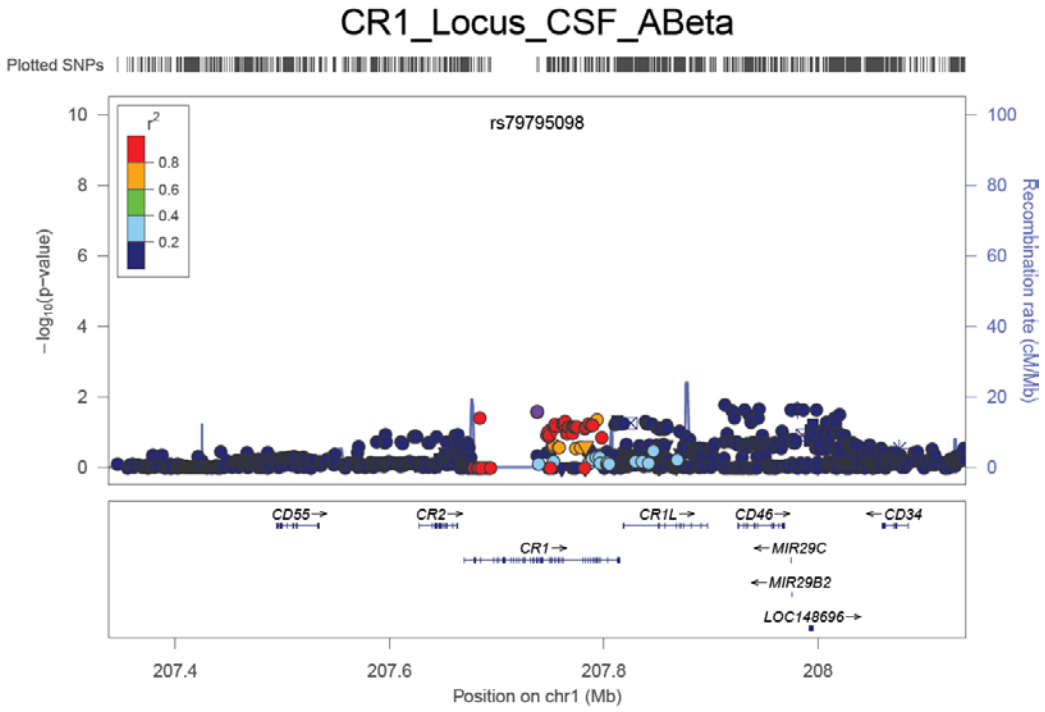


Figure S14. Regional plot for the *INPP5D* locus with CSF $A\beta_{42}$ levels.

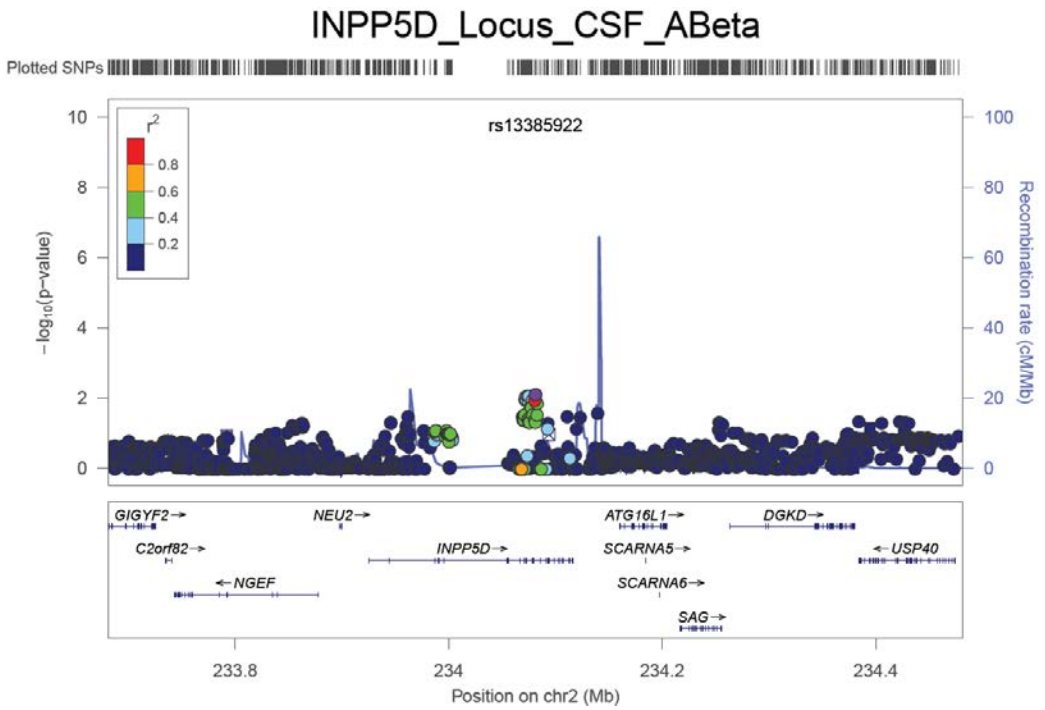


Figure S15. Regional plot for the *HLA-DRB5/DRB1* locus with CSF $A\beta_{42}$ levels.

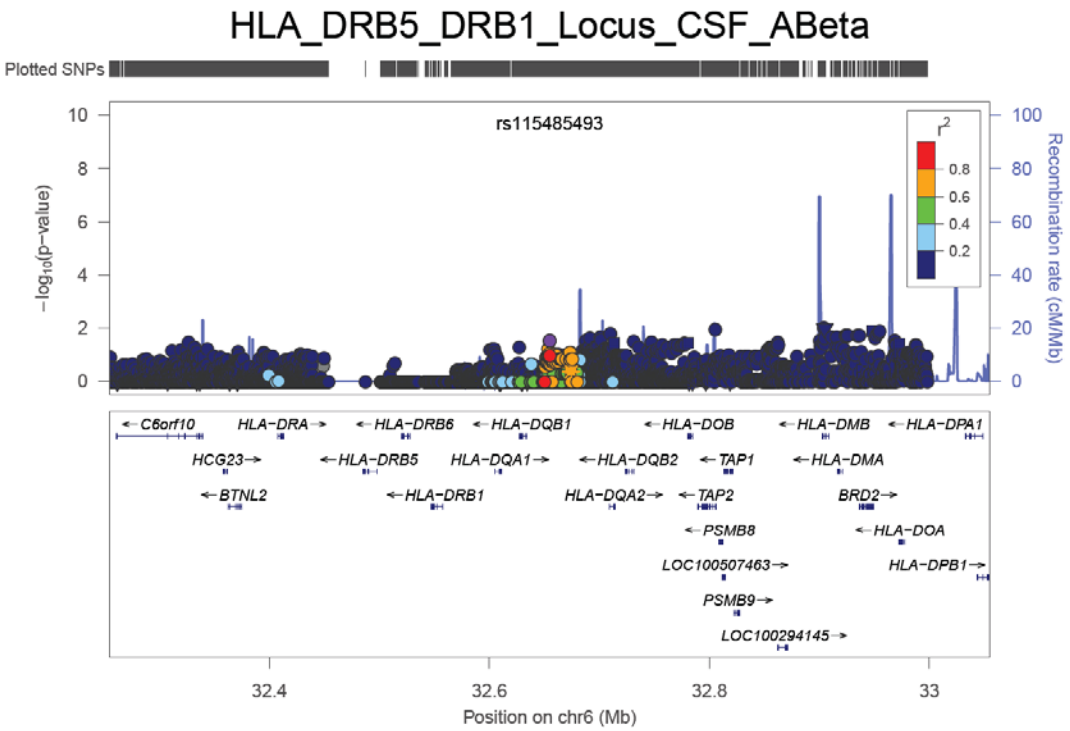


Figure S16. Regional plot for the *INPP5D* locus with CSF $A\beta_{42}$ levels.

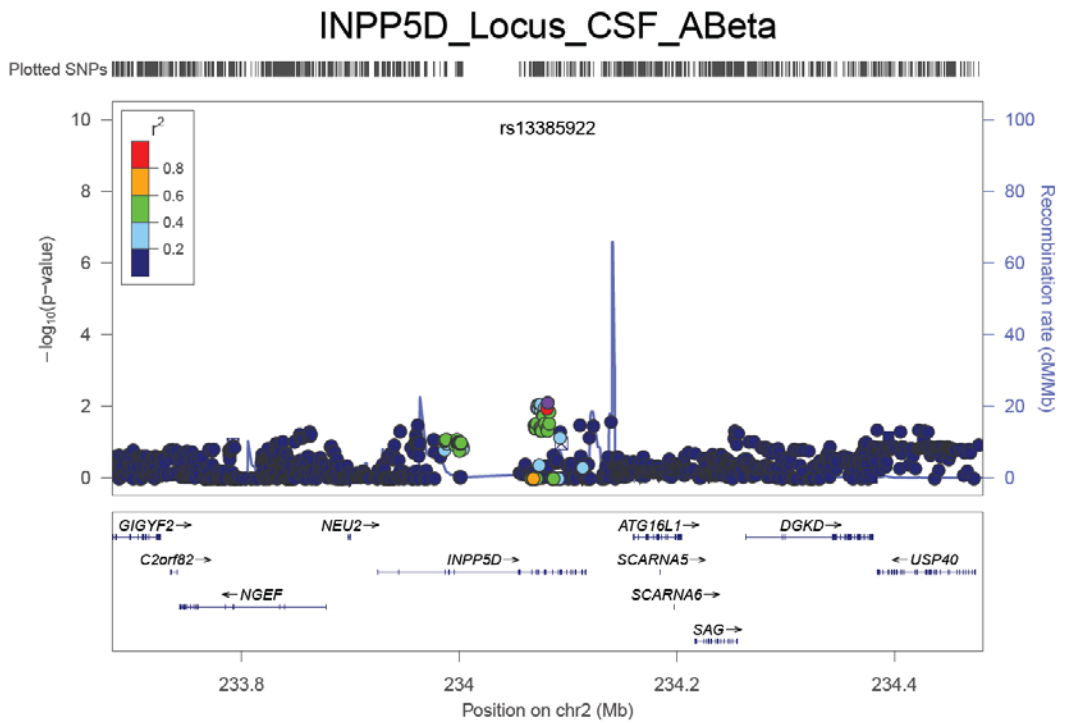


Figure S17. Regional plot for the *MEF2C* locus with CSF $A\beta_{42}$ levels.

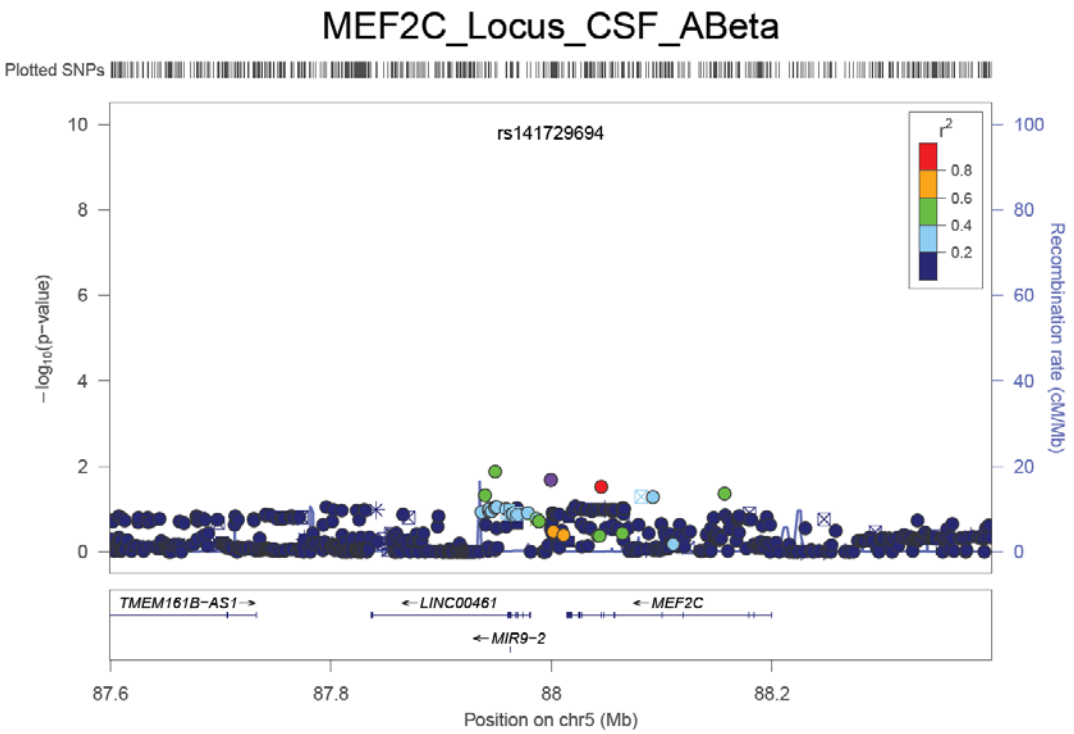


Figure S18. Regional plot for the *MS4A6A* locus with CSF $A\beta_{42}$ levels.

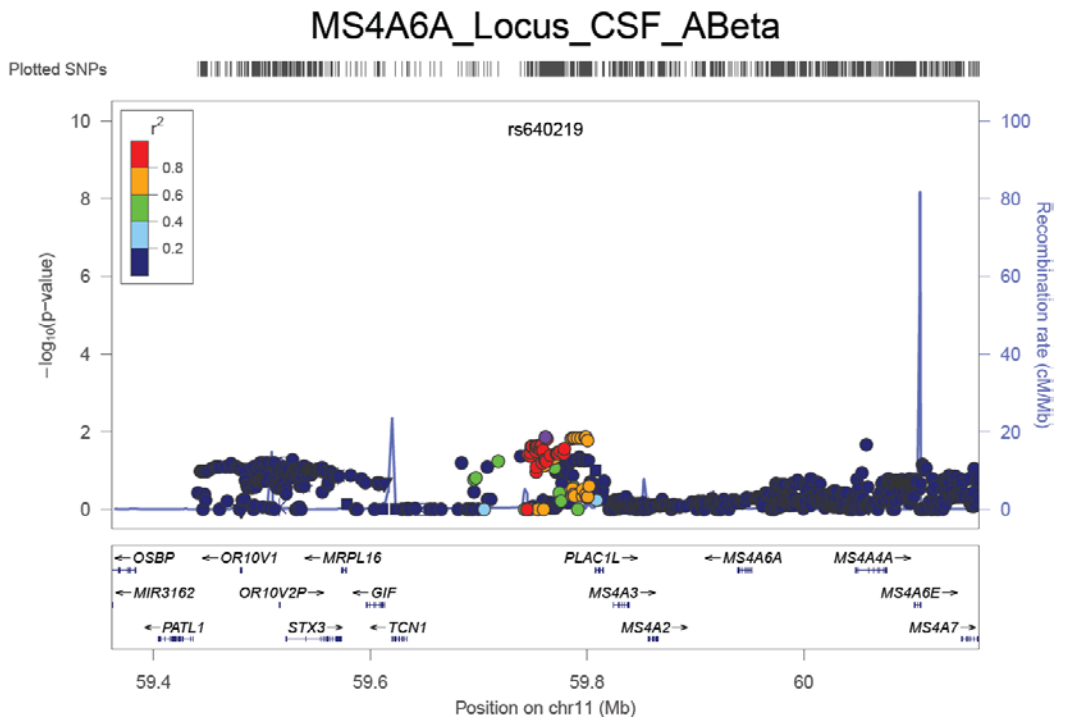


Figure S19. Regional plot for the *NME8* locus with CSF $A\beta_{42}$ levels.

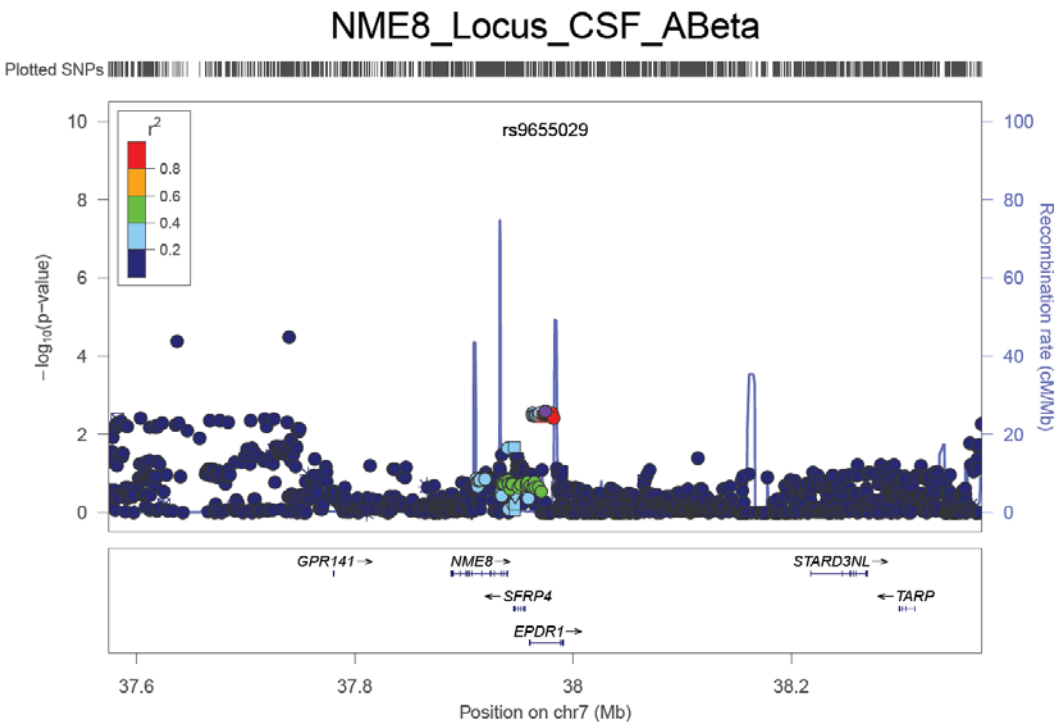


Figure S20. Regional plot for the *PICALM* locus with CSF $A\beta_{42}$ levels.

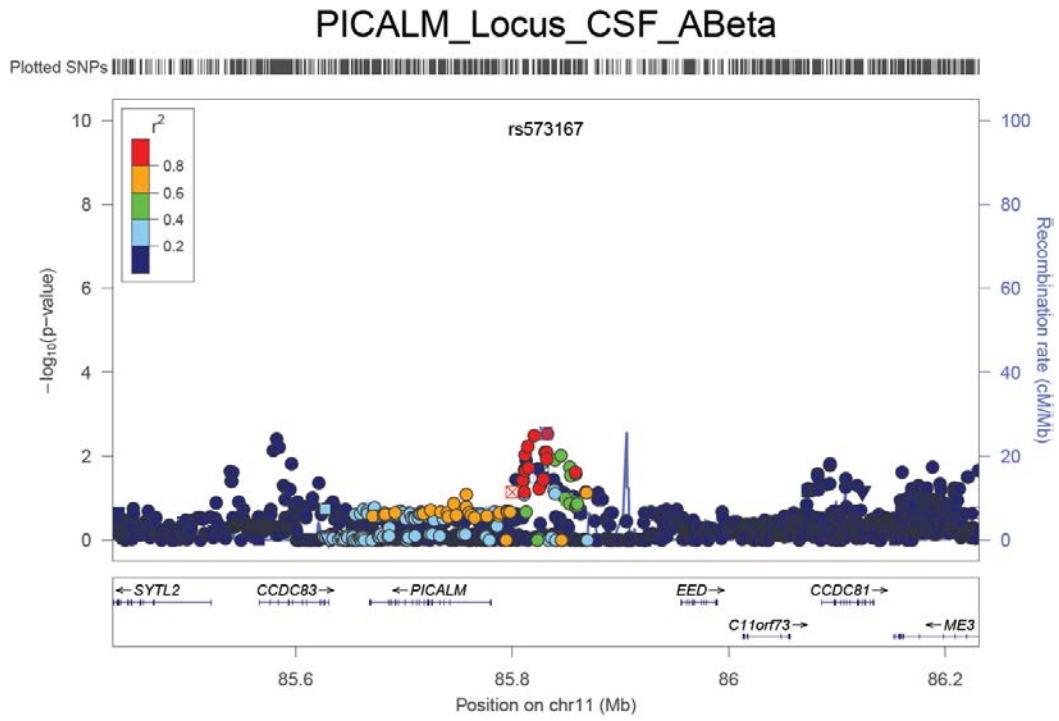


Figure S21. Regional plot for the *PTK2B* locus with CSF $A\beta_{42}$ levels.

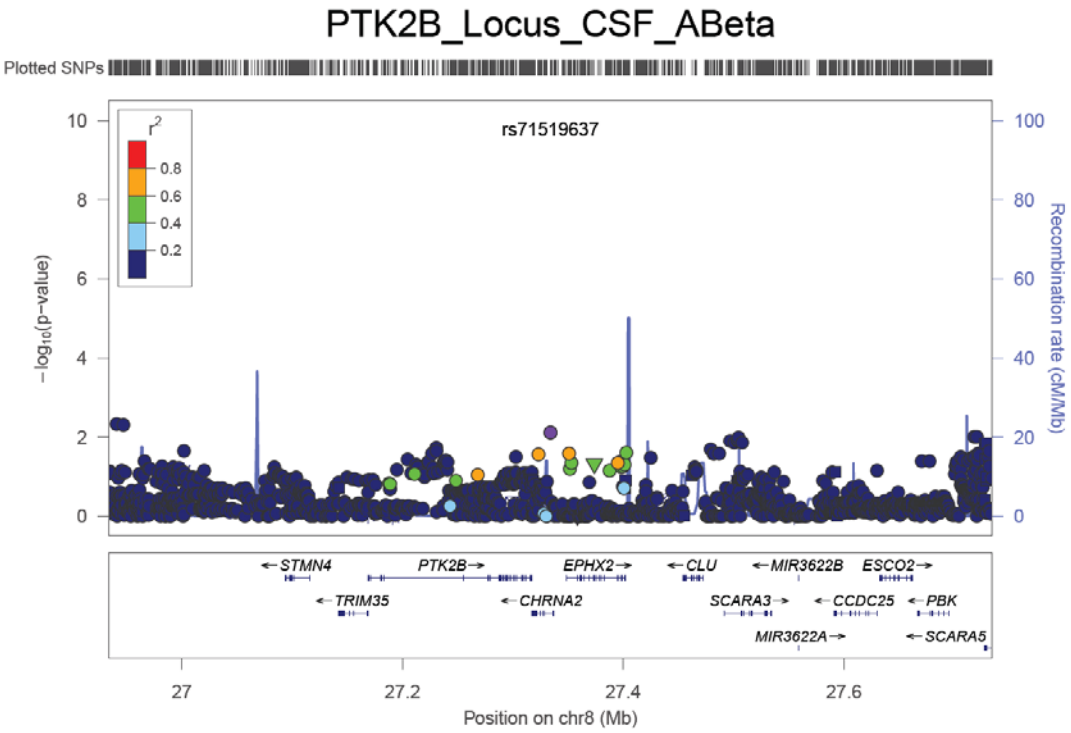


Figure S22. Regional plot for the *SLC24A4* locus with CSF $A\beta_{42}$ levels.

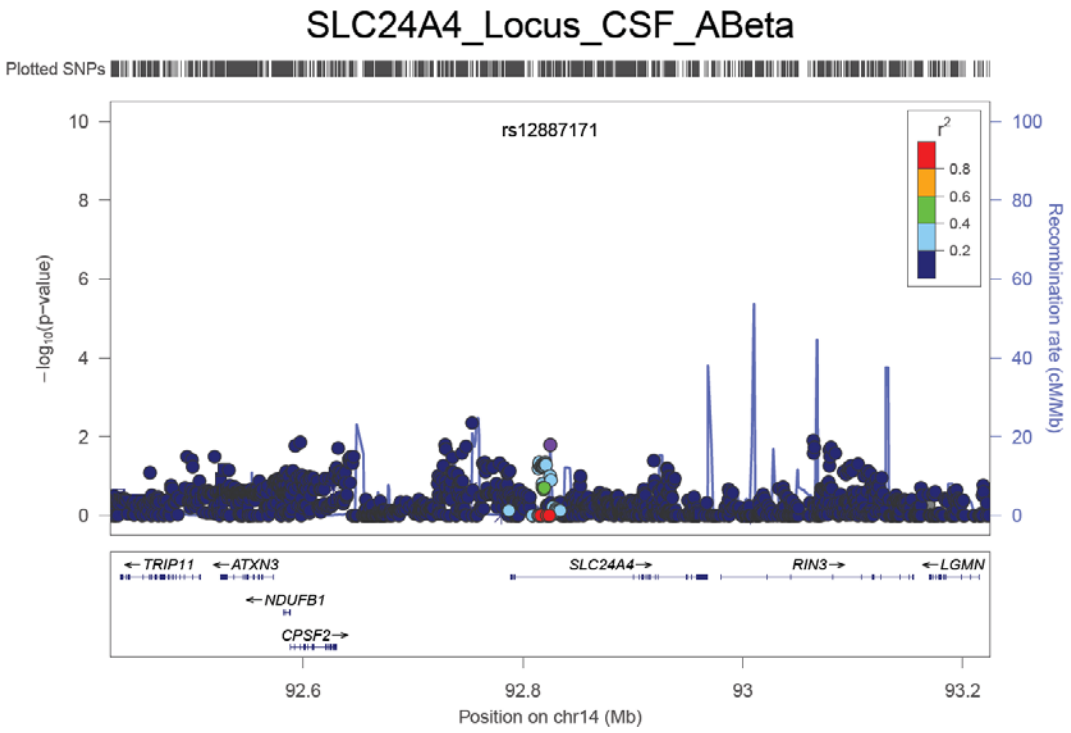


Figure S23. Regional plot for the *SORL1* locus with CSF $A\beta_{42}$ levels.

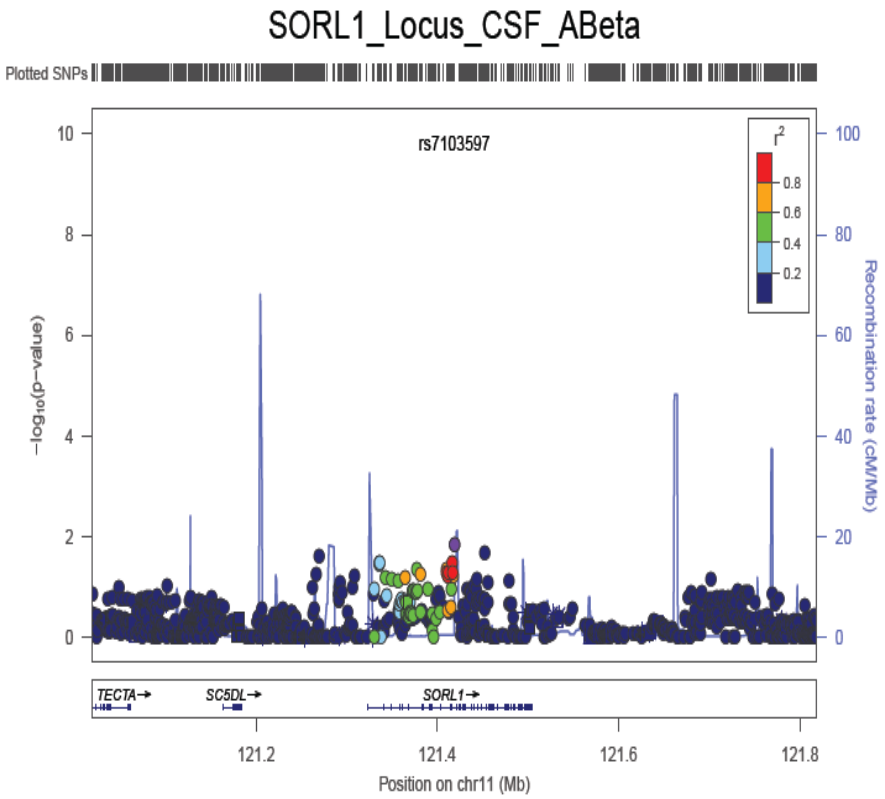


Figure S24. Regional plot for the *TP53INP1* locus with CSF $A\beta_{42}$ levels.

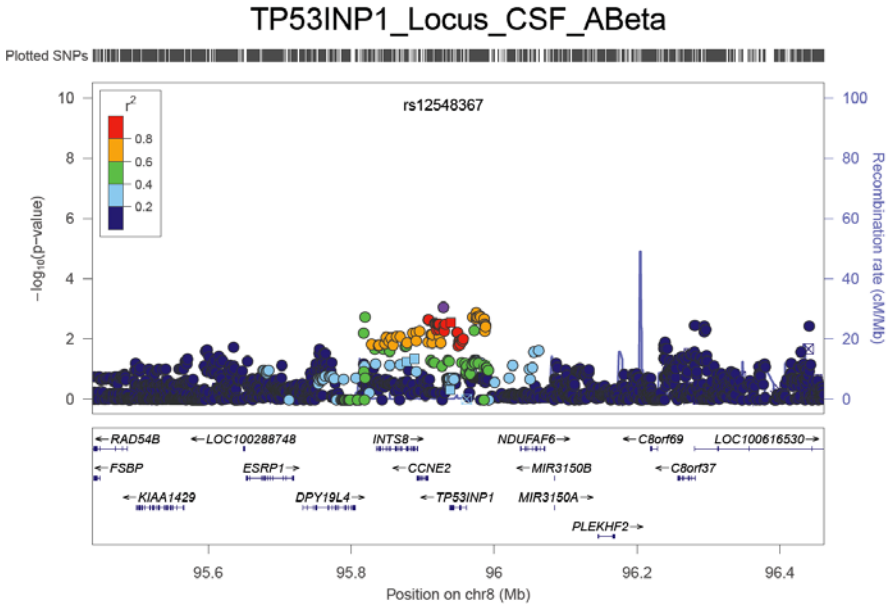


Figure S25. Regional plot for the *ZCWPW1* locus with CSF $A\beta_{42}$ levels.

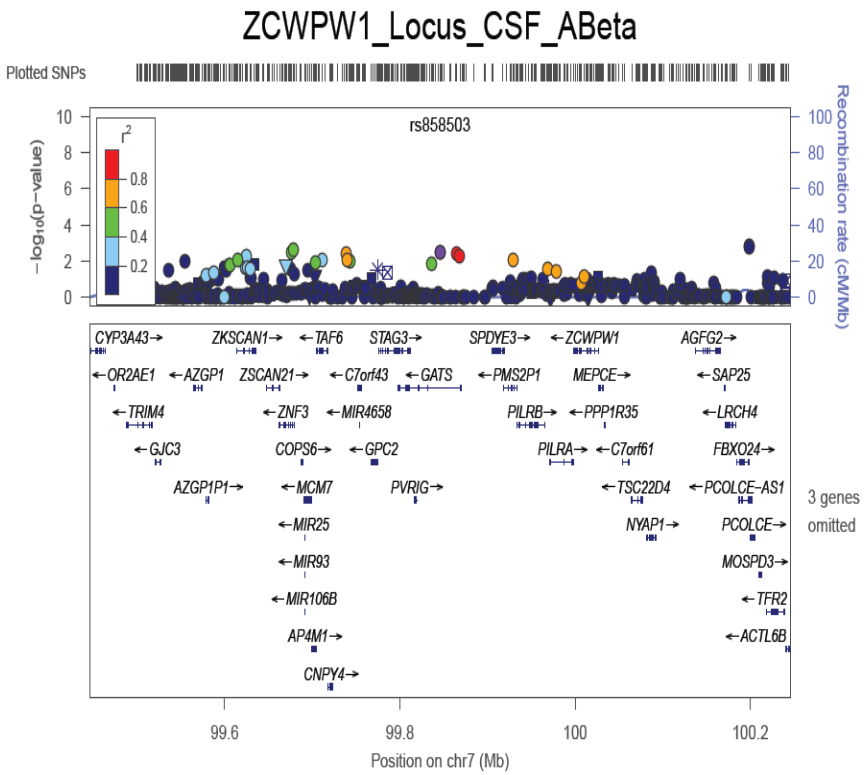
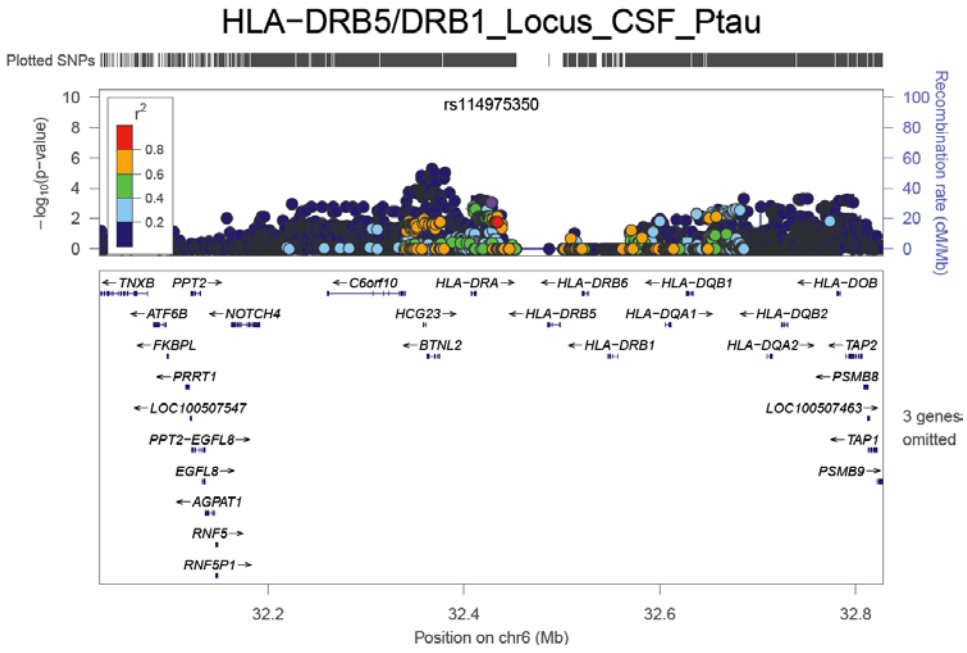
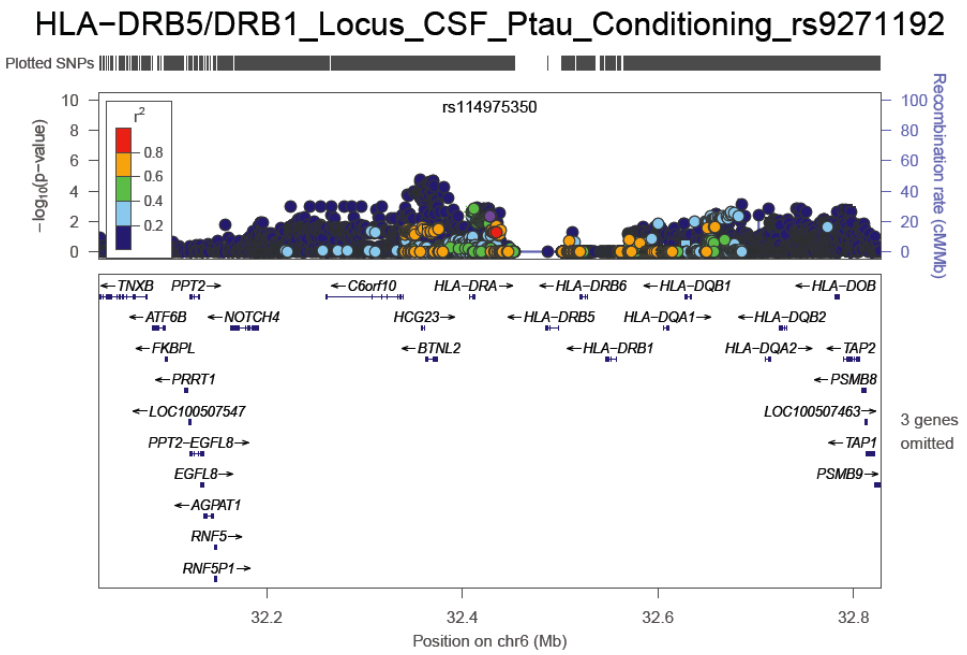


Figure S26. Regional plots for the *HLA-DRB5/DRB1* locus with CSF $ptau_{181}$ levels before (A) and after conditioning on (B) rs9271192 and (C) rs114975350. (D) IGAP regional plot for the *HLA-DRB5/DRB1* (rs9271192).

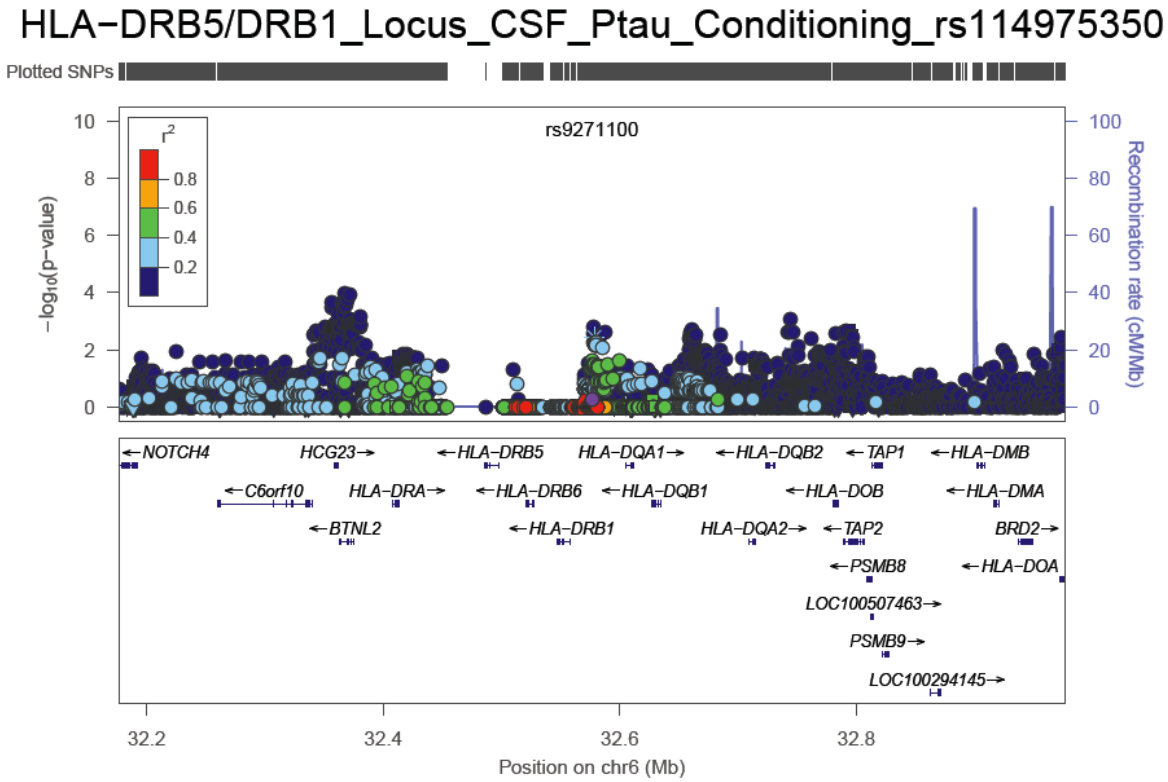
(A)



(B)


HLA−DRB5/DRB1_Locus_CSF_Ptau_Conditioning_rs9271192
Plotted SNPs
rs114975350
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
TNXB
PPT2
C6orf10
HLA−DRA
HLA−DRB6
HLA−DQB1
HLA−DOB
ATF6B
NOTCH4
HCG23
HLA−DRB5
HLA−DQA1
HLA−DQB2
FKBPL
BTNL2
HLA−DRB1
HLA−DQA2
TAP2
PRRT1
PSMB8
LOC100507547
LOC100507463
PPT2−EGFL8
TAP1
EGFL8
PSMB9
AGPAT1
RNF5
RNF5P1
3 genes omitted
32.2
32.4
32.6
32.8
Position on chr6 (Mb)

(C)


HLA−DRB5/DRB1_Locus_CSF_Ptau_Conditioning_rs9271100
Plotted SNPs
rs9271100
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
NOTCH4
HCG23
HLA−DRB5
HLA−DQA1
HLA−DQB2
TAP1
HLA−DMB
C6orf10
HLA−DRA
HLA−DRB6
HLA−DQB1
HLA−DOB
HLA−DMA
BTNL2
HLA−DRB1
HLA−DQA2
TAP2
BRD2
PSMB8
HLA−DOA
LOC100507463
PSMB9
LOC100294145
32.2
32.4
32.6
32.8
Position on chr6 (Mb)

(D)

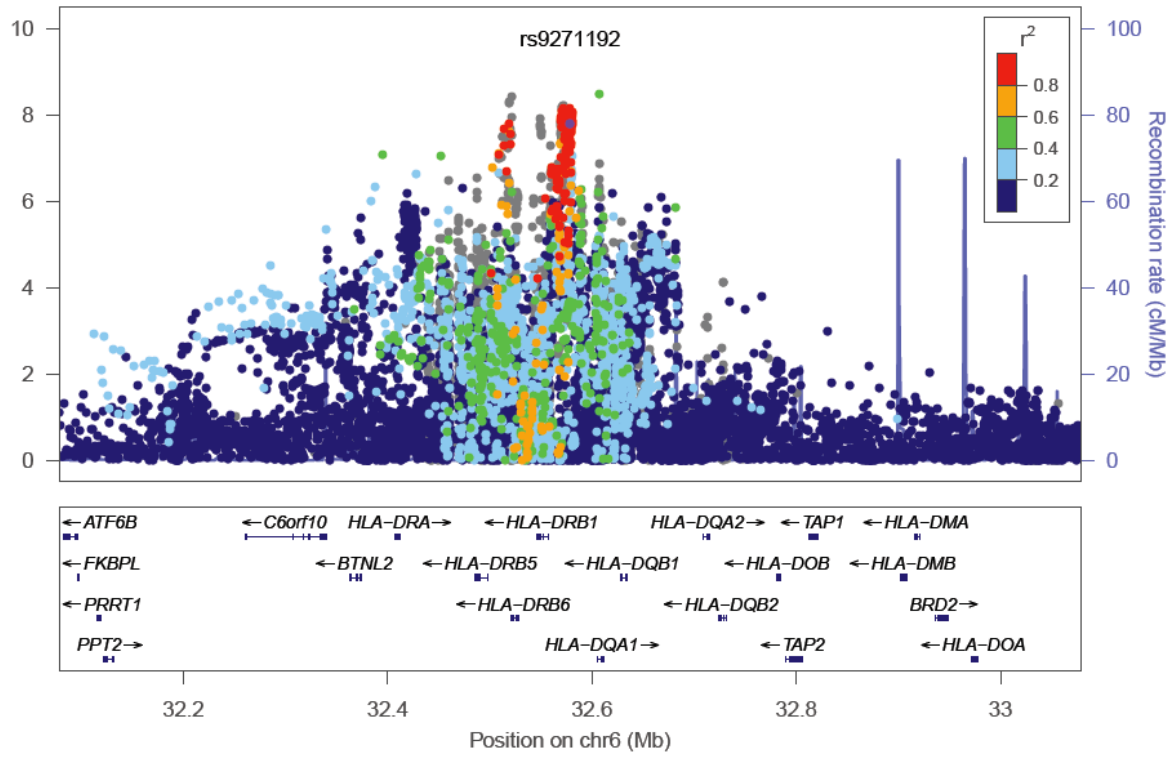
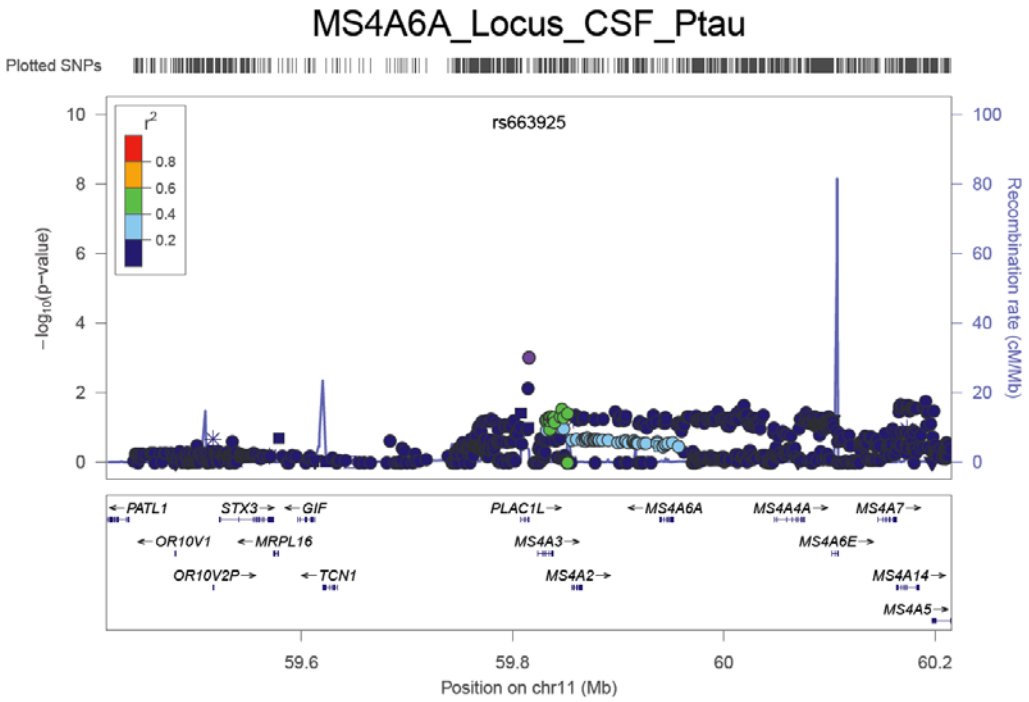
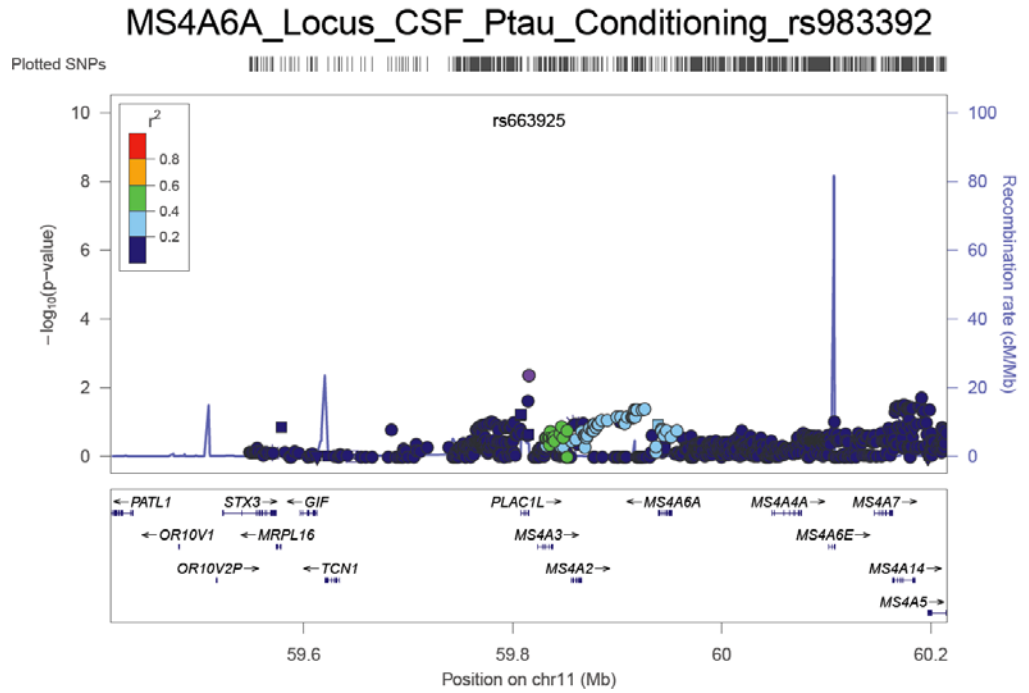

rs9271192
r2
0.8
0.6
0.4
0.2
Recombination rate (cM/Mb)
ATF6B
C6orf10
HLA-DRA
HLA-DRB1
HLA-DQA2
TAP1
HLA-DMA
FKBPL
BTNL2
HLA-DRB5
HLA-DQB1
HLA-DOB
HLA-DMB
PRRT1
HLA-DRB6
HLA-DQB2
BRD2
PPT2
HLA-DQA1
TAP2
HLA-DOA
Position on chr6 (Mb)

Figure S27. Regional plots for the *MS4A6A* locus with CSF ptau$_{181}$ levels before (A) and after conditioning on (B) rs983392 and (C) rs663925. (D) IGAP regional plot for the *MS4A6A* (rs983392).

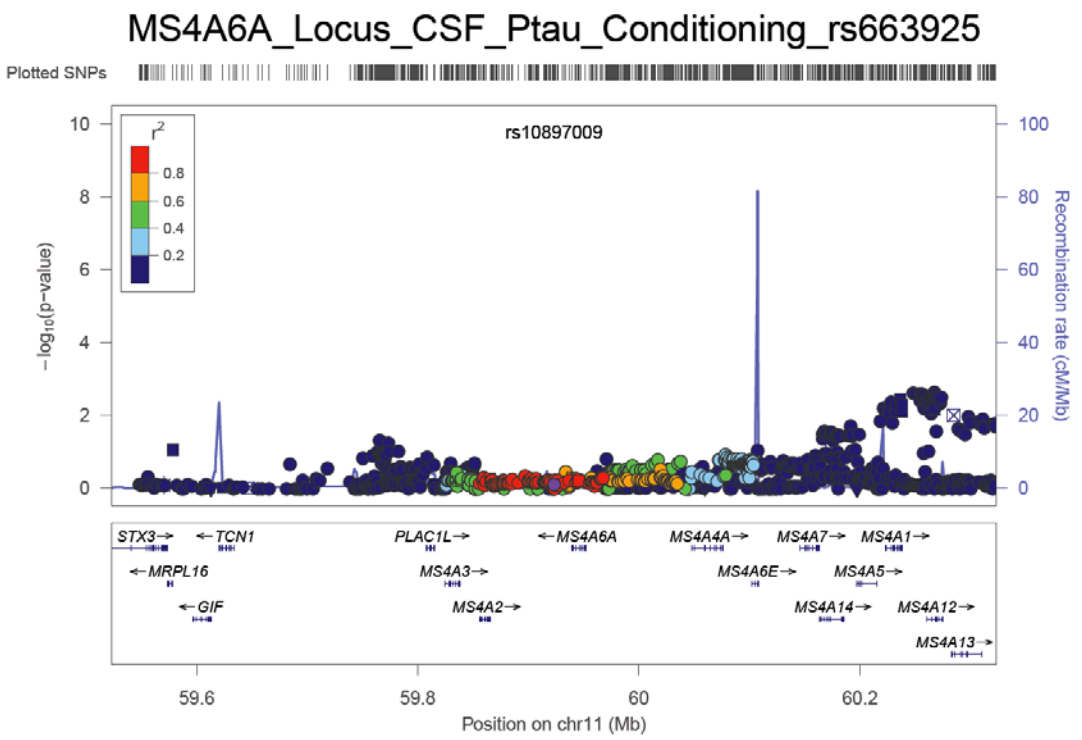
(A)



(B)



(C)


MS4A6A_Locus_CSF_Ptau_Conditioning_rs663925
Plotted SNPs
rs10897009
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
STX3→
←TCN1
PLAC1L→
←MS4A6A
MS4A4A→
MS4A7→
MS4A1→
←MRPL16
MS4A3→
MS4A6E→
MS4A5→
←GIF
MS4A2→
MS4A14→
MS4A12→
MS4A13→
59.6
59.8
60
60.2
Position on chr11 (Mb)

(D)

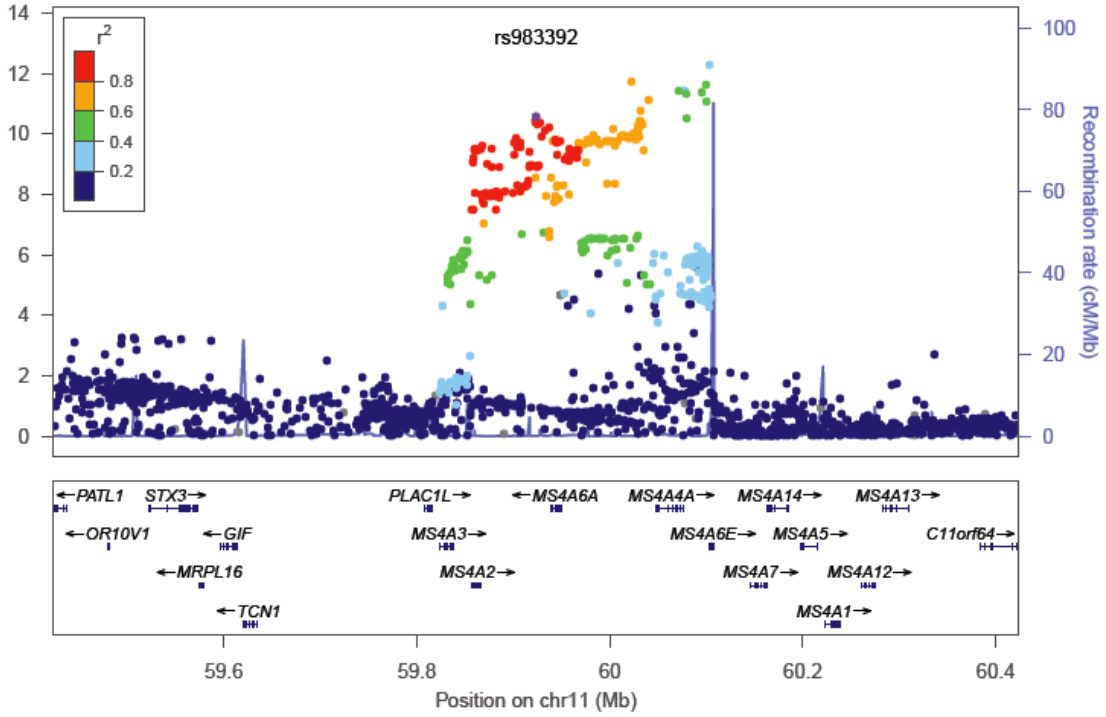

rs983392
r²
0.8
0.6
0.4
0.2
Recombination rate (cM/Mb)
←PATL1
STX3→
PLAC1L→
←MS4A6A
MS4A4A→
MS4A14→
MS4A13→
←OR10V1
←GIF
MS4A3→
MS4A6E→
MS4A5→
C11orf64→
←MRPL16
MS4A2→
MS4A7→
MS4A12→
←TCN1
MS4A1→
59.6
59.8
60
60.2
60.4
Position on chr11 (Mb)

Figure S28. Regional plot for the *ABCA7* locus with CSF ptau$_{181}$ levels

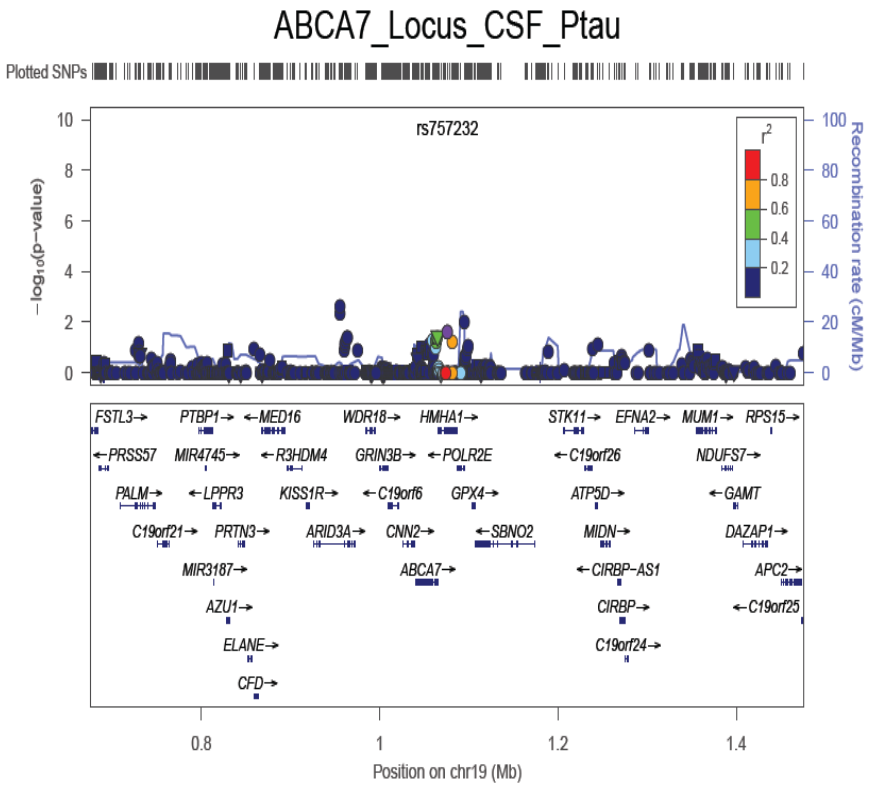


Figure S29. Regional plot for the *BIN1* locus with CSF ptau$_{181}$ levels

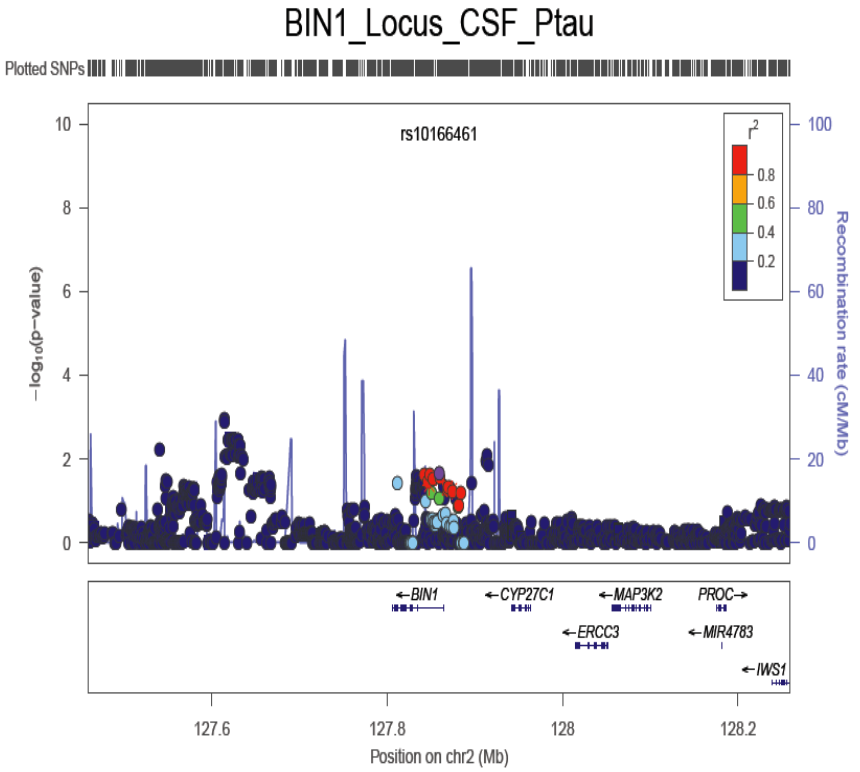


Figure S30. Regional plot for the *CASS4* locus with CSF ptau$_{181}$ levels

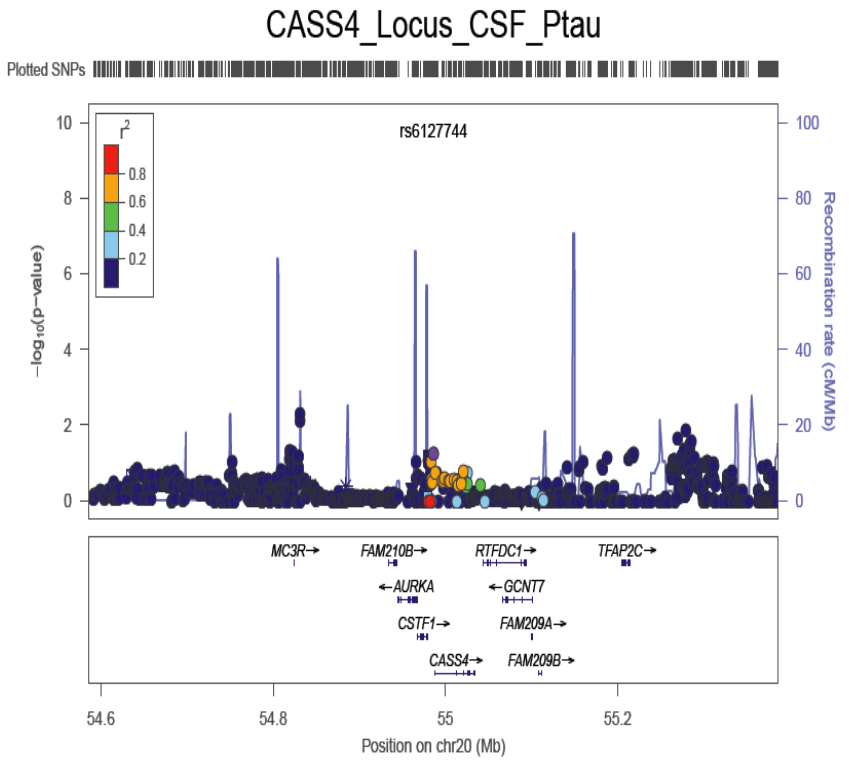


Figure S31. Regional plot for the *CD2AP* locus with CSF ptau$_{181}$ levels

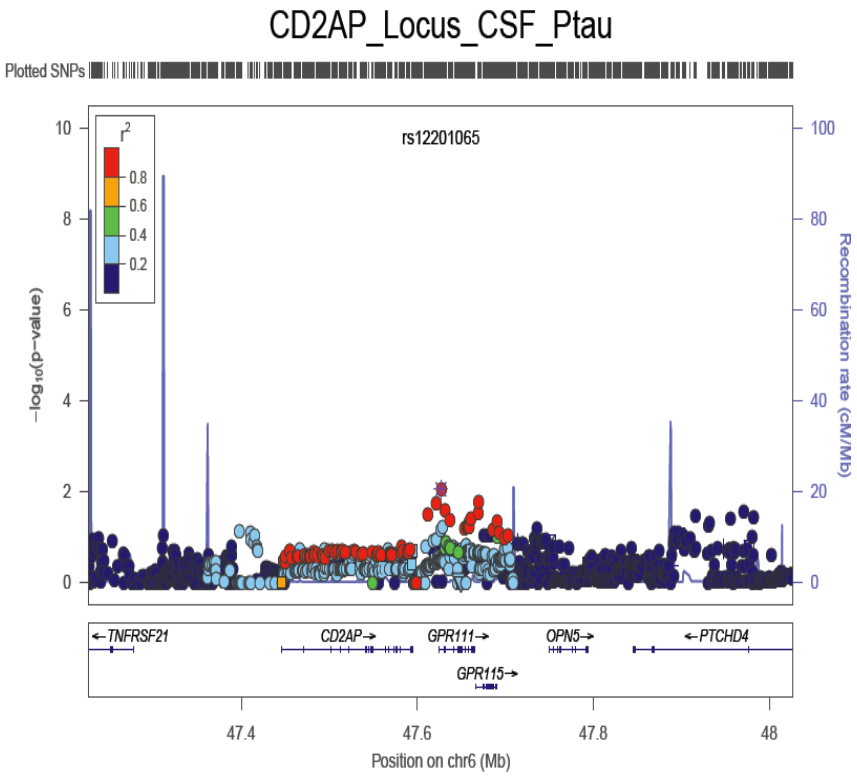


Figure S32. Regional plot for the *CD33* locus with CSF ptau$_{181}$ levels

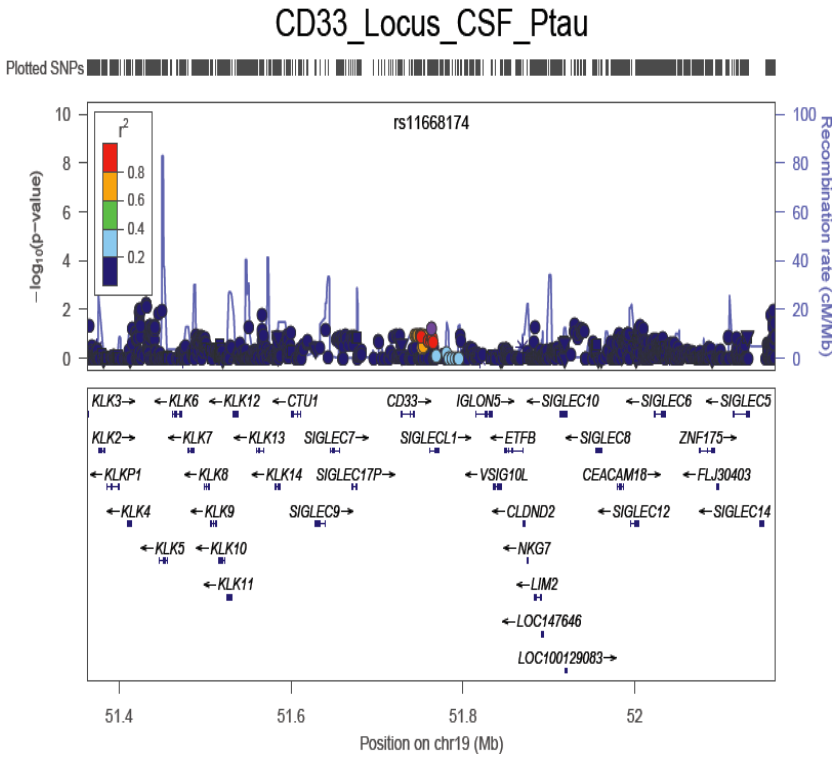


Figure S33. Regional plot for the *CELF1* locus with CSF ptau$_{181}$ levels

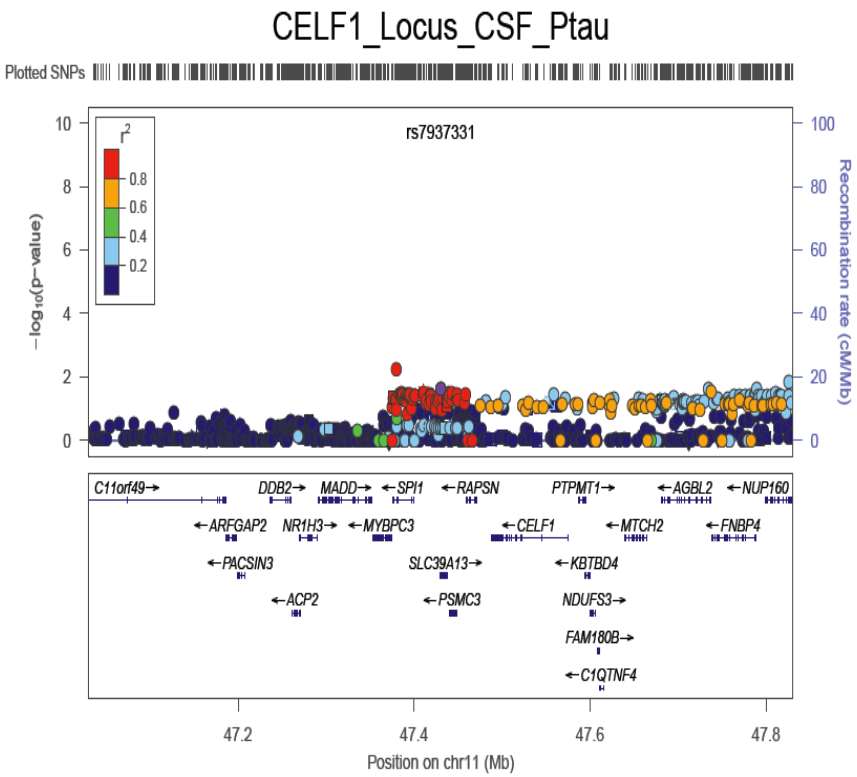


Figure S34. Regional plot for the *CLU* locus with CSF ptau$_{181}$ levels

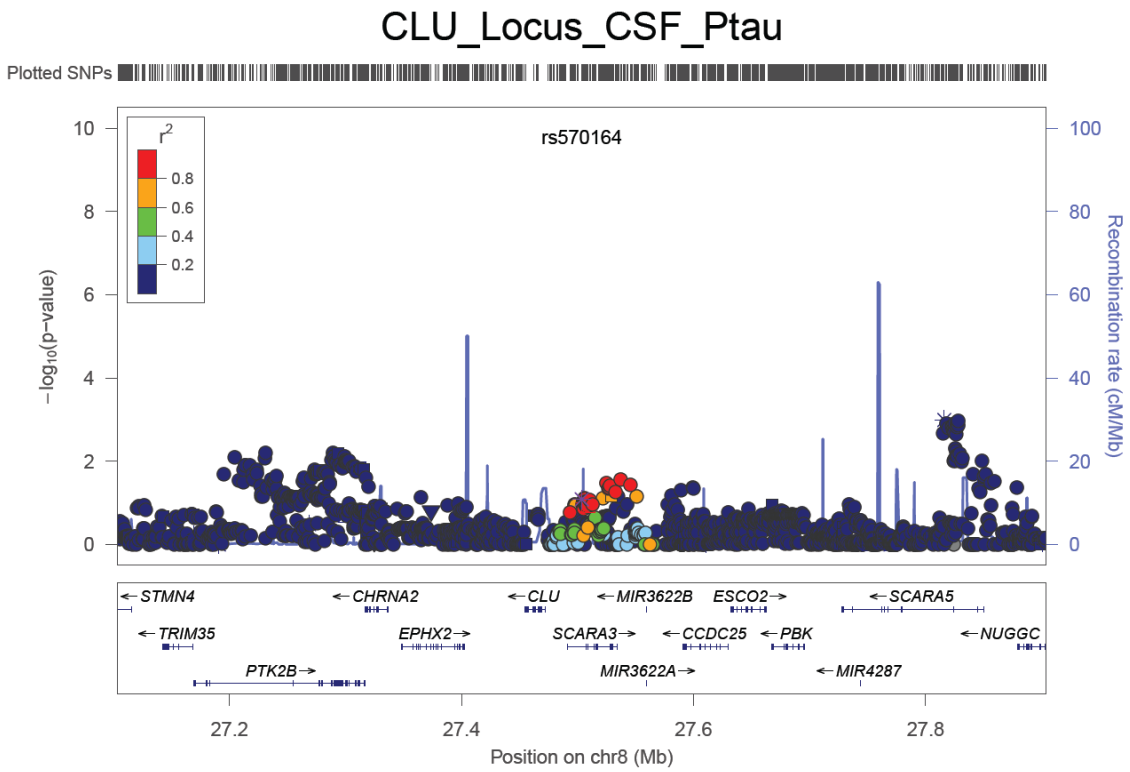


Figure S35. Regional plot for the *CR1* locus with CSF ptau$_{181}$ levels

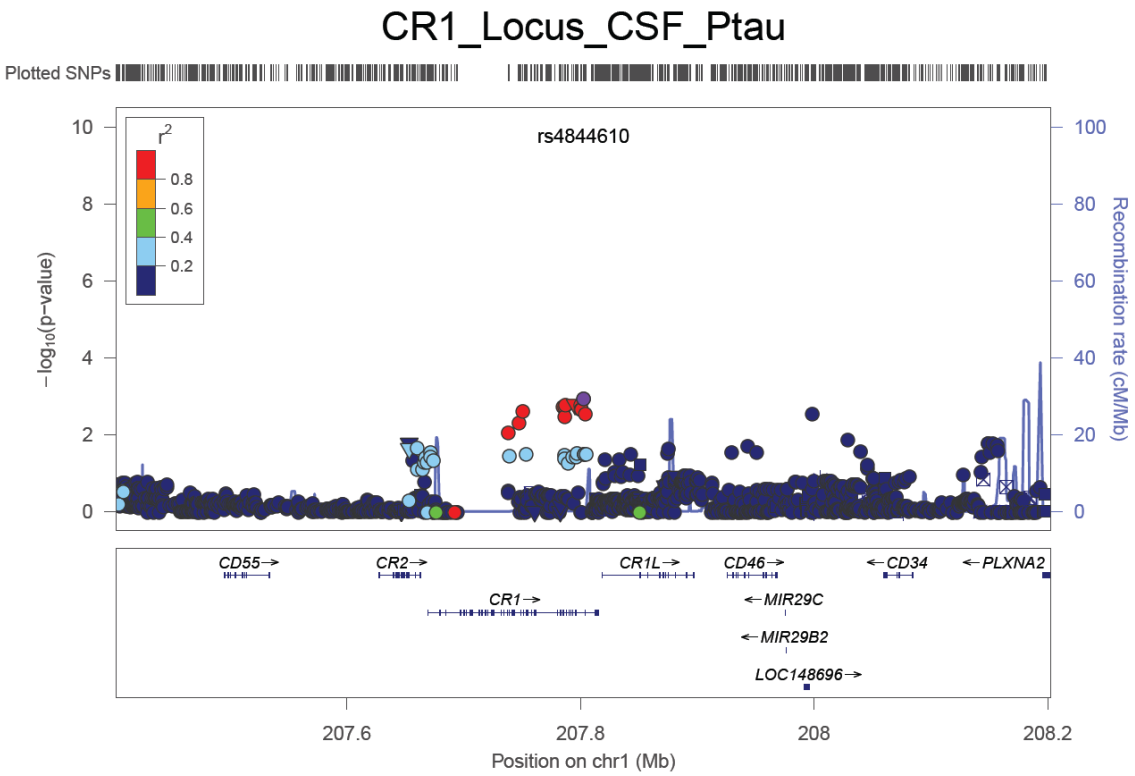


Figure S36. Regional plot for the *EPHA1* locus with CSF ptau$_{181}$ levels

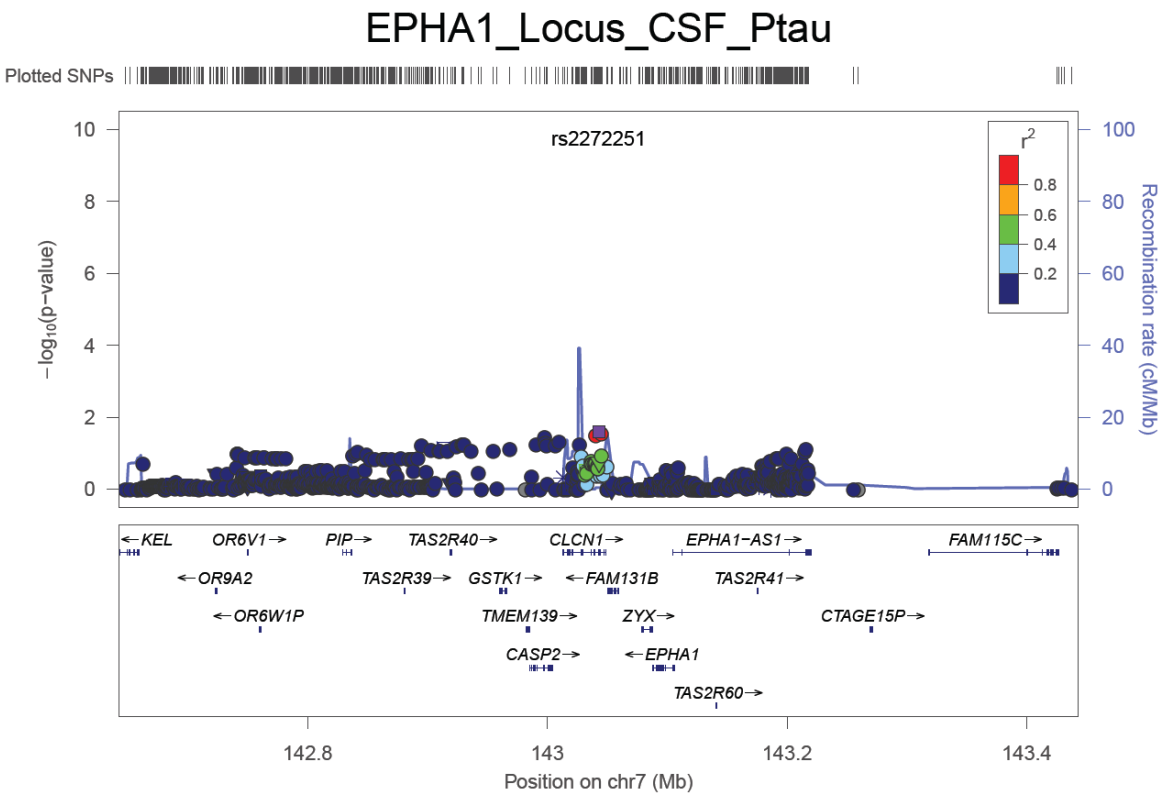


Figure S37. Regional plot for the *FERMT2* locus with CSF ptau$_{181}$ levels

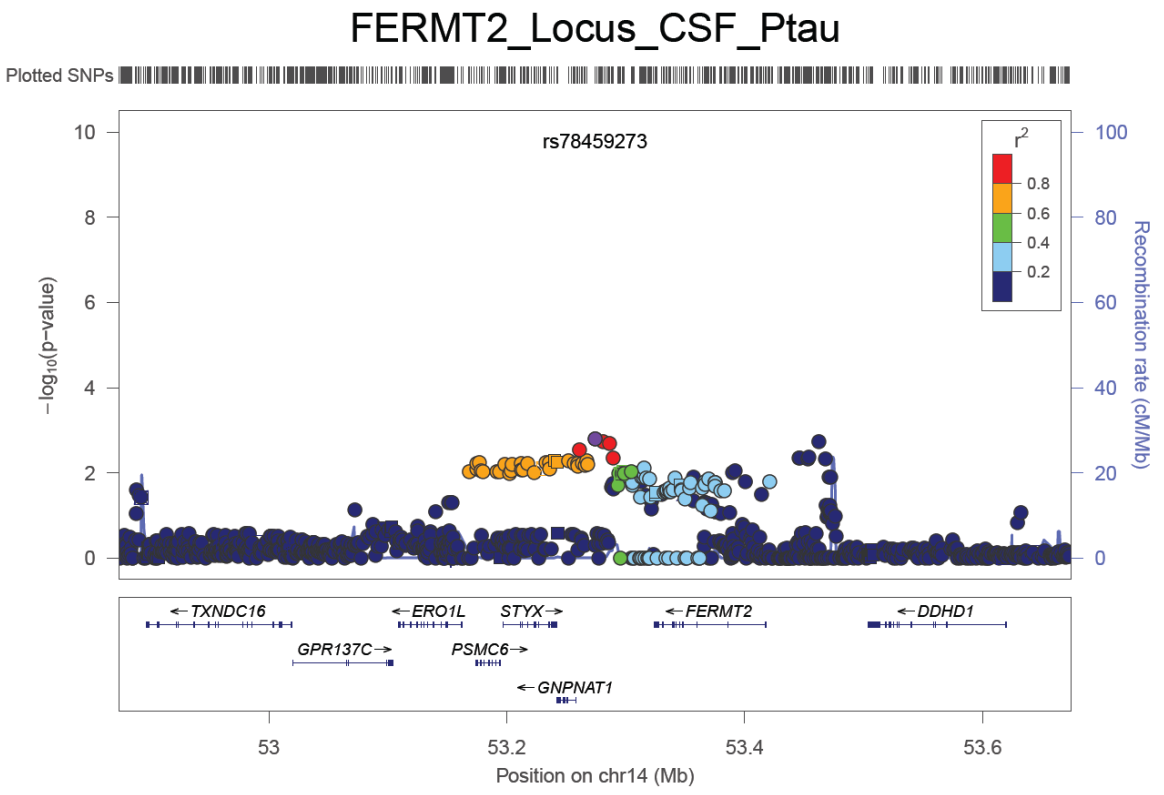


Figure S38. Regional plot for the *INPP5D* locus with CSF ptau$_{181}$ levels

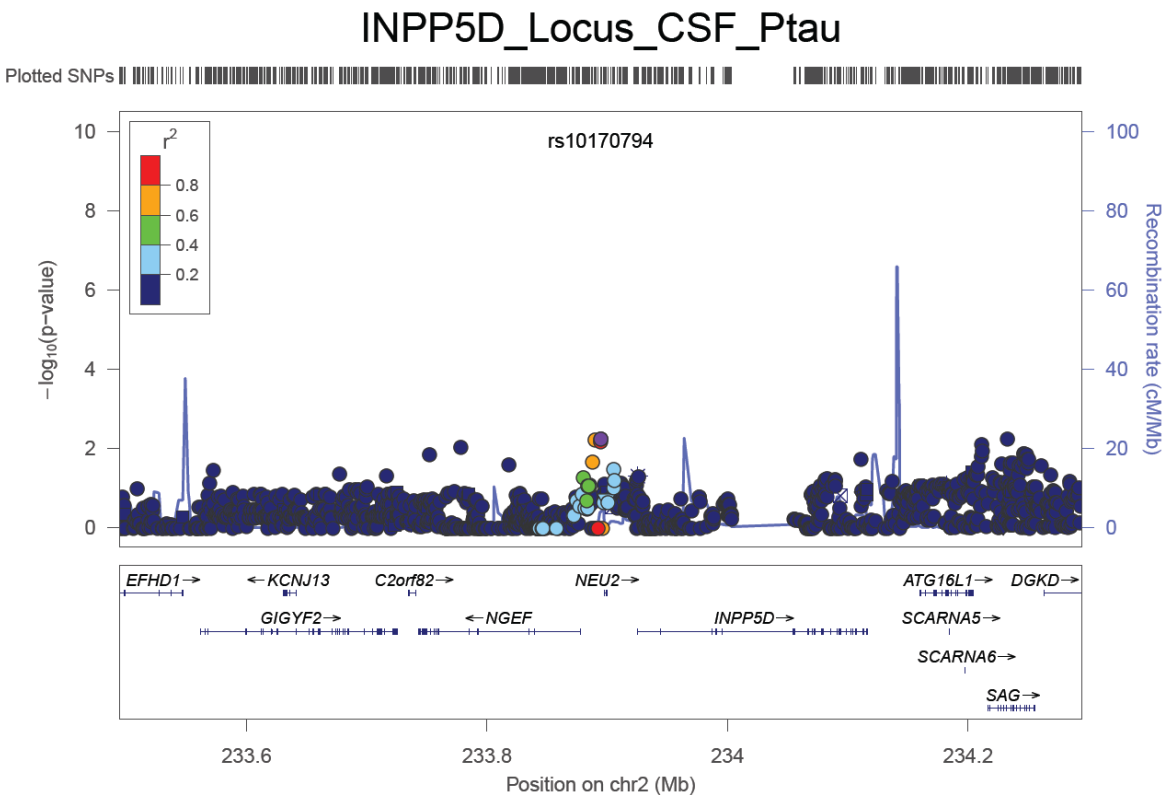


Figure S39. Regional plot for the *MEF2C* locus with CSF ptau$_{181}$ levels

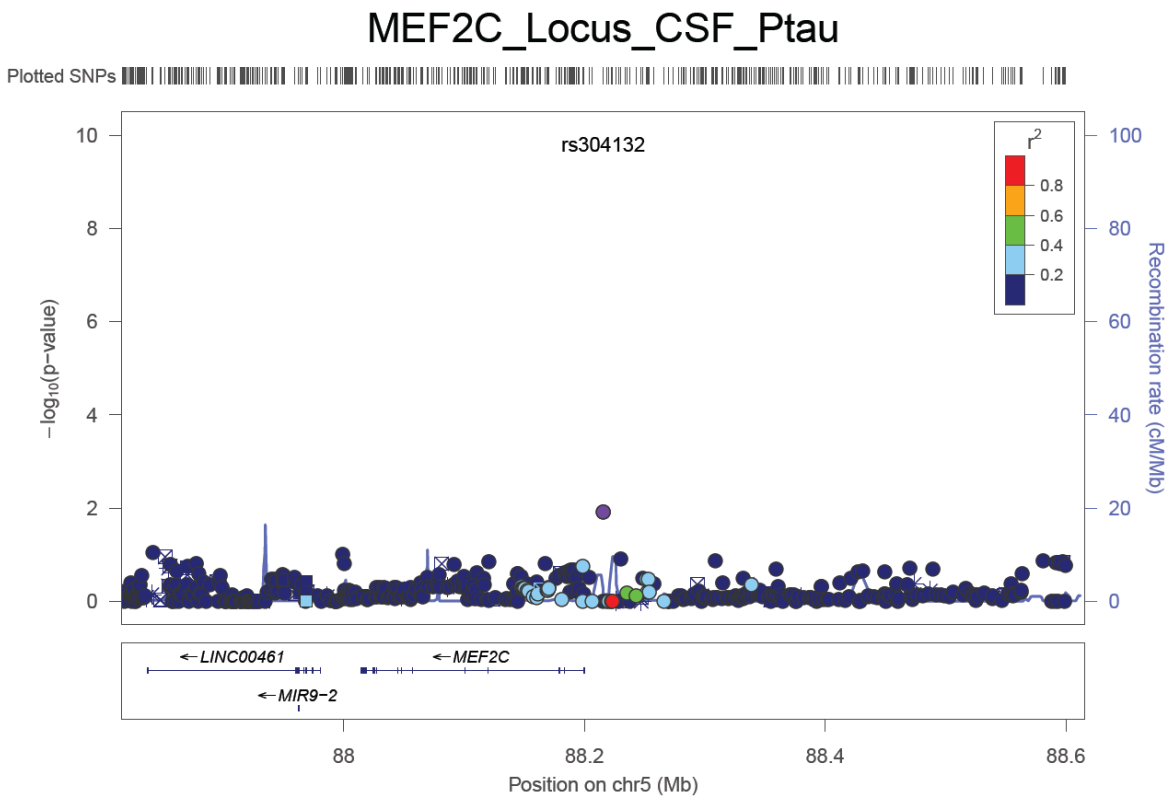


Figure S40. Regional plot for the *MS4A6A* locus with CSF ptau$_{181}$ levels

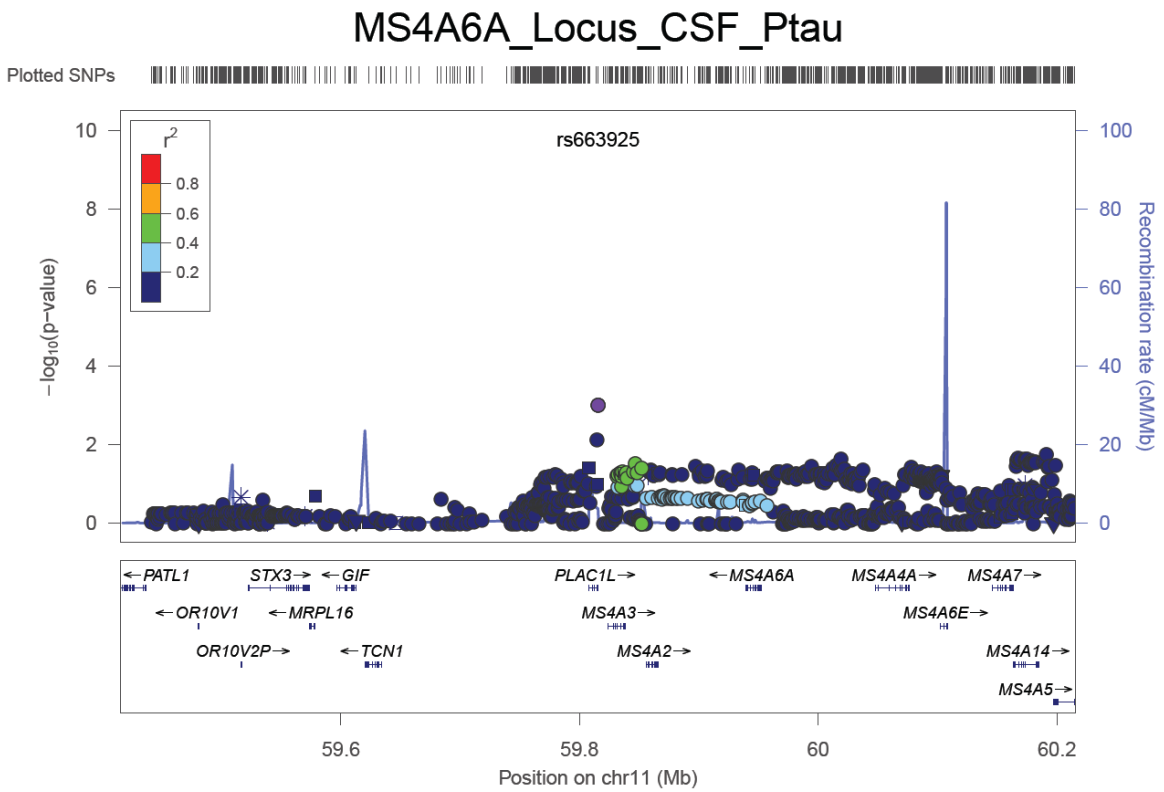


Figure S41. Regional plot for the *NME8* locus with CSF ptau$_{181}$ levels

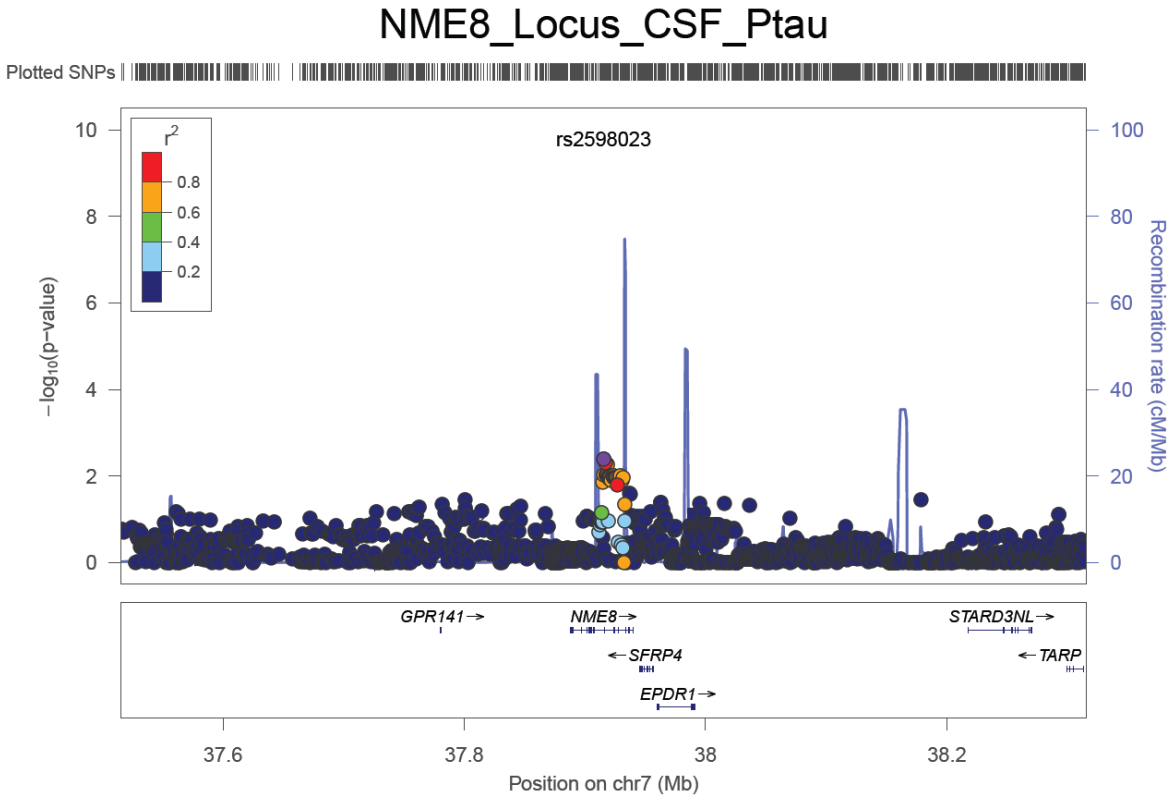


Figure S42. Regional plot for the *PICALM* locus with CSF ptau$_{181}$ levels

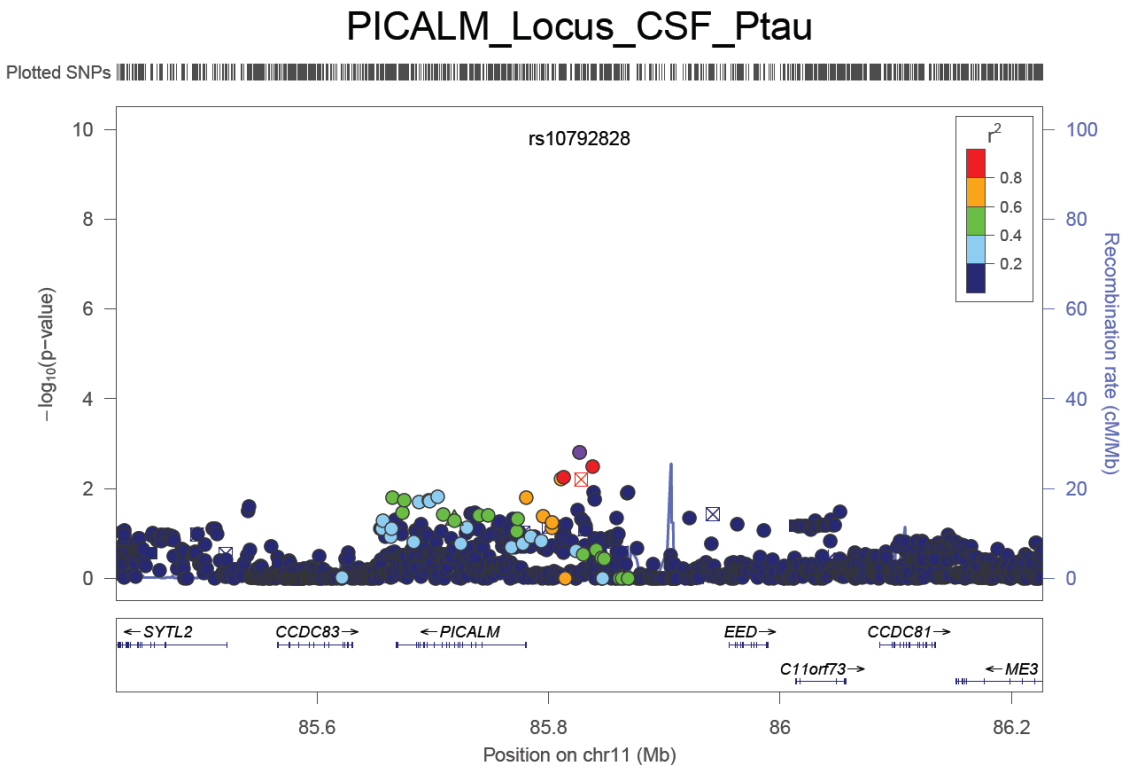


Figure S43. Regional plot for the *PTK2B* locus with CSF ptau$_{181}$ levels

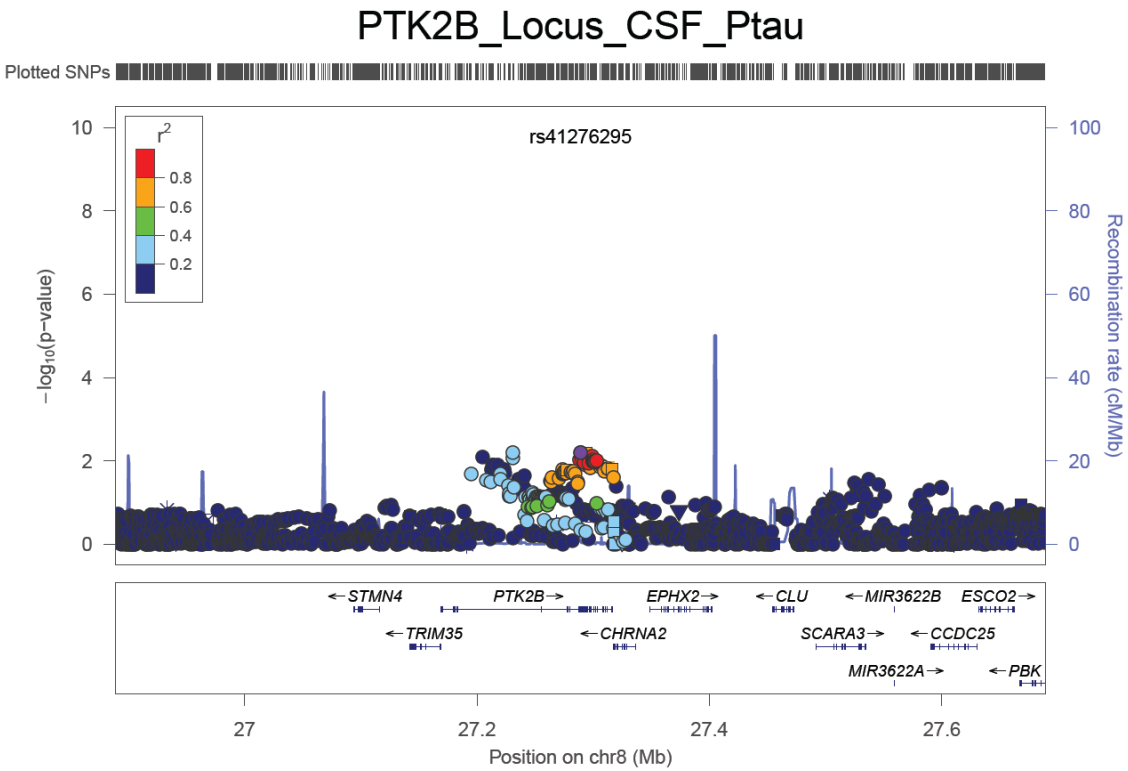


Figure S44. Regional plot for the *SLC24A4* locus with CSF ptau$_{181}$ levels

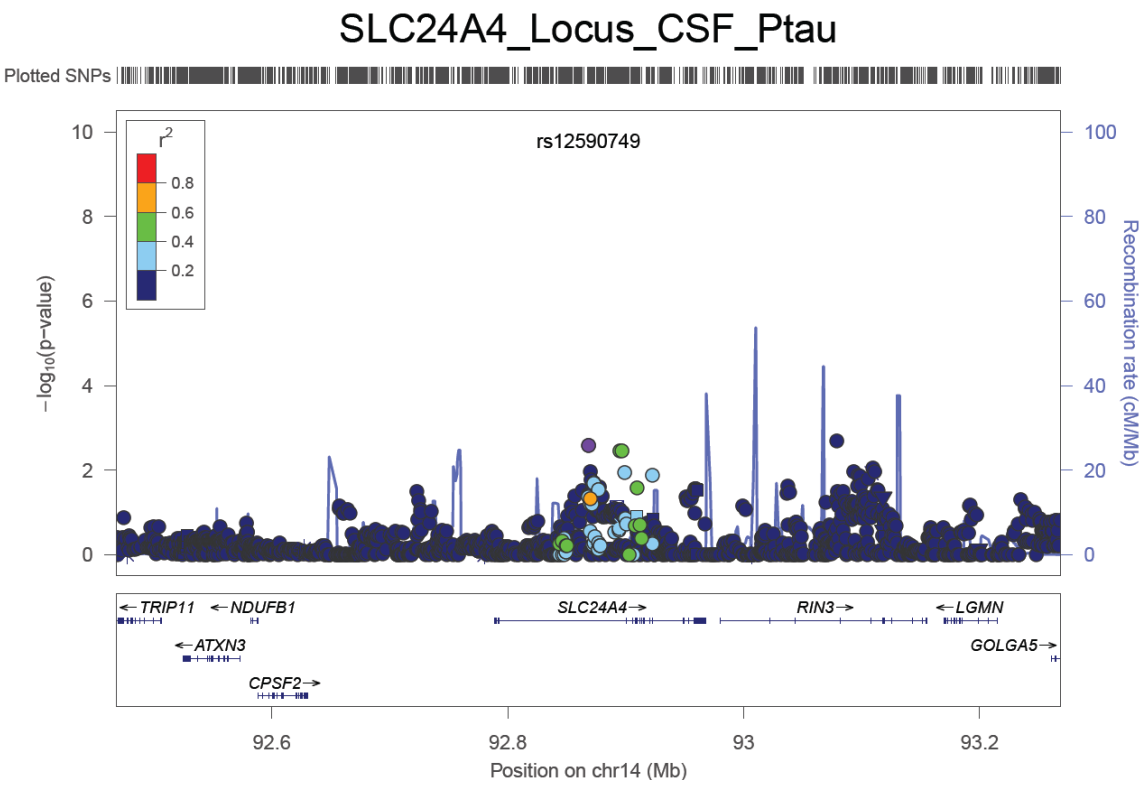


Figure S45. Regional plot for the *SORL1* locus with CSF ptau$_{181}$ levels

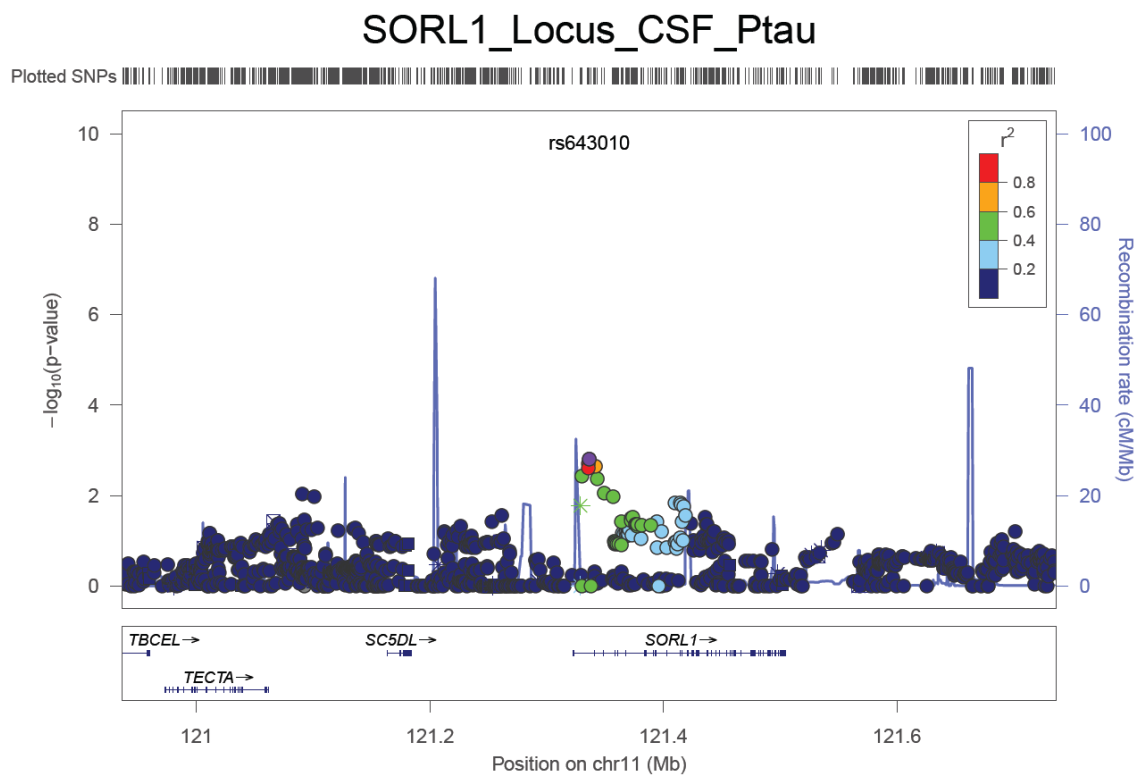
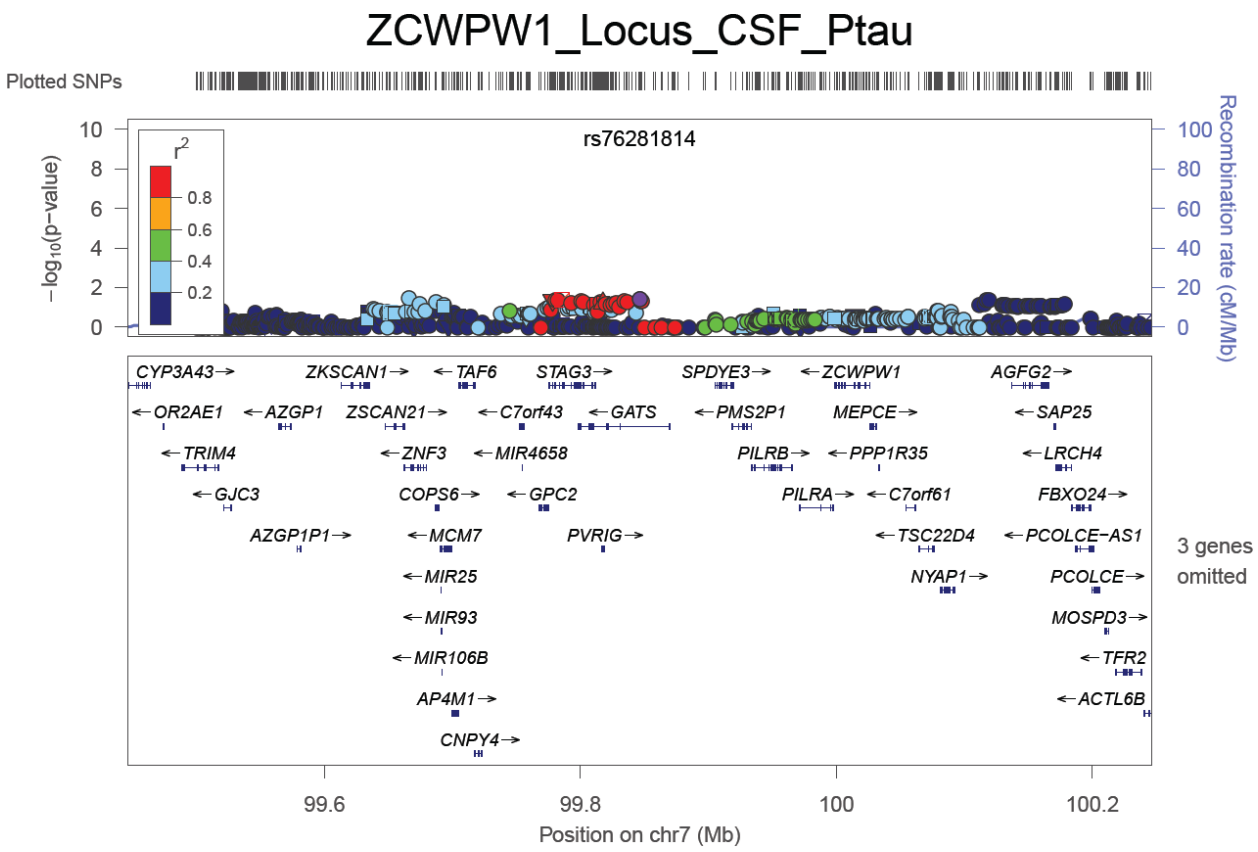


Figure S46. Regional plot for the *ZCWPW1* locus with CSF ptau$_{181}$ levels



Chapter 4

Functional Studies:

Cis- acting expression quantitative trait loci analyses

TREM2 cell surface expression studies

4.1 ABSTRACT

While our deep re-sequencing studies have identified numerous variants that can be underlying functional variants, the mechanisms by which these associated variants influence AD pathogenesis remain largely unclear. Follow-up functional studies are critical to understand the mechanisms behind genetic association and thus link disease-associated variants to disease pathogenesis. Our recent work has confirmed that the triggering receptor expressed on myeloid *2* (*TREM2*) is a *bona fide* AD risk gene. We found that p.R47H and p.R62H are significantly associated with AD risk and that several *TREM2* variants are only present in AD patients. Bioinformatic and protein conservation analyses also predicted that some of these variants are potentially functional. However, the mechanisms by which these variants affect molecular pathways involved in AD are unclear. We hypothesized that some of these *TREM2* variants may affect TREM2 trafficking to the cell surface. To test this hypothesis, we introduced *TREM2* variants into a TREM2-DAP12 cDNA construct using site-directed mutagenesis, expressed the construct in a T cell hybridoma cell line and measured cell surface expression using an antibody against the extracellular domain of TREM2. Our flow cytometry analyses suggest that p.T66M and p.R136W variants in *TREM2* robustly impact cell surface expression of TREM2 but we did not find any differences in cell surface expression comparing p.R47H and p.R62H to the wild type (WT), which suggests that p.R47H and p.R62H may affect AD through other mechanisms.

Although most of the genome-wide associations study (GWAS)-identified variants are located within non-coding regions, several recent studies have suggested that these non-coding variants may affect disease phenotypes by regulating expression levels in cis. Our RegulomeDB analyses have suggested that polymorphisms within the *CELF1* fine-mapping region may be expression quantitative trait loci (eQTLs) for nearby genes. Therefore, we performed cis-eQTL analysis for mRNA expression levels in several brain regions using four publicly available datasets to identify genetic determinants of gene expression in human brains. We found that all of the expression-associated SNPs, including rs7124681, rs10838738, rs2290850, and rs755553, are in tight linkage disequilibrium (LD) with rs7937331, the top cerebrospinal fluid (CSF) amyloid-beta 1-42 ($A\beta_{42}$) SNP in the *CEFL1* fine-mapping region. Conditional analyses suggested that there is only one independent signal within the *CELF* fine-mapping region. Moreover, the minor allele of the underlying causal variant is associated with reduced *C1QTNF4* expression levels. Additionally, we found evidence of differential expression in the *C1QTNF4* transcript between AD cases and controls in human brains. Overall, these results illustrated that the underlying causal variants and genes may not be

the gene originally identified in GWAS studies. Our data suggest the causal variants within the *CELF1* fine-mapping region mediate *C1QTNF4* expression levels and may affect AD risk by affecting beta amyloid (Aβ) biology. These findings support the hypothesis that genes involved in the inflammatory response play an important role in AD pathogenesis.

4.2 INTRODUCTION

Recent GWAS have uncovered several novel loci significantly associated with AD risk[8-11]. However, most of these loci are located in non-coding or gene-dense regions, making it difficult to identify causal genes responsible for the association. LD between the GWAS top SNPs with other SNPs also prevents the identification of casual risk variants for AD. Recent studies have suggested that eQTL-based analyses can relate the GWAS-identified risk variants to gene expression from a tissue or cell line, generating hypothesis regarding the putative mechanism associated with risk variants and genes[106-108,167,168]. Additionally, even though recent sequencing studies have shown that *TREM2* is a *bona fide* AD risk gene[12,13] and identified several potential functional variants[12,96,97], functional characterization of these novel and potentially functional *TREM2* variants in cell systems is essential to the understanding of disease mechanisms, which can eventually provide effective targets for therapeutics.

TREM2 is a type one transmembrane receptor protein expressed on myeloid cells including microglia, monocyte-derived dendritic cells, osteoclasts and bone-marrow derived macrophages [85,86]. Additionally, protein expression of TREM2 in neurons has been reported [87]. TREM2 transduces its intracellular signaling through DAP12 (TYROBP)[85,86]. Although the natural ligands of TREM2 remain unknown, upon ligand binding, TREM2 associates with DAP12 to mediate downstream signaling. In the brain, TREM2 is primarily expressed on microglia and has been shown to control two signaling pathways: regulation of phagocytosis and suppression of inflammatory reactivity[88-90]. A previous study used microarray and laser microdissection of Aβ plaque-associated areas in an animal model of AD and found that *TREM2* is differentially expressed in Aβ plaque-associated versus Aβ plaque-free tissue [91]. Several studies have shown that homozygous loss-of-function mutations in *TREM2* or *DAP12* are associated with Polycystic lipomembranous osteodysplasia with sclerosing leukoencephalopathy (PLOSL)[92-95]. Recent studies identified a *TREM2* variant p.R47H as a risk factor for LOAD with an OR around 3 [12,13], which is a similar effect size to the increased AD risk associated with carrying one *APOE* ε4 allele[67]. Several additional rare variants were enriched in AD cases; however, these variants failed to reach statistical significance[12,13,96].

In Chapter 2, we have confirmed that *TREM2* is a *bona fide* AD risk gene and identified two significant rare variants (p.R47H and p.R62H) for AD risk. Additionally, several *TREM2* variants were enriched or only identified in AD cases. However, it remains unclear how these risk variants are involved in AD pathogenesis. Since most of the identified variants are located in the TREM2 ectodomain, which is presumed to be involved in ligand binding, we hypothesized that these variants affect *TREM2* signaling by reducing ligand binding or by reducing cell surface expression. Here, we expressed potential functional variants in a T cell hybridoma cell line and performed flow cytometry analysis to investigate the effects of *TREM2* variants on cell surface expression.

In Chapter 3, we observed that rs7937331, an intronic SNP in *SLC39A13,* was significantly associated with CSF $A\beta_{42}$ levels. RegulomeDB predicts that several proxies in LD with rs7937331 may be cis-acting eQTLs for nearby genes. However, these regulatory function predictions are primarily based on data in blood and cancer cells. In this chapter, we performed cis-eQTL analyses for all genes in the LD region surrounding *SLC39A13*, to identify potential eQTL target-gene associations in several brain regions using four large-scale datasets. This information can help to identify underlying causal variants of disease-associated loci and to understand the underlying mechanism associated with GWAS-identified risk loci.

4.3 MATERIALS AND METHODS

4.3.1 Generation of wild-type and mutant constructs and cell-surface experiments

Based on our findings in the *TREM2* sequencing project, we selected seven missense variants for follow-up functional studies. *TREM2* p.T66M, a variant shown to cause Nasu-Hakola disease, was included as a positive control[124]. p.R47H was chosen because the association with AD has been replicated across several studies[12,13,97,125,136]. *TREM2* P.R52H, p.R62H, p.R136W, p.E151K, p.H157Y were selected because they are only present in AD cases in our study or likely to affect protein structure based on SIFT/PolyPhen predictions. Each mutation was introduced into the cDNA construct pIRES-hDAP12 full length human TREM2 (hTREM2) WT (provided by Dr. Marco Colonna; **Figure 1**) using site-directed mutagenesis, per the manufacturer's protocol[85]. These retroviral vectors were transfected into plat-E cells, an efficient and stable cell line for transient packaging of retroviruses, using Lipofectamine 2000. Viral supernatants were collected after 48 hours of incubation. Nuclear Factor of Activated T-cells (NFAT) 43.1 reporter cells, which express GFP under the control of the NFAT promoter,

were transduced with virus containing *TREM2* variants. The resulting cells were analyzed for cell surface expression of TREM2 by flow cytometry using a TREM2-specific monoclonal antibody developed by Dr. Colonna[86].

4.3.2 Selection of genes and variants

In Chapter 3, several proxy SNPs in LD ($R^2 \geq 0.8$) with rs7937331 in *SLC39A13* were predicted to have regulatory functions based on RegulomeDB database and may be the underlying causal variants. Examination of eQTL-gene associations provides an extra layer of information to aid in the identification of functional regulatory variants for disease-associated loci. To investigate the cis-eQTL effects on nearby genes, we examined the regional plot for CSF $A\beta_{42}$ association within the *SLC39A13* region (see **Figure S4 of Chapter 3**) and decided to evaluate the association with probes tagging 19 genes: *DDB2*, *ACP2*, *NR1H3*, *MADD*, *MYBPC3*, *SPI1*, *SLC39A13*, *PSMC3*, *RAPSN*, *CELF1*, *PTPMT1*, *KBTBD4*, *NDUFS3*, *FAM180B*, *C1QTNF4*, *MTCH2*, *AGBL2*, *FNBP4*, and *NUP160*. We retrieved all genotyped SNPs which were located between chr11:47226493 (10 kb upstream of the 5' end of *DDB2*) and chr11: 47880057(10 kb downstream of the 3' end of *NUP160)*. Variants within this region were tested for association with mRNA expression levels of aforementioned genes.

4.3.3 Datasets

We downloaded four publicly available expression datasets including GSE15745[108], GSE36192[106], GSE15222[107] and GSE5281[169] from the NCBI's Gene Expression Omnibus. We obtained the genotype data via the NCBI dbGaP authorized access portal (http://www.ncbi.nlm.nih.gov/gap).

4.3.3.1 GSE15745+GSE36192

The GSE15745 dataset was originally utilized in a study performed by Gibbs et al. to annotate and understand the impact of genetic variation on gene expression in human brains[108]. The mRNA profiling was performed using the Illumina humanRef-8 v2.0 expression beadchip. SNP genotyping was performed using the Infinium HumanHap550 beadchip (Illumina). Tissue samples of the cerebellum (CRBLM), frontal cortex (FCTX), pons (PONS), and temporal cortex (TCTX) from 150 neurologically normal Caucasian individuals were received from the University of Maryland Brain Bank, Baltimore. None of the individuals were of Hispanic descent and none were previously-diagnosed with neurological or cerebrovascular disease or cognitive impairment during life. The

samples had an average age at time of death of 45.8 yrs (range, 15-101) and 31.3% of them were female. The mean post-mortem interval (PMI) equals 14.3 hours (range, 3-28).

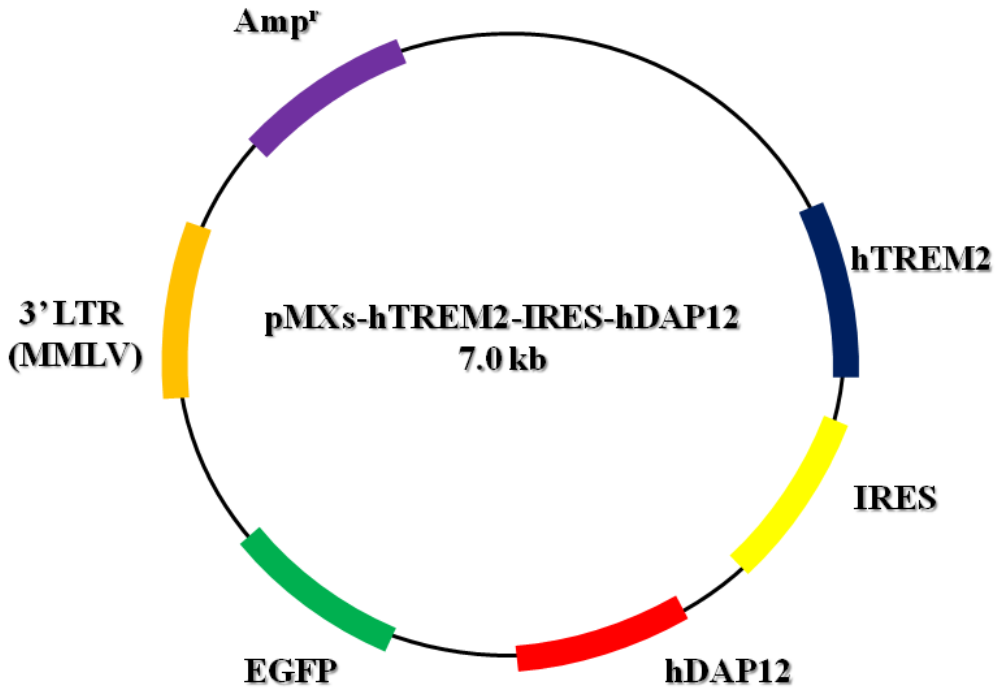


Figure 1: Schematic representation of pMXs-hTREM2-IRES-hDAP12 retroviral vector.

The GSE36192 dataset was an extension of the original study conducted by Gibbs et al[108] and included additional 712 frontal cortex and cerebellum samples from the InCHIANTI study[170]. However, genotype data was only available for 232 of these individuals. In this new dataset, the same expression assay (Illumina HT-12 beadchips consisting of 48,000 probes) was used for all individuals. The 232 new samples had an average age at time of death of 49.5 yrs (range, 0.42-102) and an average PMI of 12.7 hours (range, 1-96). 147 brains (64.5%) were from male donors. We combined genotype data for 150 samples from GSE15745 and 232 from GSE36192, which consists of data from the Illumina HumanHap550 v3, Human610-Quad v1 or Human660W-Quad v1 Infinium BeadChip. After quality control filtering (see **Data cleaning**), 330 FCTX and 330 CRBLM were used for eQTL analyses to investigate the association between 19 mRNA transcripts and 48 single nucleotide polymorphisms (SNPs). Since there was no new pons and temporal cortex data from GSE36192, we performed eQTL analyses for

the same samples (142 PONS and 144 TCTX) as those used in Gibbs et al[108]. The numbers of brain samples and numbers of SNPs used for analyses in each brain region were summarized in **Table 1.**

4.3.3.2 GSE15222

We obtained the GSE15222 dataset from a previous study investigating the association between human brain transcriptome and genetic variants[107]. Brains from neuropathologically normal postmortem controls (N = 188) and pathologically-confirmed AD cases (N=176) were received from 20 National Alzheimer's Coordinating Center (NACC) brain banks and from the Miami Brain Bank. The 188 control brains consisted of 21% frontal cortex, 73% temporal cortex, and 2% parietal cortex. 45% of the control samples were female with an average age of 82.3 yrs (range, 65-102) and an average PMI of 9.3 hours (range, 1.17-54). The 176 AD brains contained 19.5% frontal cortex, 66.5% temporal cortex, and 5.5% parietal cortex and 8.5% cerebellum. The Affymetrix GeneChip Human Mapping 500K Array Set was performed to generate genotype data and the Illumina Human RefSeq-8 Expression BeadChip was used to generate the RNA expression data. After quality control (QC), 188 controls and 176 cases were used for eQTL analysis. The numbers of brain samples and numbers of SNPs used for analyses in each brain region are summarized in **Table 1.** For detailed descriptions about sample preparation and study design, please check Webster et al[107].

4.3.3.3 GSE5281

The original study using the GSE5281 dataset investigated the association between RNA expression in neuronal nuclear genes and mitochondrial energy metabolism[169]. The GSE5281 dataset included laser-capture microdissected non-tangle bearing neuronal cells from 47 human brains of Caucasian origin that were collected at three Alzheimer's Disease Centers (Washington University, Duke University, and Sun Health Research Institute). Samples are composed of clinical and neuropathologically-confirmed AD cases or neurologically confirmed non-demented controls. Thirty-three samples were AD cases and 54.5% of them were female with an average age 79.9 ± 6.9 years. Fourteen samples were controls and 28.6% of them were female. The mean age was 79.8 ± 9.1 years. Samples from six brain regions were collected (mean PMI = 2.5): the entohinal cortex (EC; Broca's area [BA] 28 and 34), superior frontal gyrus (SFG; BA 10 and 11), hippocampus (HIP), primary visual cortex (VC; BA 17), middle temporal cingulate cortex (MTG; BA 21 and 37), and posterior cingulate cortex (PCC; BA 23 and 31). RNA

Table 1. Summary counts of number of individuals, probes, and SNPs tested per brain region

GSE15745+GSE36192	FCTX	CRBLM	PONS	TCTX
# Samples	330	330	142	144
# probes tested	22	22	18	18
# SNPs tested	48	48	52	52

GSE15222	FCTX	Parietal	TCTX	CRBLM
# Samples	71	20	242	31
# Probes tested	1	1	1	1
# SNPs testedd	30	30	30	30

GSE5281	Entorhinal cortex	Hippocampus	Middle temporal gyrus	Posterior cingulate cortex	Superior frontal gyrus	Visual cortex
# Samples	23	23	28	22	34	31
# Probes tested	1	1	1	1	1	1

FCTX: frontal cortex; CRBLM: cerebellum; PONS: pons; TCTX: temporal cortex.

expression profiling was performed using the Affymetrix Human Genome U133 Plus 2.0 Array. The numbers of brain samples and numbers of SNPs used for analyses in each brain region were summarized in **Table 1.** See Liang et al. for more details[169,171].

4.3.4 Data cleaning

We only performed data cleaning for GSE15745[108] and GSE36192[106] datasets as QC has been conducted for the GSE15222[107] and GSE5281[169,171] datasets. We identified unanticipated duplicates and cryptic relatedness using pair-wise genome-wide estimates of proportion identity by descent (IBD) using the PLINK program[122]**.** When duplicate samples or a pair of samples with cryptic relatedness was identified, priority was given to samples with higher SNP call rates. In order to control for population substructure, a principal component analysis (PCA) was conducted using the EIGENSTRAT software[121]. HapMap samples (CEU: CEPH Europeans from Utah; JPT: Japanese in Tokyo; YRI: Yoruba in Ibadan, Nigeria) were included in the analyses in order to remove outliers and confirm self-reported ethnicity. Individuals were excluded if they were not located within the CEU cluster (see **Figure S1**). We also checked whether the gender was discordant by analyzing the X-chromosome SNPs using PLINK. Individuals were removed if the recorded gender did not match the gender reported by our analyses. As previous studies have shown that younger brains may present abnormal RNA expression in genes related to cell cycle, DNA damage repair, and cell differentiation[172], in the analyses, samples were excluded if the age was under 15 years. Samples were excluded if they were outliers based on mean expression levels. For GSE15745 and GSE36192 datasets, final analyses included 330 FCTX, 330 CRBLM, 142 PONS, and 144 TCTX. Stringent quality thresholds were applied to the genotype data. SNPs were dropped if they fulfilled any one of the following criteria: i) genotyping success rate<98% per SNP or per individual; ii) HWE ($p<1\times10^{-3}$); iii) MAF<0.02. QCs were carried out jointly after combining SNPs common in both datasets (GSE15745 and GSE36192). A total of 52 SNPs were analyzed for PONS and TCTX and 48 SNPs were used for FCTX and CRBLM.

4.3.5 Statistical analysis

For the analysis of GSE15745 and GSE36192, a log 10 transformation of the mRNA expression measurements was applied in order to normalize the distribution. We then performed a two-step analysis for each of

the four brain regions. The first step involves correction for known biological and methodological covariates by performing the following multivariate regression model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \varepsilon$$

Where Y is the log10 transformed expression value and X_1 …… X_n indicate the biological covariates, which include age and gender, and methodological covariates, which include PMI, which Brain Bank these samples was from and which preparation/hybridization batch the sample was processed in. Within this model, gender, tissue bank, and batch were coded as dummy variables. After fitting each trait to the model, the residuals from the model were kept and represent the dependent variable in the second regression. This step is essential as variance attributable to gender, age, PMI, tissue source and hybridization batch were removed prior to eQTL analyses. Next, the second-step analysis was performed by regressing the residuals against allele dosages for each SNP as follows.

$$Y = \beta_0 + \beta_1 \mathrm{ADD} + \varepsilon$$

Where Y represents the residuals and ADD represents the genotype encoded as allele dosage. The R software (version 3.0.2) was used to perform all of the analyses.

We then analyzed the GSE5281 dataset for association between mRNA expression levels and AD disease status for specific genes of interest. We performed joint analysis for $\log_{10}$ transformed mRNA expression values adjusting for age, gender, AD disease status, and which brain region the sample was from. Within the model, gender, AD disease status and brain region were treated as dummy variables. We also performed separate analyses for each brain region by conducting a two-tailed unpaired *t*-test.

In order to replicate the suggestive SNP-transcript pair association found in GSE15745 and GSE36192 datasets, we analyzed the GSE15222 dataset using 188 neuropathologically normal postmortem controls in order to prevent the confounding effects due to AD disease status. For consistency, we first regressed out the effects due to known biological and methodological covariates and then fitted the residuals against allele dosage for each SNP. We also compared the expression level between 176 AD cases and 188 controls for the probes of interest.

For GSE15745 and GSE36192, the Bonferroni correction was used to control for type-1 error rate (0.05). Since 52 SNPs were tested for association with 18 transcripts within 4 brain regions, the multiple-testing correction threshold is approximately equal to $0.05/(52\times4\times18) = 1.33\times10^{-5}$.

4.3.6 Bioinformatic and linkage disequilibrium analyses

We used the SeattleSeq Annotation server (http://snp.gs.washington.edu/SeattleSeqAnnotation138/) and the SNP and CNV Annotation Database (http://www.scandb.org) to annotate the variants. The LocusZoom tool was used to create the regional plots[173]. The Haploview software[174] was used to estimate the LD structure and to generate the LD plots for suggestive and significant eQTLs.

4.4. RESULTS

4.4.1 TREM2 cell surface expression

TREM2 is a cell surface receptor protein expressed on myeloid cells and upon ligand binding, TREM2 transduces its intracellular signaling through association with DAP12[85,86]. In this study, we tested the effects of rare variants in *TREM2* on TREM2 cell surface expression. NFAT 43.1 reporter cells, a cell line expressing GFP under the control of the NFAT promoter, were transduced with viral particles containing hDAP12 and either hTREM2 WT, p.T66M (a known Nasu-Hakola causal mutation served as the positive control), p.R47H, P.R52H, p.R62H, p.R136W, p.E151K, or p.R157Y. In hTREM2-WT expressing cells, 74.8 % of TREM2 is detected at the cell surface (see **Figure 2A**). In contrast, the positive control cells expressing a Nasu-Hakola mutation, hTREM2-p.T66M, have only 16.8% of expressed TREM2 at the cell surface (see **Figure 2H**). Cells expressing p.R47H (60.8%, **Figure 2B**), p.R52H (67.5%, **Figure 2D**), p.R62H (67.9%, **Figure 2F**), p.E151K (72.3%, **Figure 2C**), and p.H157Y (52.2%, **Figure 2E**) exhibited slightly lower levels of cell surface TREM2 expression but they were not significantly different from hTREM2 WT. Interestingly, we observed a robust effect on TREM2 cell surface expression in cells expressing p.R136W (28.8%, **Figure 2G**), comparable to cells expressing p.T66M. Our data suggest that *TREM2* p.T66M and p.R136W variants have a robust effect on cell surface expression of TREM2. Replication studies have

been performed and the results were similar to those in the first run.

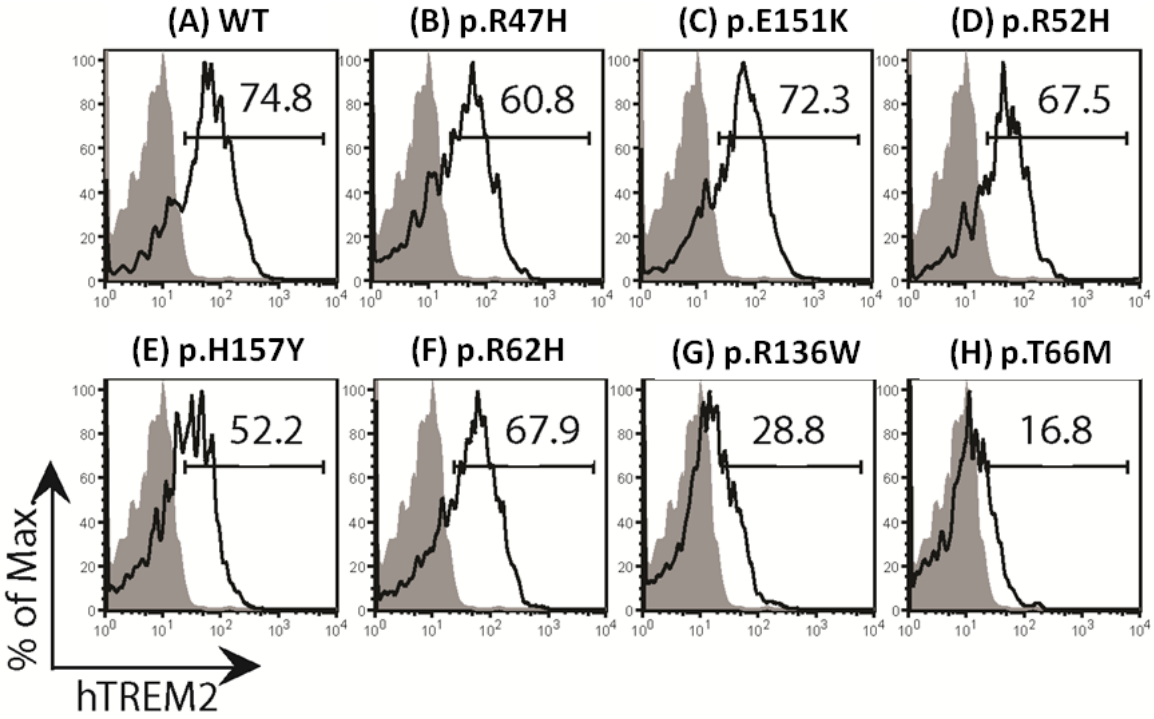


Figure 2: Flow cytometry of NFAT43.1 reporter cells from (A) hTREM2-WT (B) p.R47H (C) p.E151K (D) p.R52H (E) p.H157Y (F) p.R62H (G) p.R136W and (H) p.T66M. Analyses suggested that *TREM2* p.R136W and p.T66M variants significantly impair the cell surface expression of TREM2.

4.4.2 Cis-actings eQTLs of *C1QTNF4* expression levels

To identify genetic variants within the *CELF1* fine-mapping region that regulate gene expression in normal human brains, we conducted cis-eQTL analysis for four brain regions (FCTX, CRBLM, PONS, and TCTX) using two large microarray datasets (GSE15745 and GSE36192)[106,108]. For PONS and TCTX, we analyzed the GSE15745 dataset with 150 neurologically normal Caucasian individuals obtained from the University of Maryland Brain Bank, Baltimore. For FCTX and CRBLM, we combined the GSE15745 dataset with the GSE36192 dataset, which consists of FCTX and CRBLM genotype and phenotype data from 232 neurologically normal Caucasian individuals in the InCHIANTI study[170] to maximize the power of detecting cis-eQTLs. We extracted genotype data between chr11:47226493 (10 kb upstream of the 5' end of *DDB2*) and chr11: 47880057(10 kb downstream of the 3' end of *NUP160)*. After QC (see **MATERIALS AND METHODS**), 52 variants were analyzed in 142 samples and 144

samples for association with 18 transcripts in PONS and TCTX respectively while 48 variants were analyzed in 330 samples for 22 transcripts in FCTX and CRBLM (see **Table 1**).

To minimize brain region specific effects on expression levels, we performed analyses separately for each brain region and applied a two-step regression approach to cis-eQTLs. We also checked whether the identified variants were located within the probe sequences, which could result in inaccurate mRNA expression due to differential hybridization (see **Table S1** for *C1QTNF4* probe sequences). After multiple test correction (Bonferroni correction cutoff=1.33×10^{-5}), *C1QTNF4* (C1q and tumor necrosis factor related protein 4), expression levels were significantly associated with rs7124681 (located in the intronic region of *CELF1*; minor allele frequency [MAF]=40.46%, p = 8.63×10^{-8}; **Table 2** and **Figure 3B**), rs10838738 (located in the intronic region of *MTCH2*, MAF=35.35%, p=1.67×10^{-7}; **Table 2** and **Figure 3B**), and rs2290850 (located in the intronic region of *NUP160*, MAF=35.89%, p=2.65×10^{-6}; **Table 2** and **Figure 3B**) in cerebellum. After adjusting for the most-significant SNP rs7124681, no SNPs within the region remained significant. In order to determine whether these eQTL signals are independent of rs7937331, the strongest SNP within the region associated with CSF $A\beta_{42}$ levels, we performed conditional analysis adjusted for rs755553, a proxy for rs7937331 (D'=1, R^2=0.99). When conditioning on rs755553, no SNPs in the region remained significant (see **Figure 3D**). Linkage disequilibrium analyses suggest that rs7124681, rs10838738, rs2290850, and rs755553 are in tight LD (see **Figure 4**). Together, these results suggest that there is only one independent driver within this region and that the minor allele of the causal variant is associated with reduced *C1QTNF4* expression levels. Additionally, RegulomeDB predicts that rs10769258, rs4752993, and rs755553 are eQTLs for *C1QTNF4* in blood and cancer cells (see **Table 7 in Chapter 3**), which is in line with our cis-eQTL analyses for several brain regions. Regional plots for the cis-association with mRNA expression of other genes can be found in **Figures S2-S22**.

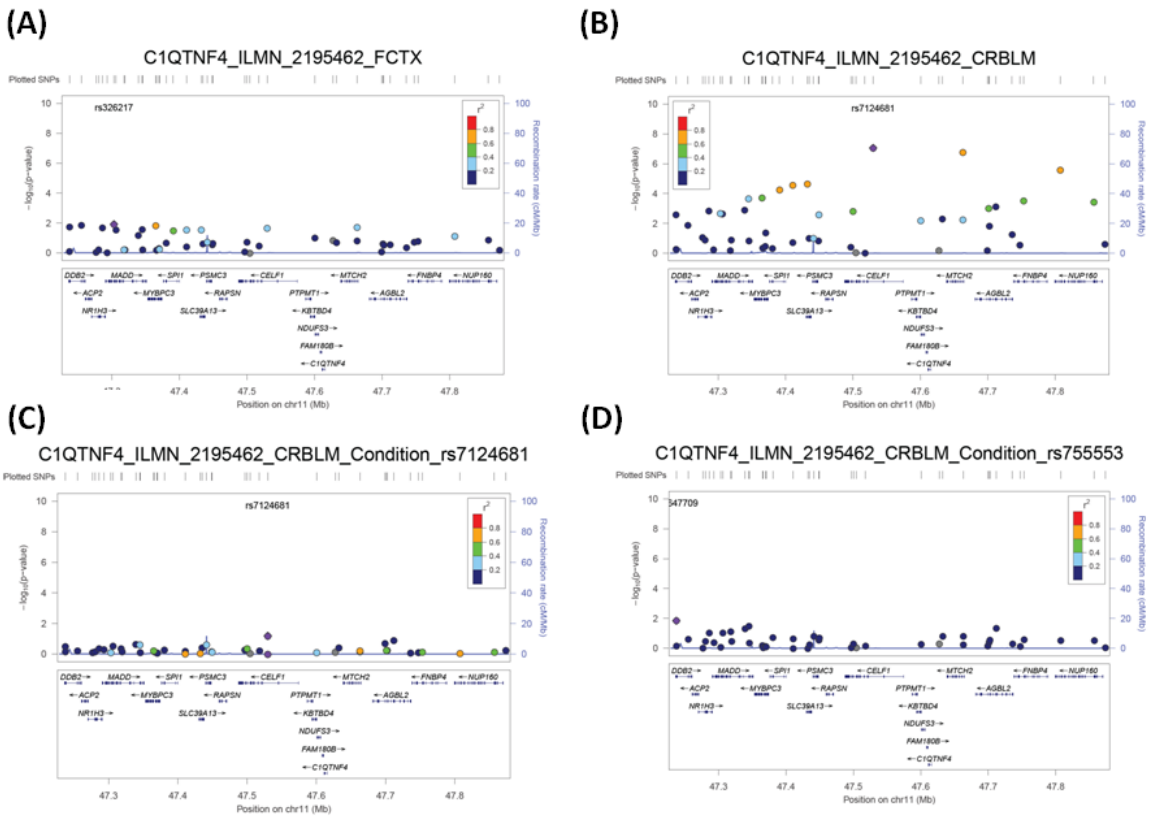


Figure 3: Cis association for *C1QTNF4* expression levels in (A) FCTX and (B) CRBLM. (C) and (D) represent cis association for *C1QTNF4* expression levels in CRBLM after conditioning on rs7124681 and rs755553, a proxy for the CSF top SNP in the *CELF1* fine-mapping region.

Table 2. SNP-Transcript pairs with a suggestive or significant association in GSE15745 and GSE36192

Gene	Probe ID	SNP	Chr:BP	MAF (%)	β_{FCTX}	P_{FCTX}	β_{CRBLM}	P_{CRBLM}	β_{TCTX}	P_{TCTX}	β_{PONS}	P_{PONS}
C1QTNF4	ILMN_2195462	rs7124681	11:47529947	40.5	-0.028	2.26×10^{-2}	-0.047	8.63×10^{-8}	NA	NA	NA	NA
C1QTNF4	ILMN_2195462	rs10838738	11:47663049	35.4	-0.029	1.98×10^{-2}	-0.047	1.67×10^{-7}	NA	NA	NA	NA
C1QTNF4	ILMN_2195462	rs2290850	11:47807774	35.9	-0.022	7.43×10^{-2}	-0.042	2.65×10^{-6}	NA	NA	NA	NA
C1QTNF4	ILMN_2195462	rs755553	11:47432303	32.5	-0.029	2.89×10^{-2}	-0.040	2.24×10^{-5}	NA	NA	NA	NA
C1QTNF4	ILMN_2195462	rs4752993	11:47410951	32.8	-0.029	2.95×10^{-2}	-0.040	2.80×10^{-5}	NA	NA	NA	NA
C1QTNF4	ILMN_2195462	rs10769258	11:47391039	32.9	-0.028	3.36×10^{-2}	-0.038	5.42×10^{-5}	NA	NA	NA	NA

Association between mRNA transcript and SNP pair with significant evidence within at least one brain region using the GSE15745 and GSE36192 datasets. The multiple-testing significant threshold is 1.33×10^{-5}. A total of 330 FCTX, 330 CRBLM, 142 PONS and 144 TCTX were used for analyses.

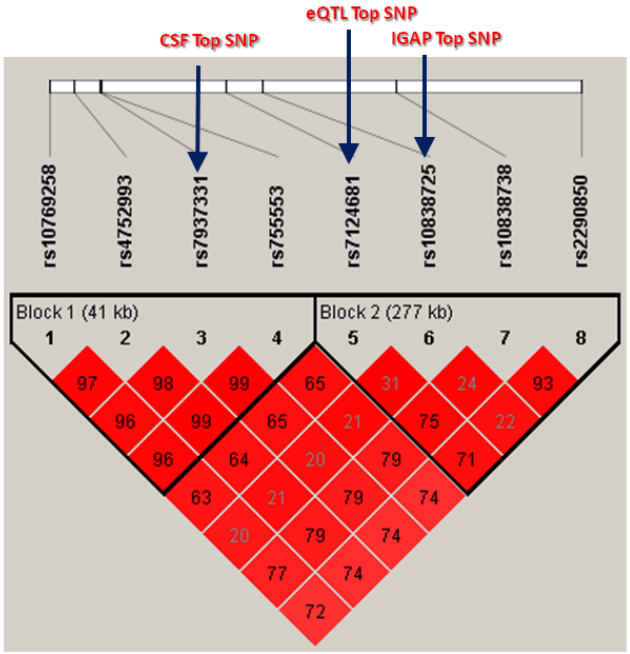


Figure 4: Linkage disequilibrium (LD) analyses for potential eQTLs, CSF top SNP, and IGAP top SNP in the *CELF1* region. Six variants were found to be suggestively ($p < 10^{-4}$) or significantly associated with the *C1QTNF4* transcript. D' which ranges from 0 to 1 (white to red) represents the LD between SNPs and numbers indicate R^2 between SNPs.

4.4.3 Alzheimer's disease status is significantly associated with *C1QTNF4* expression level

To determine whether Alzheimer's disease status affects *C1QTNF4* expression levels, we used the GSE5281 dataset to compare *C1QTNF4* expression levels between 47 AD brains and 14 normal brains in 6 brain regions. In the joint analysis, we conducted multivariate linear regression adjusted for disease status, brain region, age, and gender. Our results showed that disease status was significantly associated with reduced *C1QTNF4* expression levels (β=-0.73, $p=1.76\times10^{-13}$, adjusted R^2=28.91%; **Table 3**). When we analyzed each brain region separately using a two-tailed unpaired *t*-test, we found that the *C1QTNF4* expression levels were significantly lower in AD cases compared to controls in all tested brain regions ($p=8.82\times10^{-4}$ in entorhinal cortex, $p=6.73\times10^{-2}$ in hippocampus, $p=1.08\times10^{-2}$ in middle temporal gyrus, $p=1.17\times10^{-2}$ in posterior cingulate cortex, $p=3.03\times10^{-2}$ in superior frontal gyrus; $p=1.15\times10^{-2}$ in visual cortex; **Table 4**). Together, these results suggest that reduced *C1QTNF4* mRNA expression levels are associated with AD disease status.

Table 3. Association between AD disease status and *C1QTNF4* transcripts in the joint analysis in GSE5281

Probe-Symbol	Probe-ID	β (SE)	P	Adjusted R^2 (%)
C1QTNF4	223708_at	-0.73 (0.09)	1.76×10^{-13}	28.91

Multivariate linear regression was performed to estimate the association with *C1QTNF4* mRNA expression levels adjusting for brain region, age and gender. Brian region, age, and gender covariates were coded as categorical variables. The effect size, standard error, p values, and adjusted R-squared values were reported. SE: Standard error.

Table 4. Association between AD disease status and *C1QTNF4* expression levels in six brain regions in GSE5281

	*C1QTNF4*_223708_at		
	AD cases	Controls	P
Entorhinal cortex (EC)	24.34	521.49	8.82×10^{-4}
Hippocampus (HIP)	62.75	141.84	6.73×10^{-2}
Middle temporal gyrus (MTG)	56.69	326.36	1.08×10^{-2}
Posterior cingulate cortex (PCC)	33.07	103.24	1.17×10^{-2}
Superior frontal gyrus (SFG)	64.85	342.69	3.03×10^{-2}
Visual cortex (VC)	98.08	191.37	1.15×10^{-2}

A two-tailed unpaired t test was performed to evaluate mRNA expression levels in AD cases versus controls. Average expression levels in AD cases and controls, and p values were reported.

4.4.4 Replication using the GSE15222 dataset

In order to confirm the observations from the GSE15745 and GSE36192 datasets we analyzed *C1QTNF4* expression in the GSE15222 dataset for replication. The GSE15222 dataset contains genotype and expression data from both AD cases (N=176) and controls (N=188) and we found 36 SNPs within the defined genomic region. The control samples were composed of 40 (21.3%) frontal cortex, 3 (1.6%) parietal cortex, 136 (72.3%) temporal cortex, and 9 (4.8%) cerebellar tissues. In the control-only analyses, a similar two-step regression approach was used to examine the cis-eQTL effects on *C1QTNF4* expression levels for each brain region. In the GSE15222 dataset, we did not find the top six SNPs which were associated with *C1QTNF4* expression levels in the GSE15745 and GSE36192 datasets. The most significant SNP, rs7102372, located in an intronic region of *CELF1*, showed evidence of association with *C1QTNF4* expression levels in the TCTX (MAF=15.73%, $p=6.44\times10^{-4}$; **Table 5** and **Figure 5A)**. However, we did not find any significant association of rs7102372 with *C1QTNF4* expression levels in FCTX, CRBLM, and parietal cortex which may due to the small sample sizes (N=40, 9, and 3 for FCTX, CRBLM, and parietal cortex respectively; **Figure 5B-5D**). Moreover, rs7102372 was not associated with CSF $A\beta_{42}$ levels ($p=0.62$) and AD risk (IGAP $p=0.83$). Overall, these results suggest that rs7102372 may be a false positive signal. We next investigated whether there are significant differences in the *C1QTNF4* expression levels between AD cases and controls for each brain region (71 frontal cortex, 20 parietal cortex, 242 temporal cortex, and 31 cerebellar tissues). We found that *C1QTNF4* expression levels were significantly higher in controls compared to AD cases ($p=8.82\times10^{-16}$ in frontal cortex, $p=6.07\times10^{-8}$ in parietal cortex, $p=1.67\times10^{-36}$ in temporal cortex, $p=9.69\times10^{-14}$ in cerebellar tissues; **Table 5**), consistent with our findings in the GSE15745+GSE36192 datasets (see **Table 4**).

4.5 DISCUSSION

TREM2 is an immune phagocytic receptor expressed on brain microglia and is known to trigger phagocytosis and regulate the inflammatory response. AD cases that carry *TREM2* risk variants have a significantly higher rate of brain atrophy than those who do not carry *TREM2* risk variants[132]. A recent study showed that *TREM2* is a substrate of γ-secretase upon removal of ectodomain[175]. These studies suggest that TREM2 may play a role in AD. Taken together with our genetic findings that some rare coding variants are more frequently identified in AD cases than in controls, these findings suggest some of these rare coding variants may alter TREM2 function.

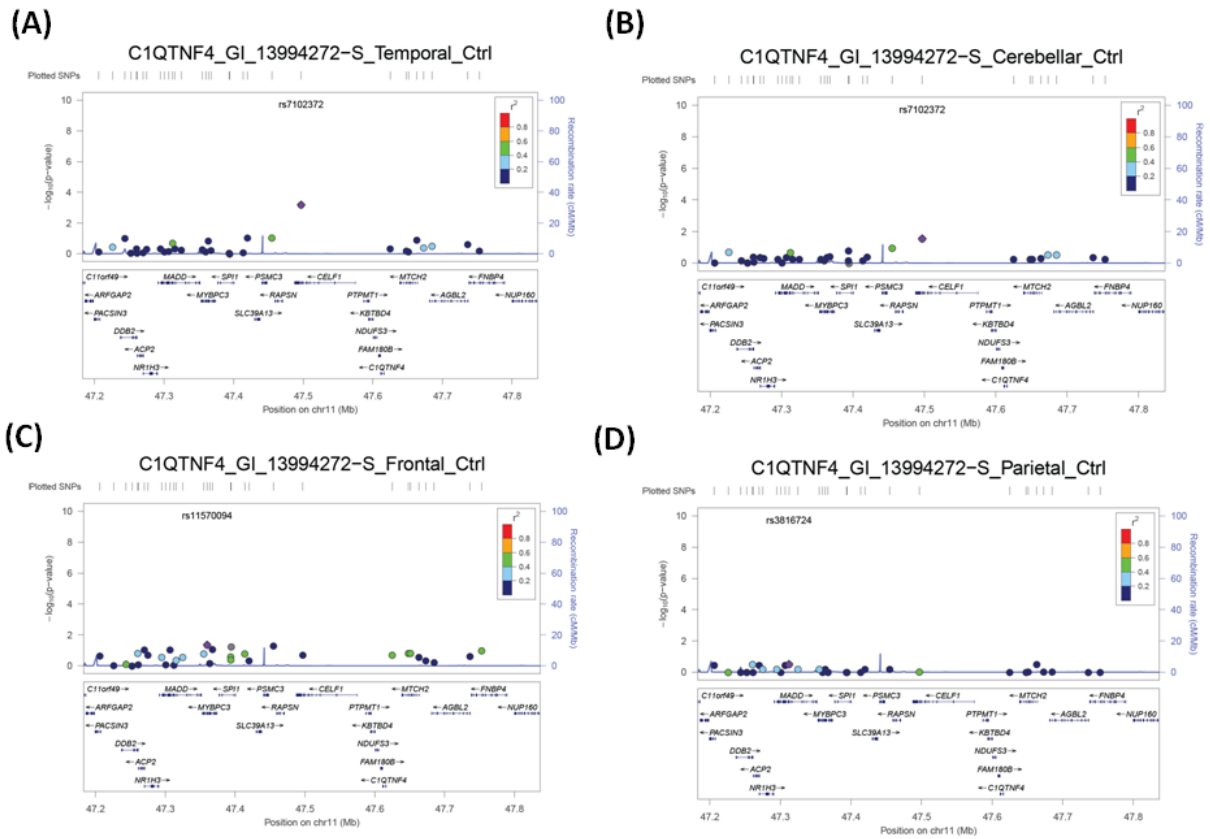


Figure 5: Cis association for *C1QTNF4* expression levels in (A) TCTX and (B) CRBLM. (C) FCTX, and (D) Parietal cortex in the GSE15222 dataset. A total of 40 FCTX, 9 CRBLM, 3 parietal cortex and 135 TCTX from control samples were used for analysis.

Table 5. Transcript- SNP Pairs with a suggestive or significant association based on the GSE15222 dataset

Probe-Symbol	Probe-ID	SNP	Chr:BP	MAF (%)	Function	Gene	β_{FCTX}	P_{FCTX}	β_{CRBLM}	P_{CRBLM}	β_{TCTX}	P_{TCTX}	$\beta_{Parietal}$	$P_{Parietal}$
C1QTNF4	GI_13994272-S	rs7102372	11:47496827	15.73	Intron	*CELF1*	0.16	2.05×10^{-1}	-0.38	2.84×10^{-2}	0.36	6.44×10^{-4}	0.01	9.76×10^{-1}

Most significant SNP associated with a probe tagging *C1QTNF4* transcript in the GSE15222 dataset. A total of 40 FCTX, 9 CRBLM, 3 parietal cortex and 135 TCTX from control samples were used for analysis. The six top SNPs identified in GSE15745 and GSE36192 datasets were not found in the GSE15222 dataset and thus were not included in the table.

Table 6. Association between AD disease status and *C1QTNF4* expression levels in four brain regions in GSE15222

	*C1QTNF4*_GI_13994272-S		
	AD cases	Controls	P
Frontal cortex (N = 71)	473.56	535.96	8.82×10^{-16}
Parietal cortex (N = 20)	472.86	641.83	6.07×10^{-8}
Temporal cortex (N = 242)	437.52	615.72	1.67×10^{-36}
Cerebellar tissues (N = 31)	485.4	491.94	9.69×10^{-14}

A two-tailed unpaired *t*-test was performed to evaluate mRNA expression levels in AD cases versus controls. Average expression levels in AD cases and controls, and p-values were reported.

In this chapter, we sought to assess the effects of novel *TREM2* risk variants on TREM2 function by transducing cells with *TREM2* risk variants and measuring cell surface expression of TREM2. Our preliminary data demonstrate that the novel risk variant p.R136W results in reduced detection of TREM2 at the cell surface, similar to the causative Nasu-Hakola mutation p.T66M. We hypothesize that both the risk variant and the causative mutation could impair TREM2 function by preventing its transport to the cell surface. TREM2 trafficking to the cell surface is important since if TREM2 does not get to the cell surface, TREM2 cannot bind to its endogenous ligand and thus cannot transduce its intracellular signal through DAP12 properly. We cannot yet rule out the possibility that p.R136W and p.T66M disrupt transcription or translation of TREM2. Since p.R136W is very close to the transmembrane domain of TREM2, it is likely that p.R136W affects folding of TREM2. Follow-up studies are on-going to distinguish between expression and trafficking defects by measuring TREM2 and DAP12 RNA and protein levels inside the cell. Surprisingly, there are no obvious differences in TREM2 cell surface expression for previously reported AD risk factors *TREM2* p.R47H and p.R62H compared to WT in our first run and following replication studies. We hypothesize that p.R47H and p.R62H may affect other mechanisms of TREM2 function such as ligand binding. One caveat to the current study is that TREM2 variants may occur in regions of the TREM2 protein that would affect the binding affinity of anti-TREM2 antibody used to measure cell surface TREM2.

Another important goal of this chapter is to identify causal variants and to generate testable hypotheses associated with GWAS-identified variants and genes. To date, the majority of these GWAS-identified AD risk variants are non-coding variants with very small effect size[8,10,11,78]; therefore, the functional impact and mechanism associated with these variants and genes remained largely unclear. In chapter 3, we analyzed our CSF datasets and found that most of these disease-associated variants were not necessarily associated with CSF biomarker levels. Several recent studies have suggested that GWAS-identified disease-associated variants may perturb mRNA expression levels of neighboring genes as eQTLs[160,167]. In chapter 3, several SNPs within the *CELF1* fine-mapping region were predicted to regulate expression levels of neighboring genes based on RegulomeDB database. To identify genetic determinants of gene expression within the *CELF1* fine-mapping region in human brains, we performed comprehensive cis-eQTL analyses using four publicly available microarray datasets with gene expression in human brain tissues. After analyzing GSE15745 and GSE36192 datasets, we found that all of the expression-associated SNPs, including rs7124681, rs10838738, rs2290850, and rs755553, are in tight LD with rs7937331, the top SNP in the *CELF1* fine-mapping region associated with CSF $A\beta_{42}$ levels. Conditional analyses suggested that

there is only one independent driver within this region and that the minor allele of the underlying causal variant mediates both *C1QTNF4* expression levels. Moreover, RegulomeDB predictions for these expression-associated SNPs, which are based on expression data in blood and cancer tissues, are consistent with our cis-eQTL analyses, demonstrating that coincident eQTLs are present between several human tissues.

Next, we analyzed the GSE5281 dataset and found that AD disease status was significantly associated with reduced *C1QTNF4* mRNA expression levels ($p=1.76\times10^{-13}$), which indicates that *C1QTNF4* is differentially expressed between AD brains and control brains. Finally, we analyzed the GSE15222 dataset in order to replicate our findings for the *C1QTNF4* transcript. We did not identify significant association for any SNP-*C1QTNF4* transcript pair in the control-only analysis; however, *C1QTNF4* expression levels were significantly higher in controls than in AD cases, which is in line with the results in the GSE5281 dataset. The possible explanation for the lack of association in the control-only analysis is the small sample size in each brain region except TCTX in the GSE5281 dataset. Additionally, the differences may result from differences in the datasets (laser-capture neurons [GSE 5281] versus brain homogenates which contain many cell types that do not control for the heterogeneity in cell composition between AD cases and controls [GSE15745, GSE36192 and GSE15222]). In a previous eQTL-analysis which consisted of 733 brain samples from the cerebellum and temporal cortex of about 200 AD cases and about 200 non-AD individuals, a genome-wide significant association of rs10838738 with expression of *C1QTNF4* ($\beta=-0.18$, $p=1.48\times10^{-9}$) was reported in cerebellum even though the association was not significant in temporal cortex ($\beta=-0.02$, $p=0.66$)[176]. The direction of effects in the cerebellum is the same as that in GSE15745 and GSE36192 datasets. When we combined the p value[176] for the rs1083878/*C1QTNF4* association with that estimated in GSE15745 and GSE36192 datasets using the Fisher's method[177], the meta-analysis p-value was 9.1×10^{-15} in cerebellum, which strongly indicates that the observed association is authentic. Although the cerebellum is rarely affected in AD brains[178], recent studies have indicated a significant overlap in cis-eQTL association between different brain regions and different tissue types[176,179]. These studies also showed that expression patterns are more likely to be similar between two brain regions than two different tissue types[176,179]. Additionally, pathological changes have been detected in the AD cerebellum[180]. However, we cannot rule out the possibilities that eQTLs are only specific to certain brain regions and that eQTLs that display similar effects in different brain regions are not associated with human diseases. Additional studies need to determine whether disease-associated eQTLs regulate

gene expression regardless of brain regions or tissue types. Overall, these findings suggest that eQTL-based analyses can reveal potential functional genes and variants and provide a hypothesis as to their functions.

The *C1QTNF4* gene encodes a protein which triggers NF-kappa-B activity[181]. A previous study has shown that over-expression of *C1QTNF4* induces the activation of NF-kappa-B, and IL6/STAT3signaling pathways and may be a tumor-promoting regulator of inflammation[181]. *TREM2* expression was recently shown to be regulated by an NF-kappa-B-sensitive miRNA[182], which suggests that *C1QTNF4* may be involved in the same pathway as *TREM2*. Our mRNA expression analyses suggested that *C1QTNF4* expression levels are significantly lower in AD cases. We also showed that expression-associated SNPs (rs7124681, rs10838738, rs2290850) within this region have a complete LD with the CSF $A\beta_{42}$ top SNP (rs7937331), which suggests that they tag the same signal. Like *TREM2*, *C1QTNF4* may be involved in clearance of Aβ aggregates or in neuroinflammation.

In conclusion, our cell-surface expression experiments found that *TREM2* p.T66M and p.R136W reduce cell surface expression of TREM2, while p.R47H and p.R62H, two confirmed risk factors for AD, have no robust impact on TREM2 cell surface expression. Our cis-eQTL analyses identified significant polymorphisms regulating *C1QTNF4* expression levels. We also found evidence of differential expression in *C1QTNF4* transcript between AD cases and controls. These results also illustrated that the underlying causal variants and genes may not be the gene originally identified in GWAS studies and that genes involved in the inflammatory response play an important role in AD pathogenesis.

SUPPLEMENTARY TABLES

Table S1. Detailed information for probes of interest

Probe-ID	Sequence	Start site	End site
C1QTNF4-ILMN_2195462	ACCGAGTTCGTCAACATTGGCGGCGACTTCGACGCGGCGGCCGGCGTGTT	11:47611701	11:47611750

SUPPLEMENTAL FIGURES

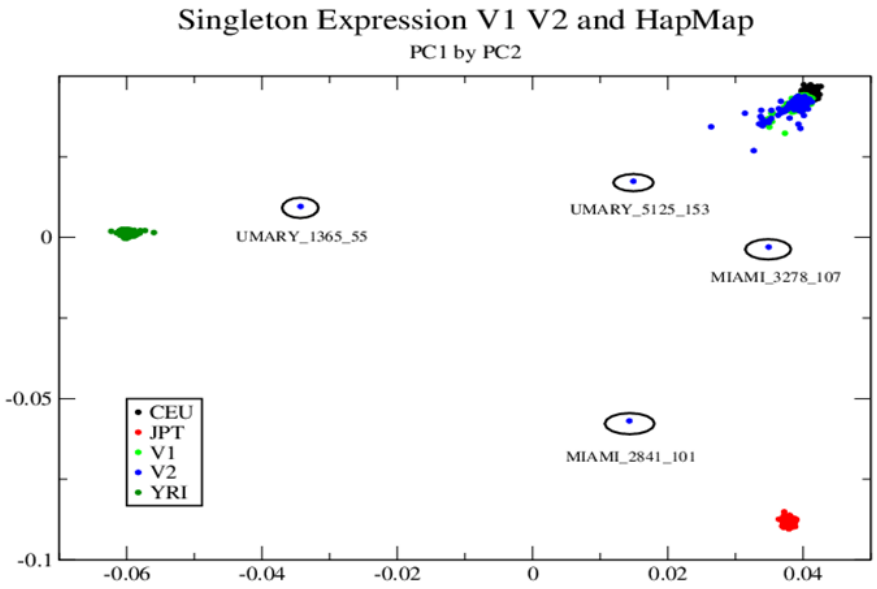
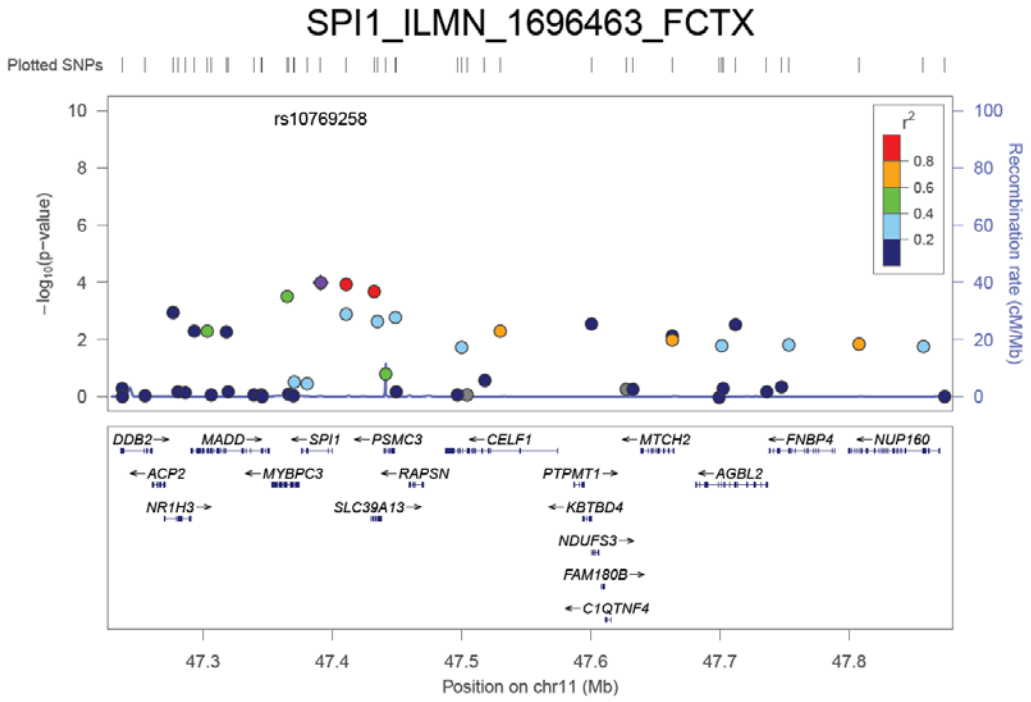


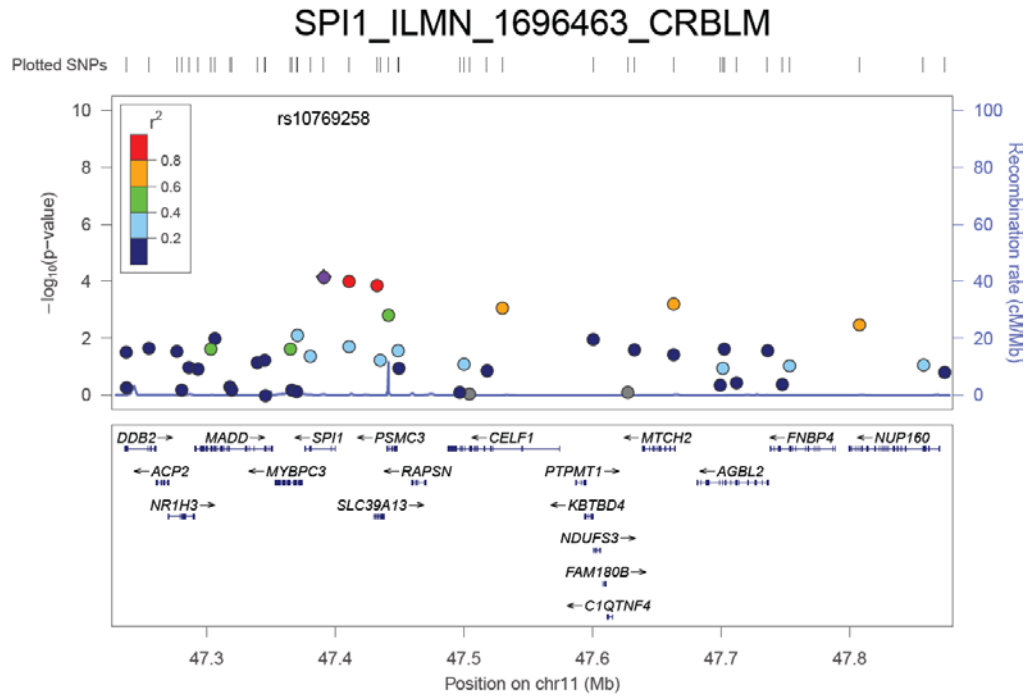
Figure S1: Principal component analyses for the GSE15745 + GSE36192 datasets (N = 382). HapMap samples (CEU: CEPH Europeans from Utah; JPT: Japanese in Tokyo; YRI: Yoruba in Ibadan, Nigeria) were included in the analyses as reference populations. V1 and V2 represent samples in the GSE15745 and GSE36192 respectively. Four outliers circled in black were excluded from final analyses.

Figure S2. Cis association for *SPI1*-ILMN_1696463 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

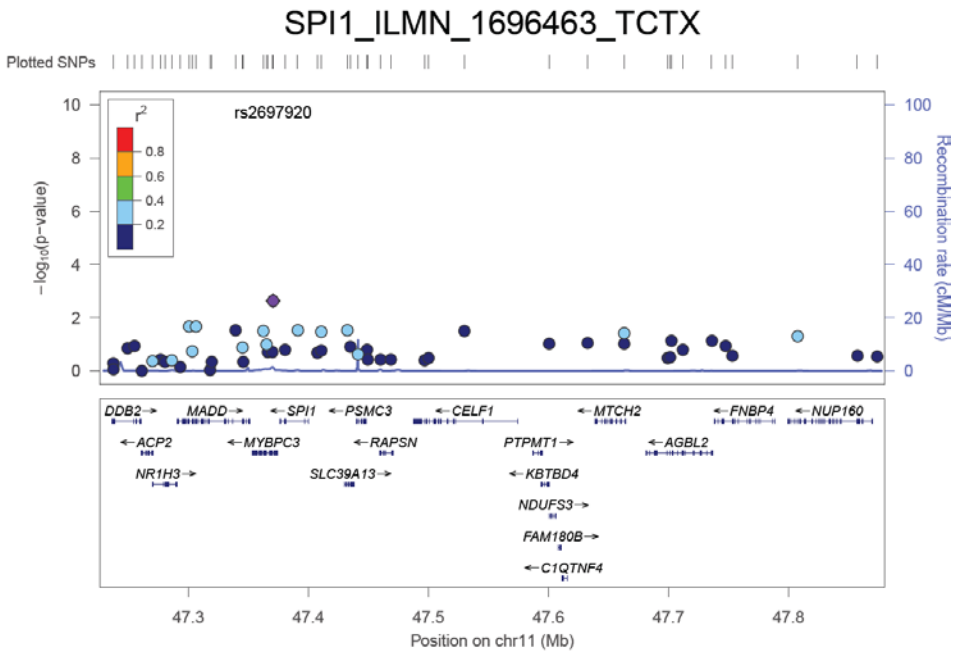
(A)



(B)



(C)


SPI1_ILMN_1696463_TCTX
Plotted SNPs
rs2697920
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

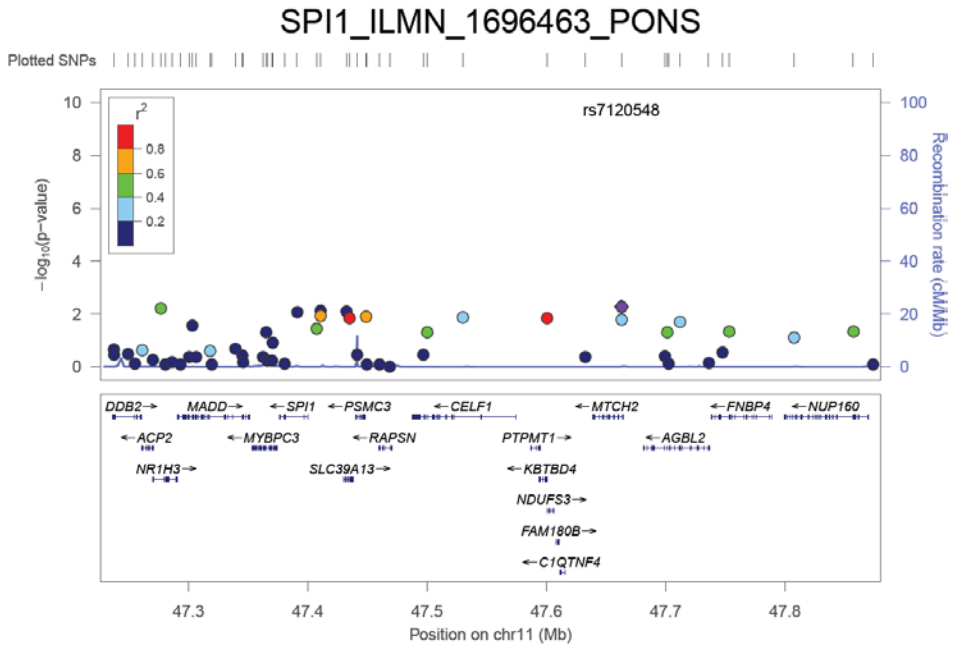
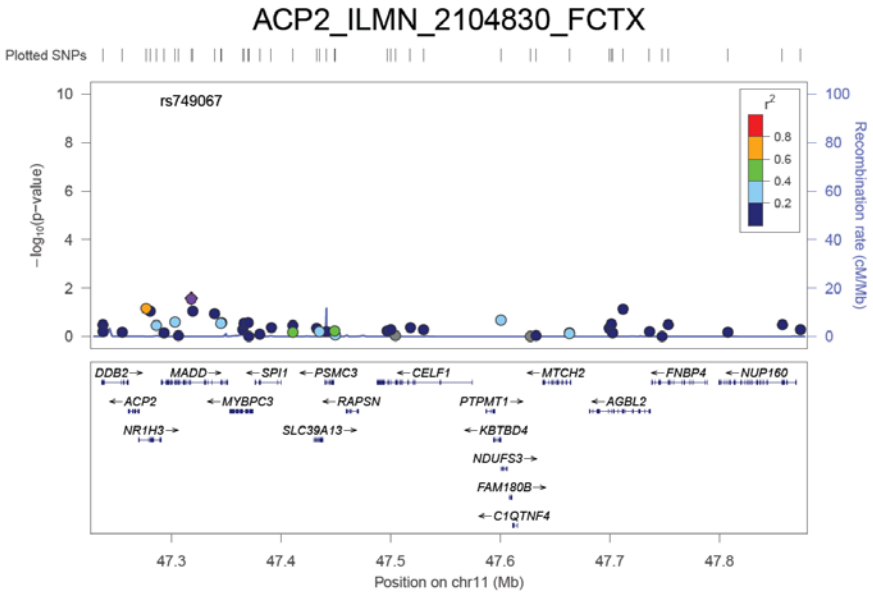
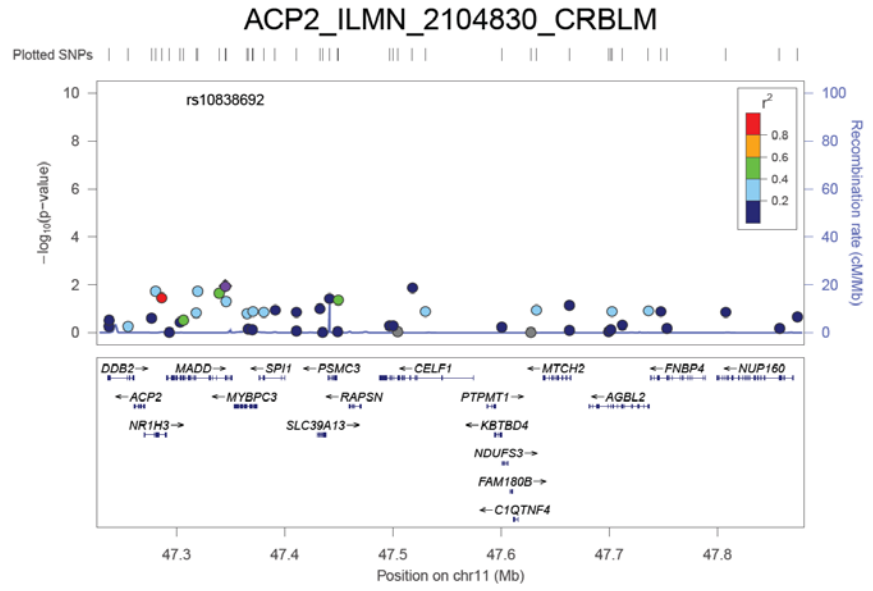

SPI1_ILMN_1696463_PONS
Plotted SNPs
rs7120548
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S3. Cis association for *ACP2*-ILMN_2104830 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

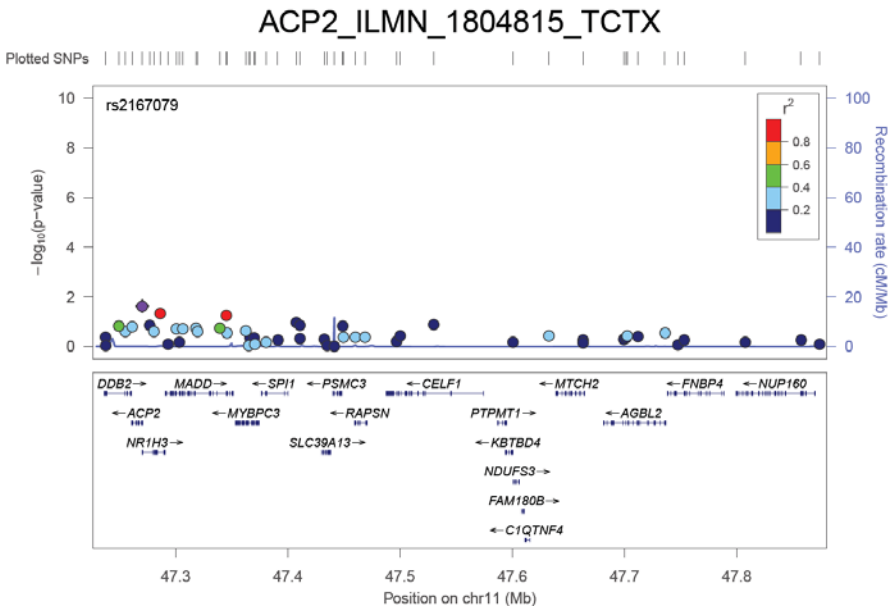
(A)



(B)



(C)


ACP2_ILMN_1804815_TCTX
Plotted SNPs
rs2167079
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
DDB2 MADD SPI1 PSMC3 CELF1 MTCH2 FNBP4 NUP160
ACP2 MYBPC3 RAPSN PTPMT1 AGBL2
NR1H3 SLC39A13 KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3 47.4 47.5 47.6 47.7 47.8
Position on chr11 (Mb)

(D)

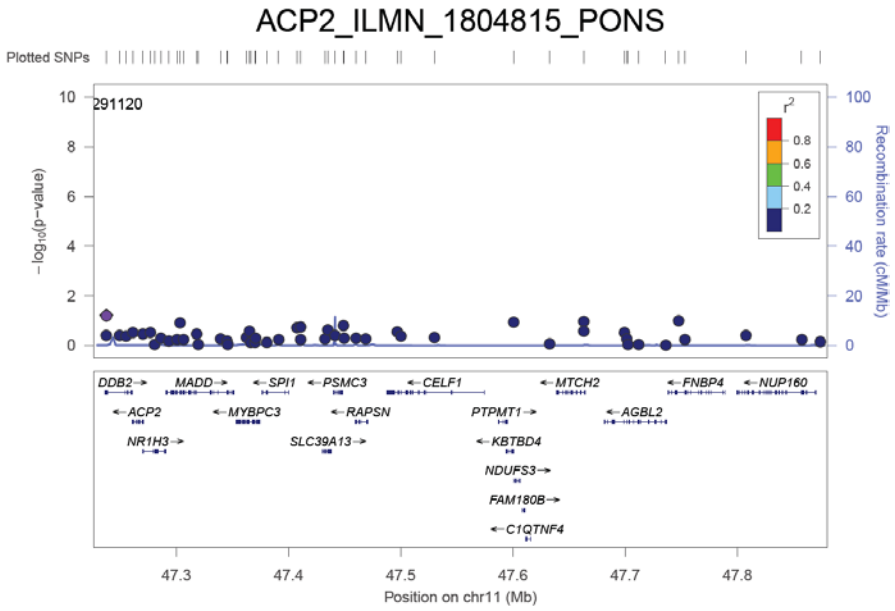
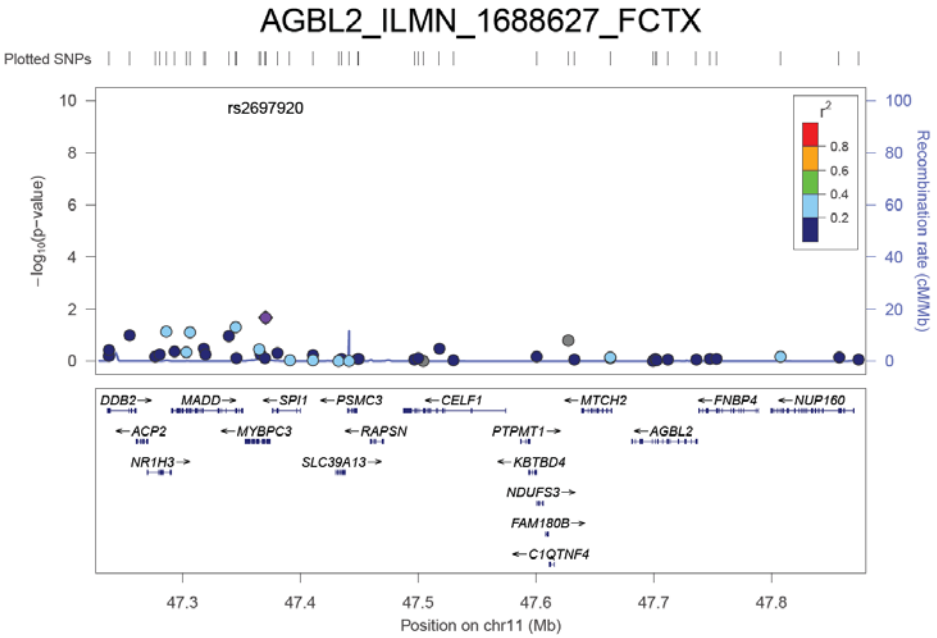
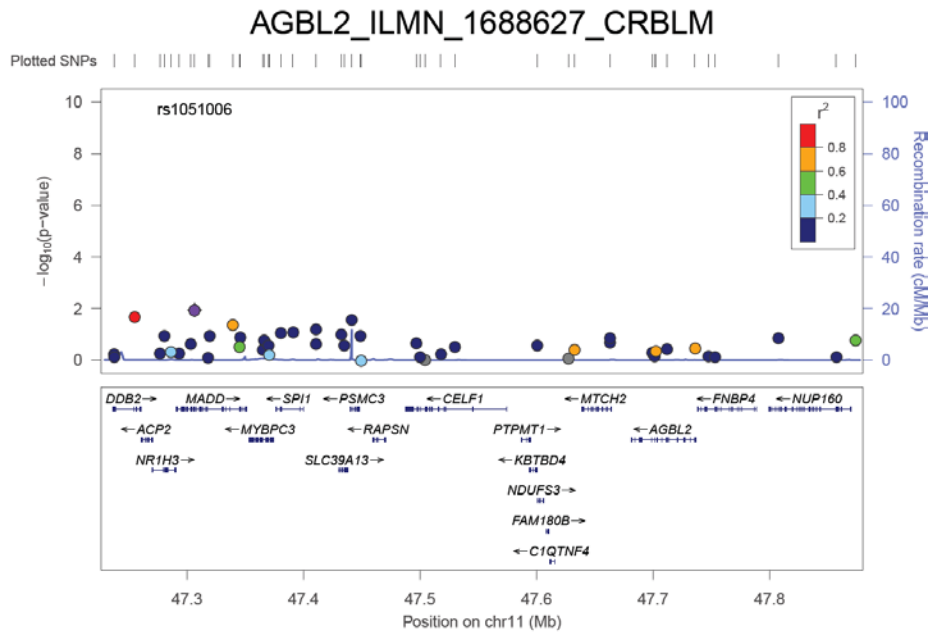

ACP2_ILMN_1804815_PONS
Plotted SNPs
291120
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
DDB2 MADD SPI1 PSMC3 CELF1 MTCH2 FNBP4 NUP160
ACP2 MYBPC3 RAPSN PTPMT1 AGBL2
NR1H3 SLC39A13 KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3 47.4 47.5 47.6 47.7 47.8
Position on chr11 (Mb)

Figure S4. Cis association for *AGBL2*-ILMN_1688627 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

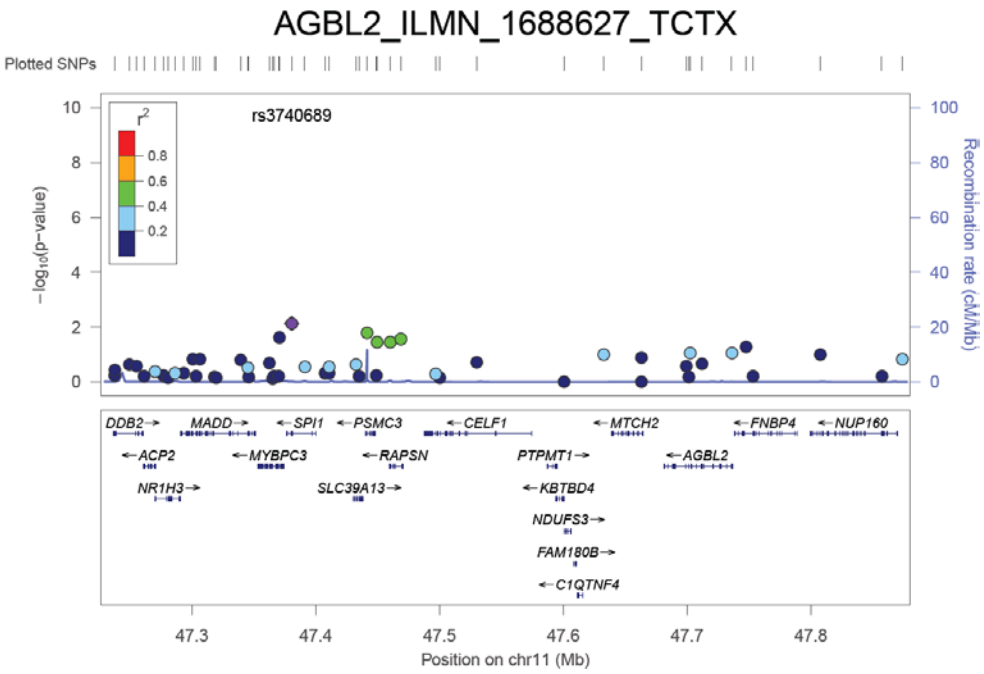
(A)



(B)



(C)


AGBL2_ILMN_1688627_TCTX
Plotted SNPs
rs3740689
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

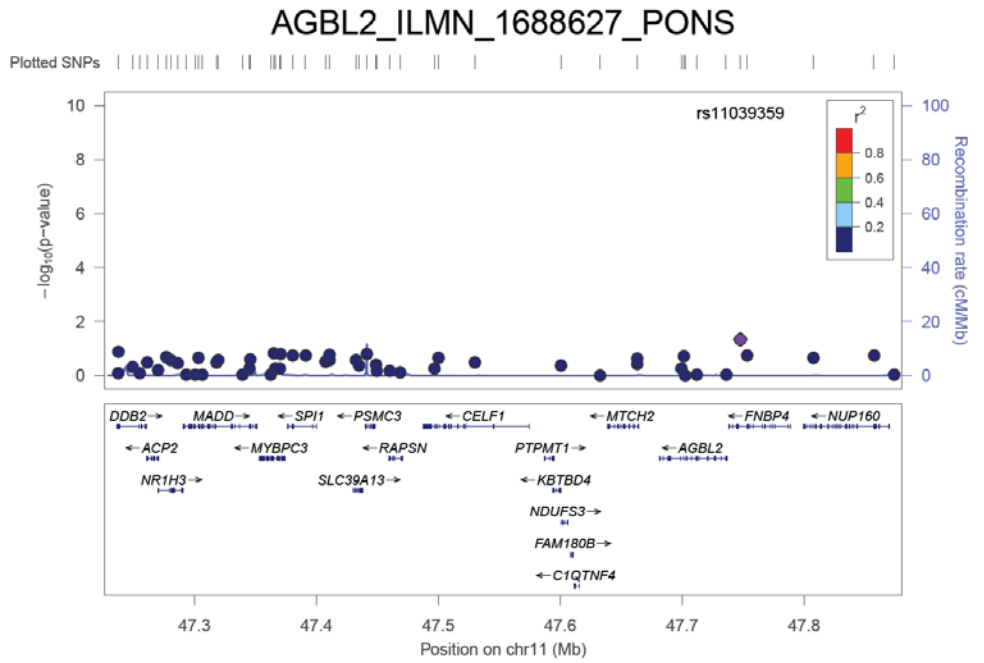
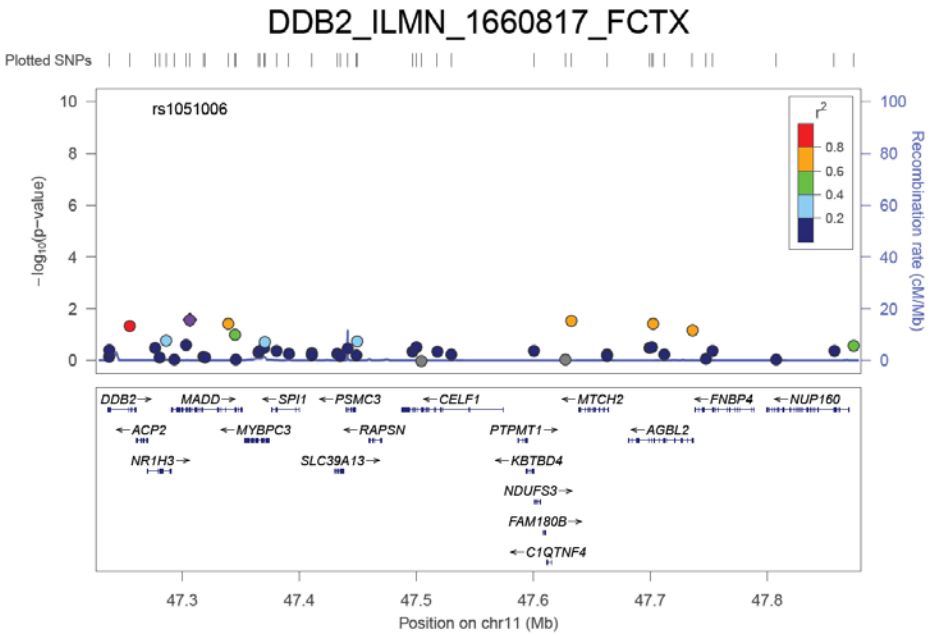
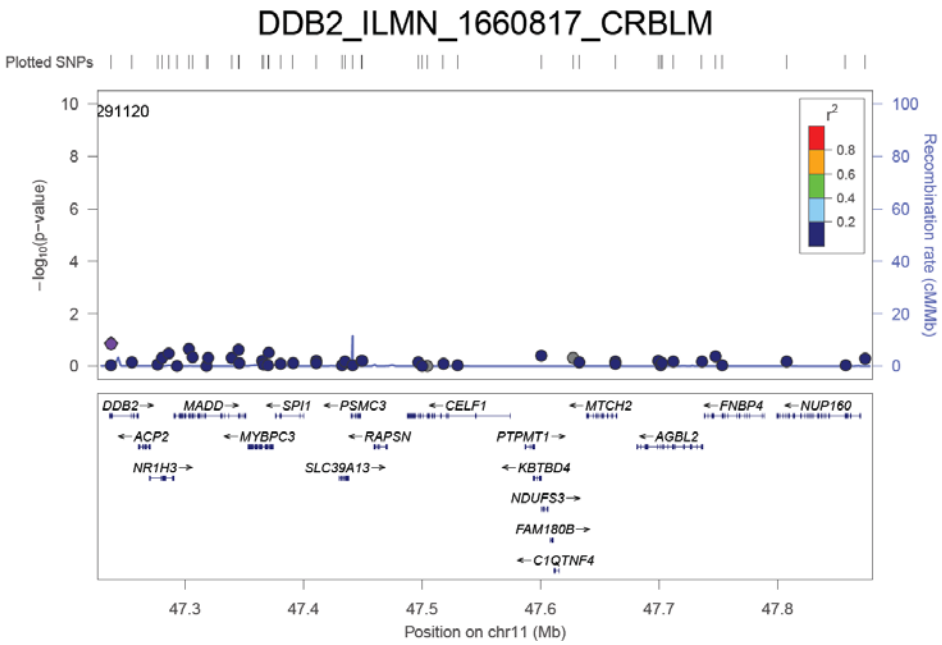

AGBL2_ILMN_1688627_PONS
Plotted SNPs
rs11039359
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S5. Cis association for *DDB2*-ILMN_1660817 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

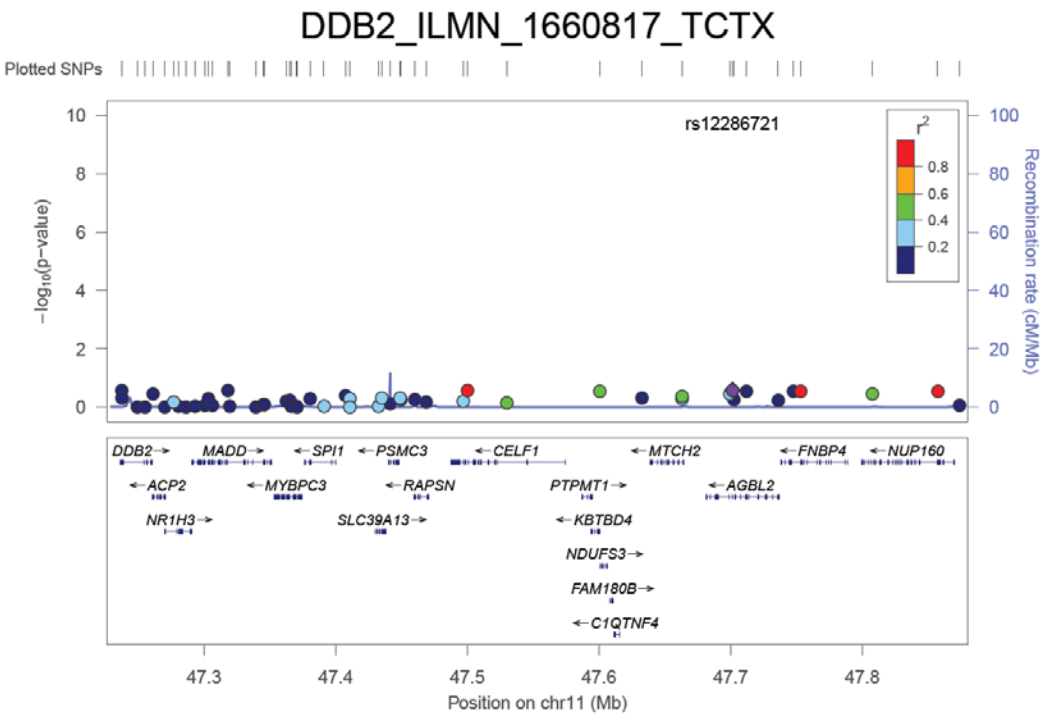
(A)



(B)



(C)


DDB2_ILMN_1660817_TCTX
Plotted SNPs
rs12286721
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

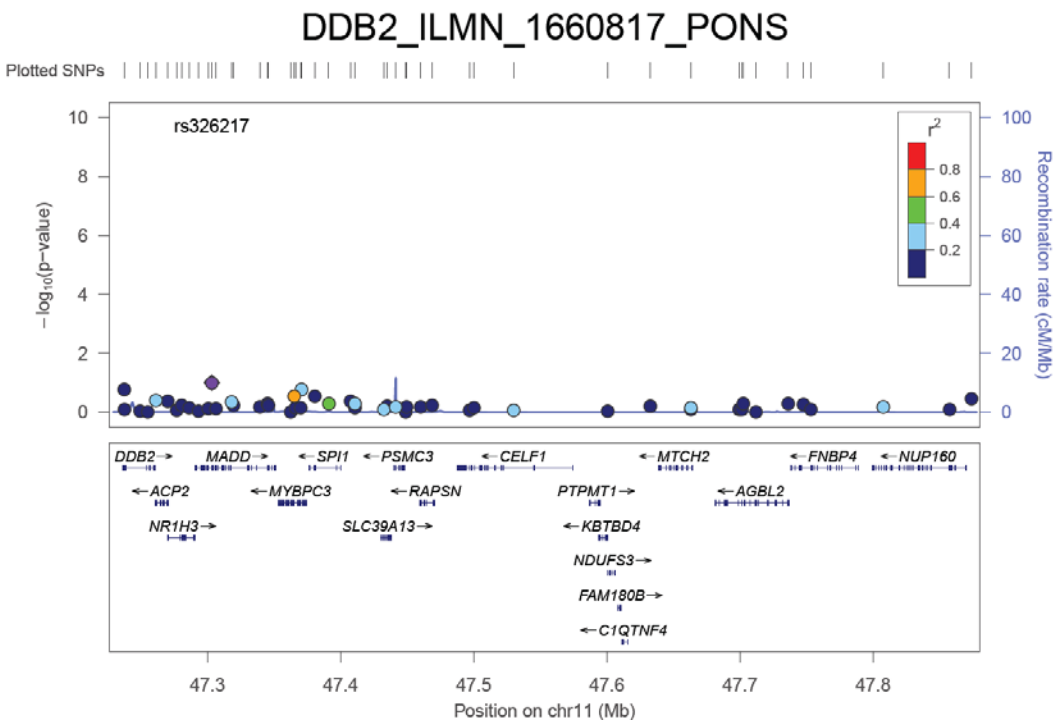
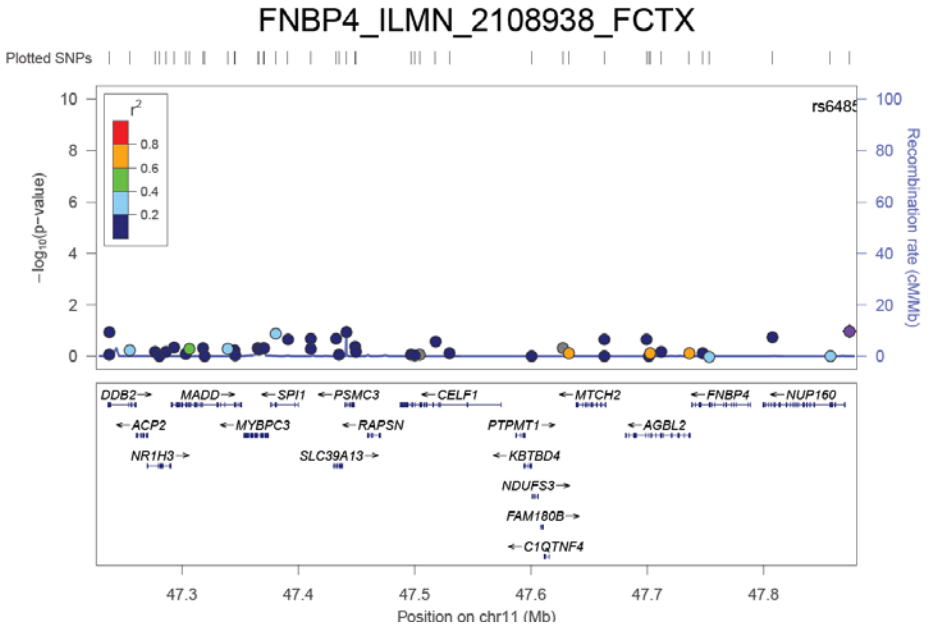
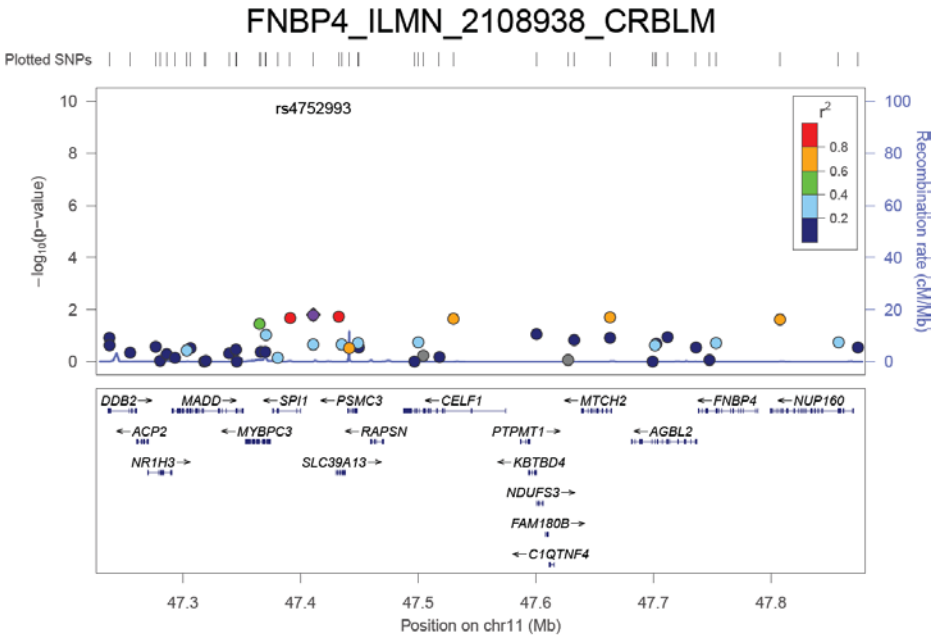

DDB2_ILMN_1660817_PONS
Plotted SNPs
rs326217
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S6. Cis association for *FNBP4*-ILMN_2108938 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

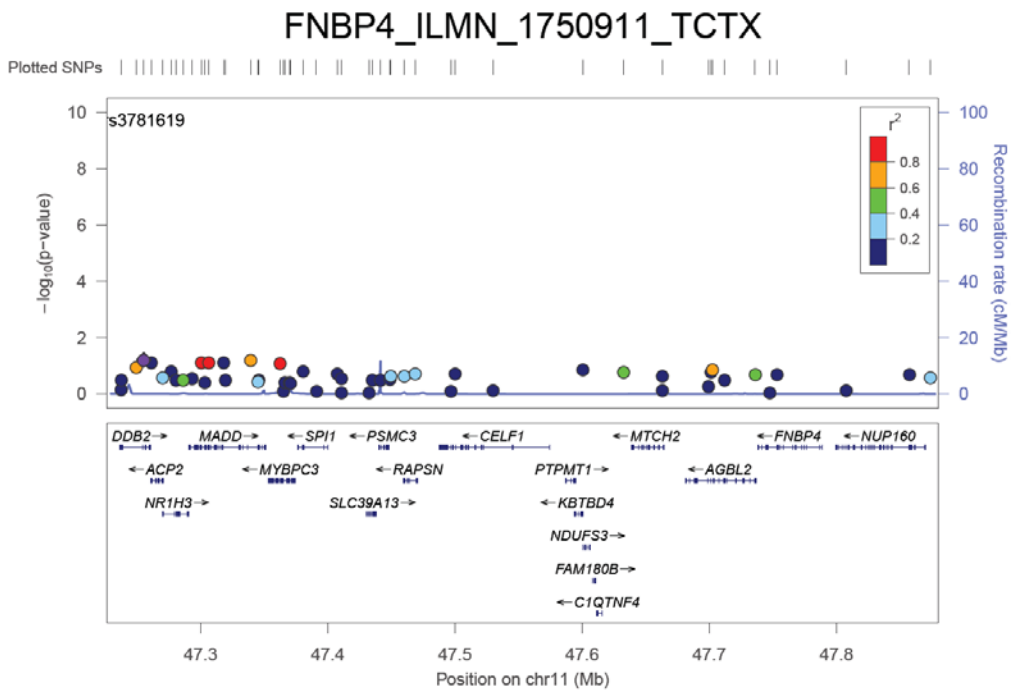
(A)



(B)



(C)


FNBP4_ILMN_1750911_TCTX
Plotted SNPs
s3781619
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

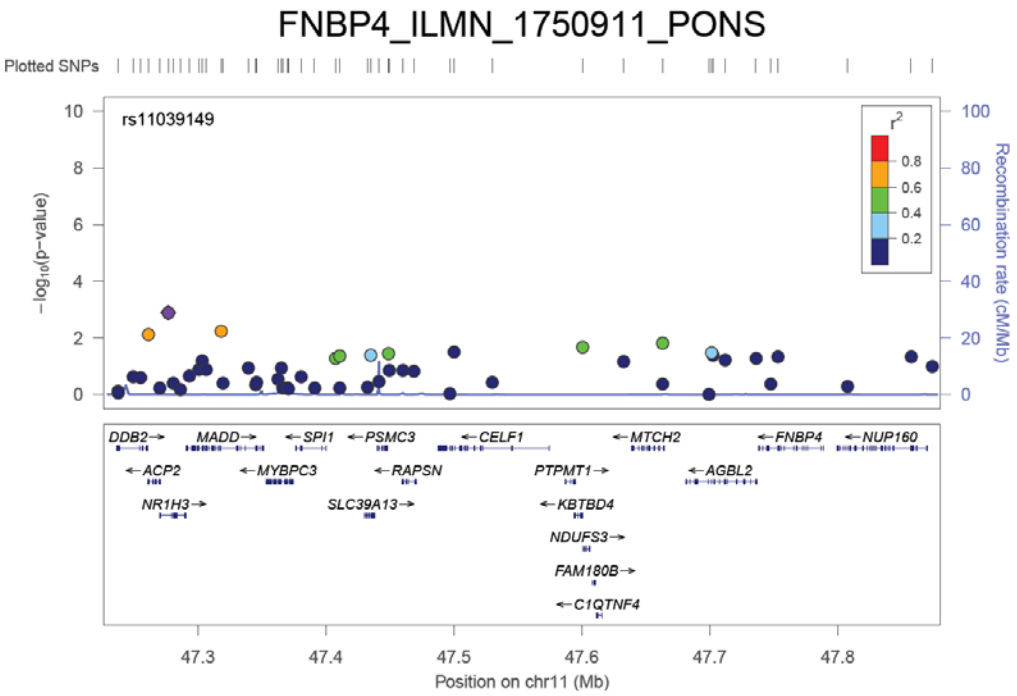
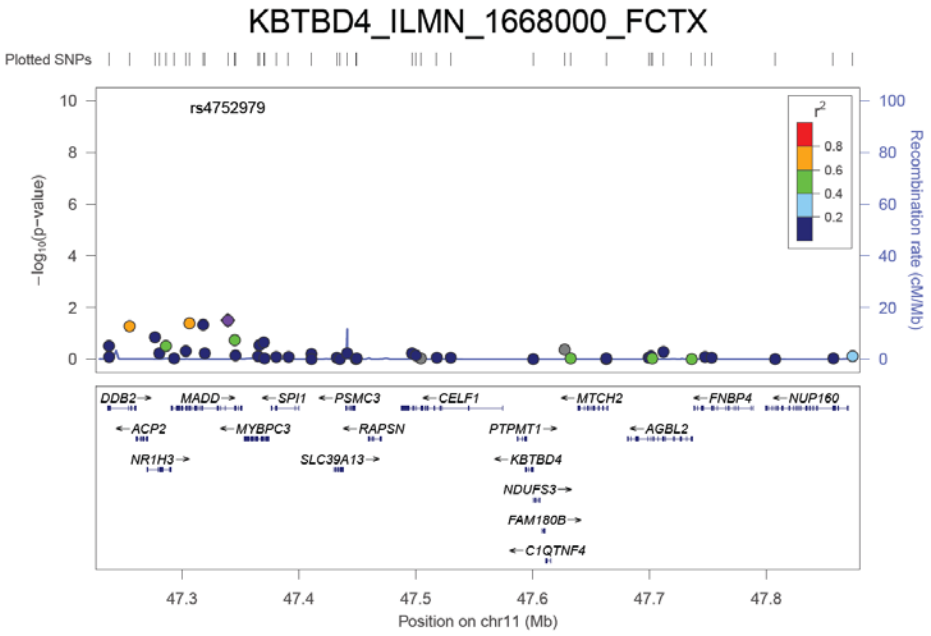
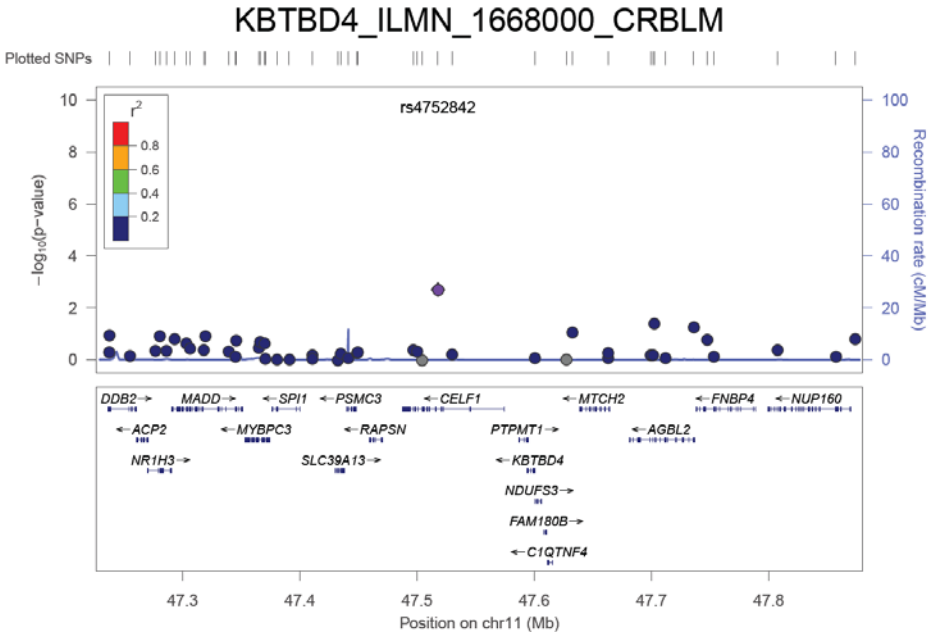

FNBP4_ILMN_1750911_PONS
Plotted SNPs
rs11039149
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S7. Cis association for *KBTBD4*-ILMN_1668000 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

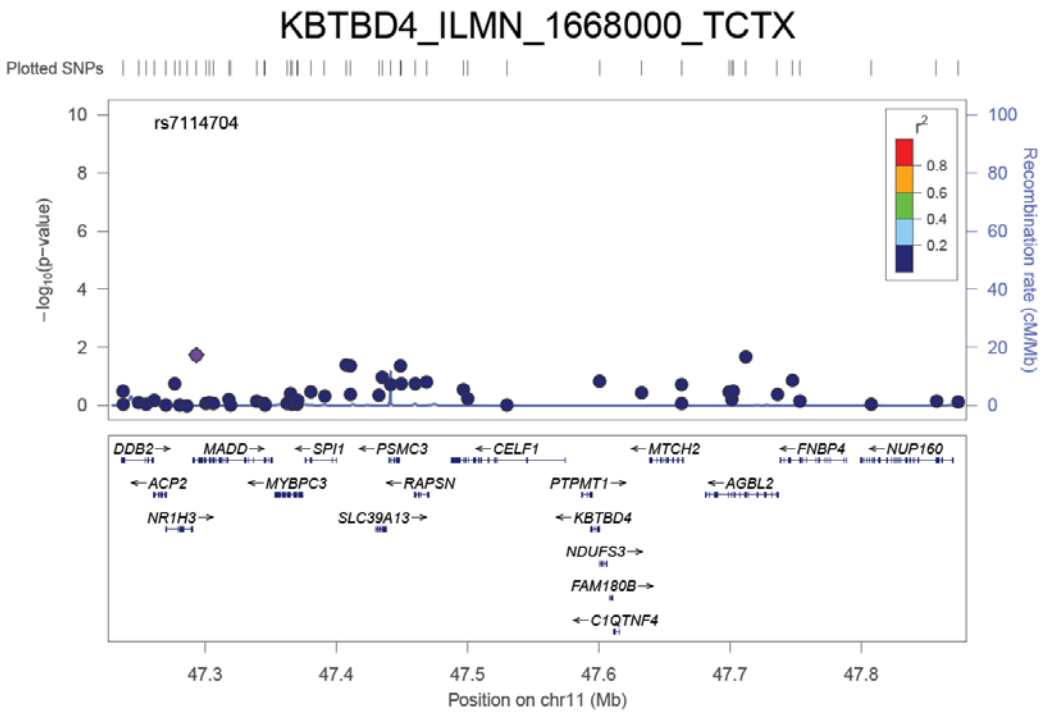
(A)



(B)



(C)


KBTBD4_ILMN_1668000_TCTX
Plotted SNPs
rs7114704
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

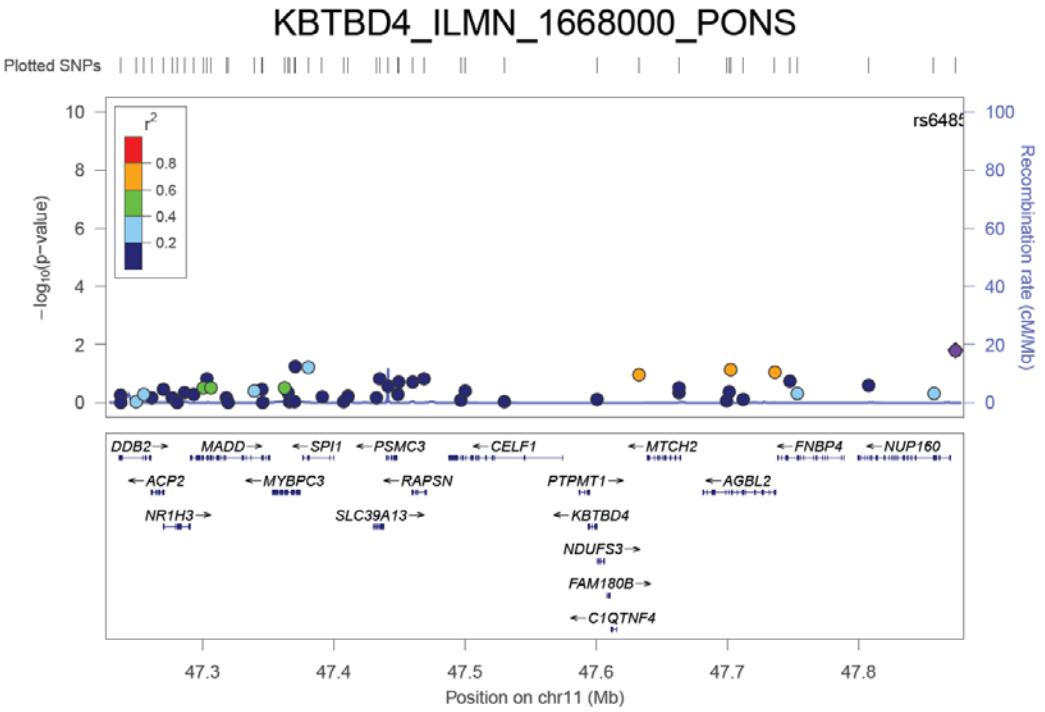
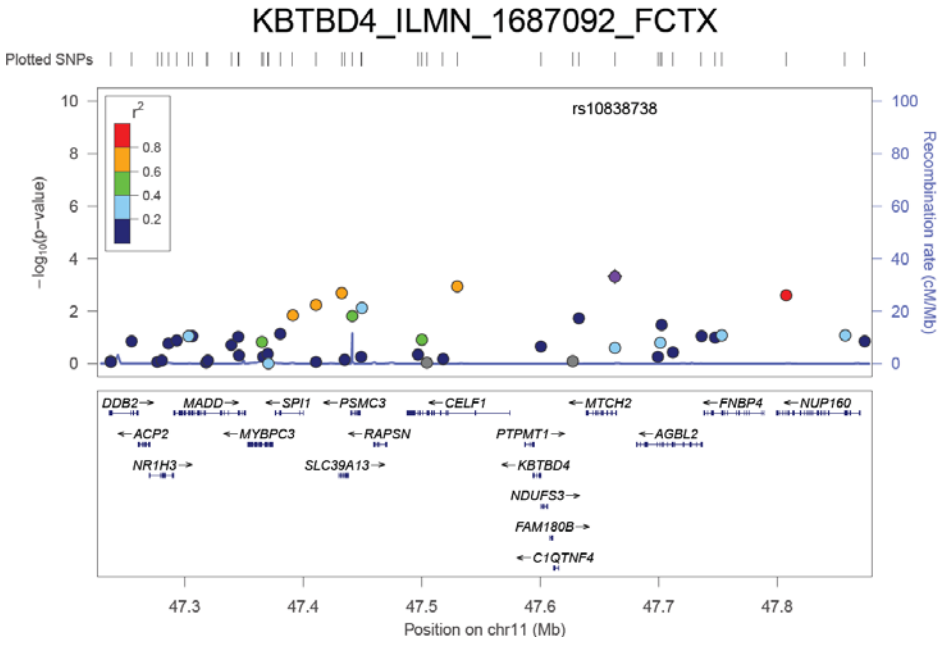
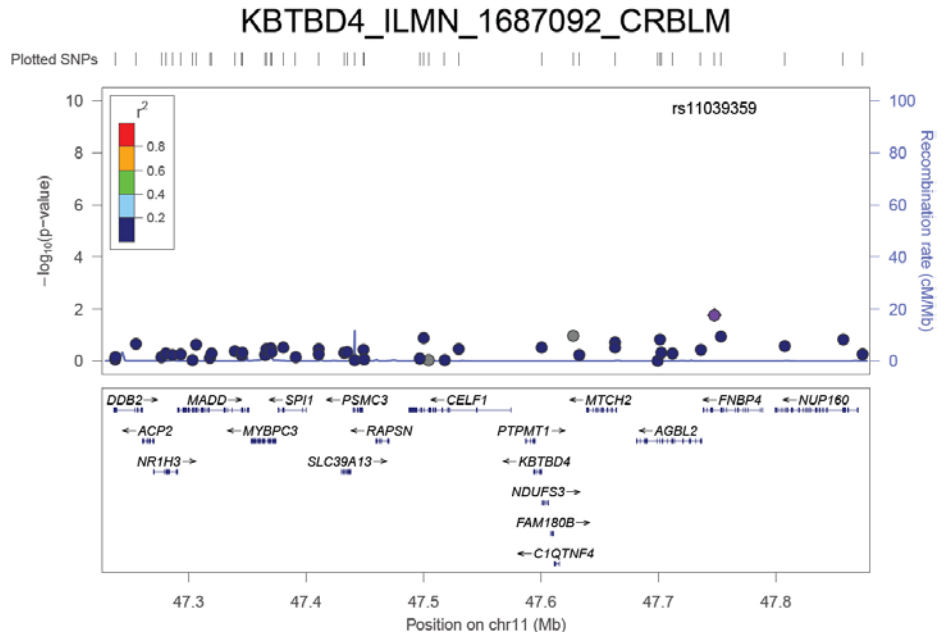

KBTBD4_ILMN_1668000_PONS
Plotted SNPs
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S8. Cis association for *KBTBD4*-ILMN_1687092 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

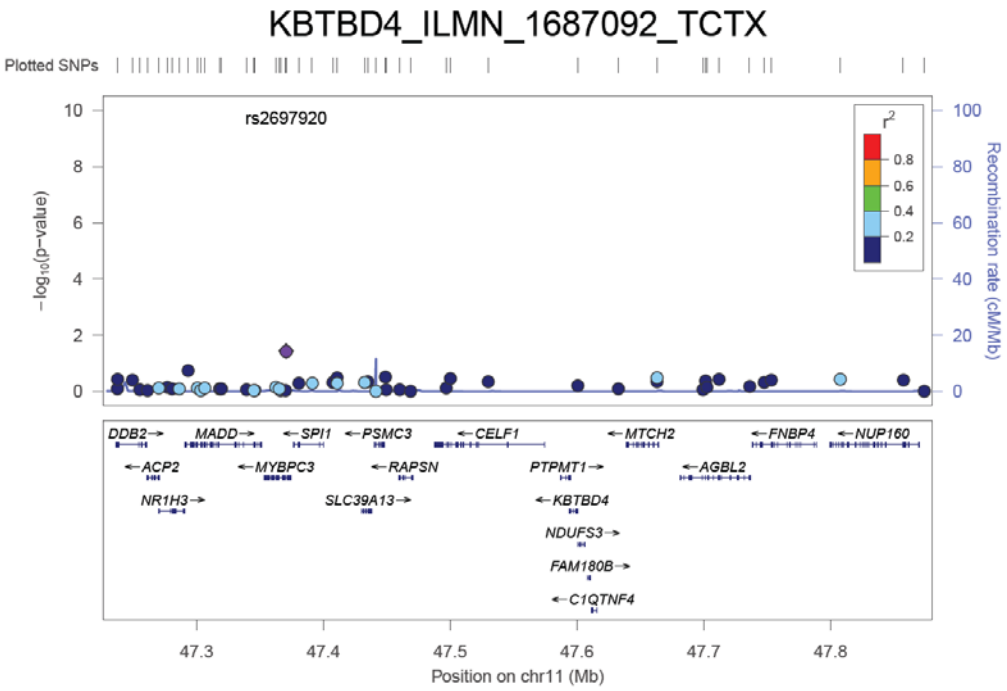
(A)



(B)



(C)


KBTBD4_ILMN_1687092_TCTX
Plotted SNPs
rs2697920
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

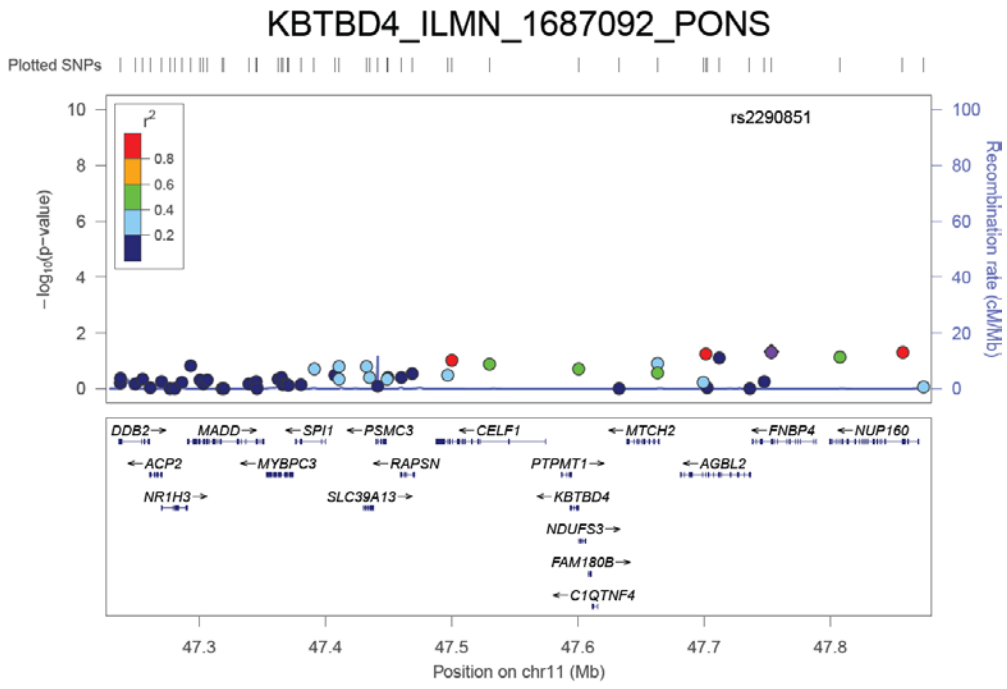
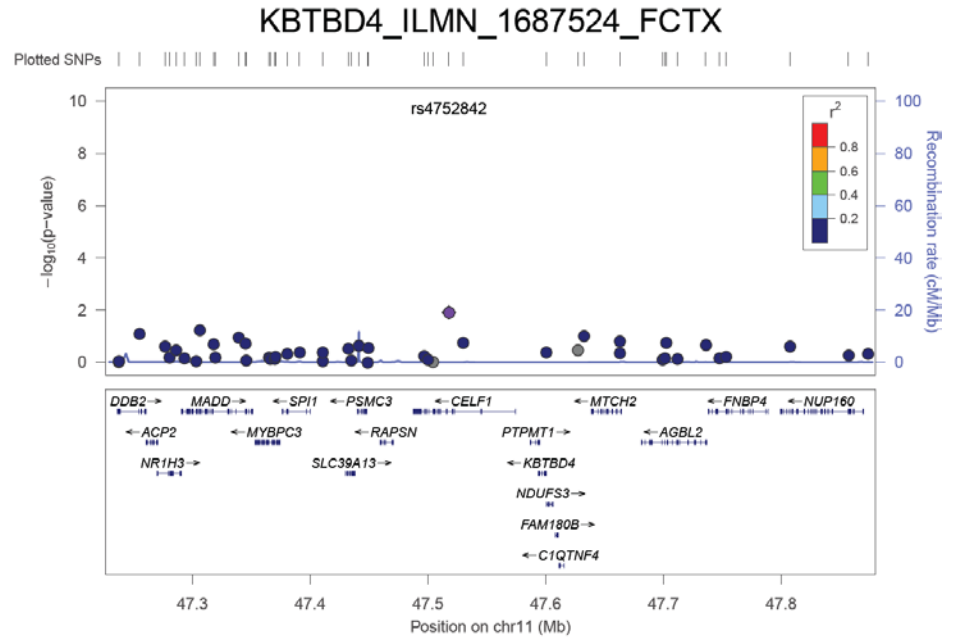

KBTBD4_ILMN_1687092_PONS
Plotted SNPs
rs2290851
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S9. Cis association for *KBTBD4*-ILMN_1687524 expression levels in (A) FCTX and (B) CRBLM.

(A)



(B)

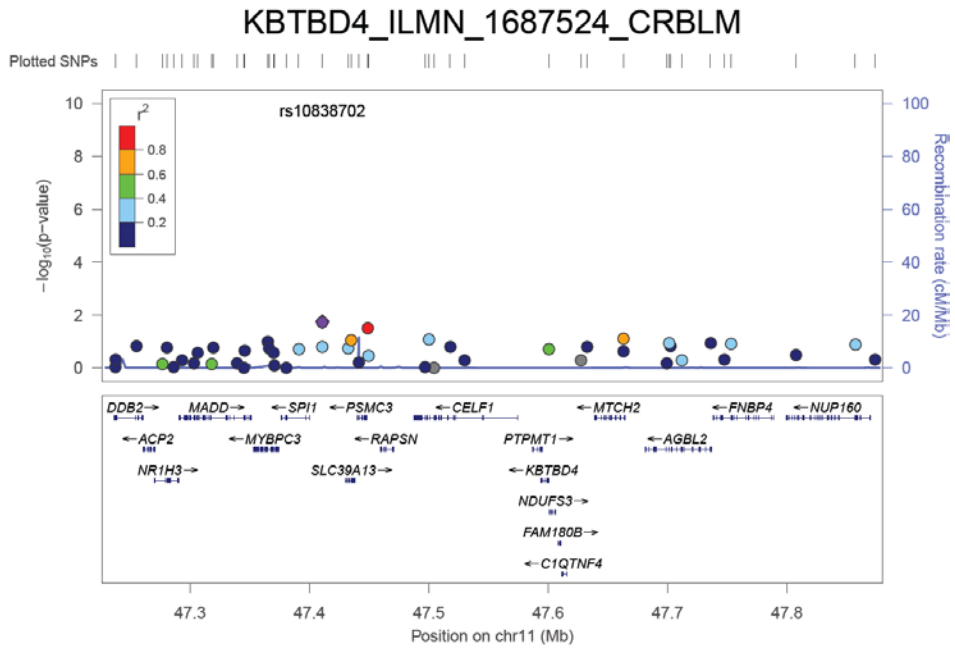
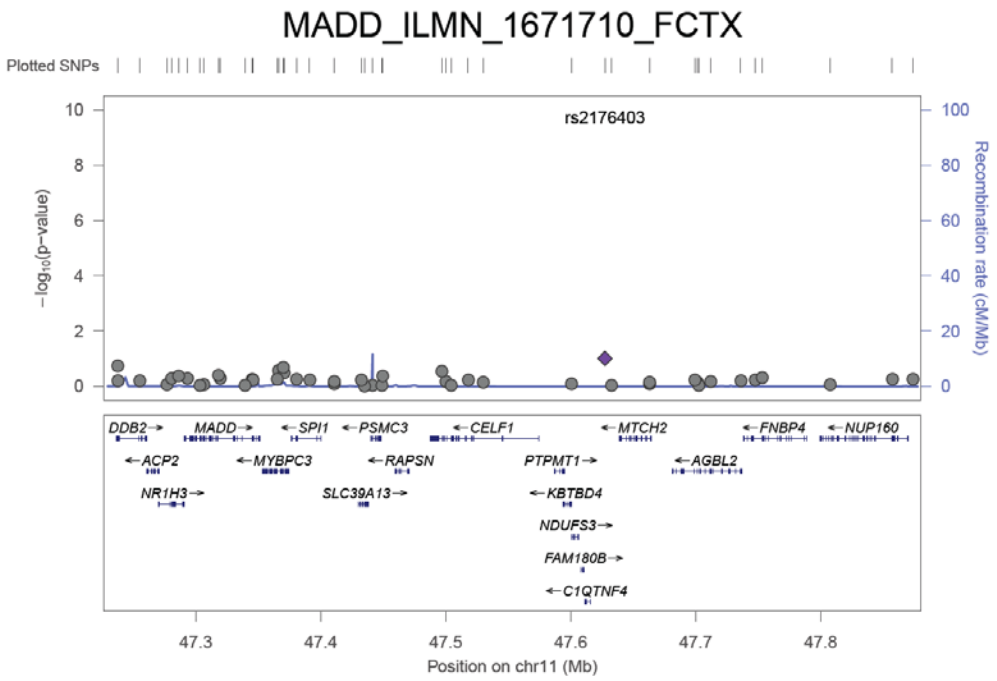
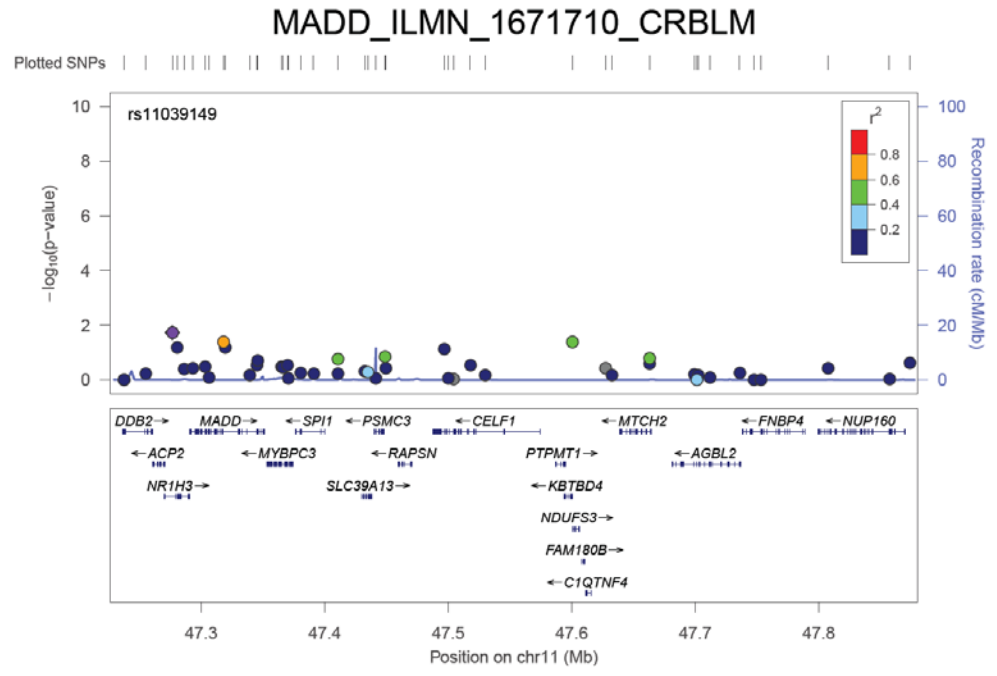


Figure S10. Cis association for *MADD*-ILMN_1671710 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

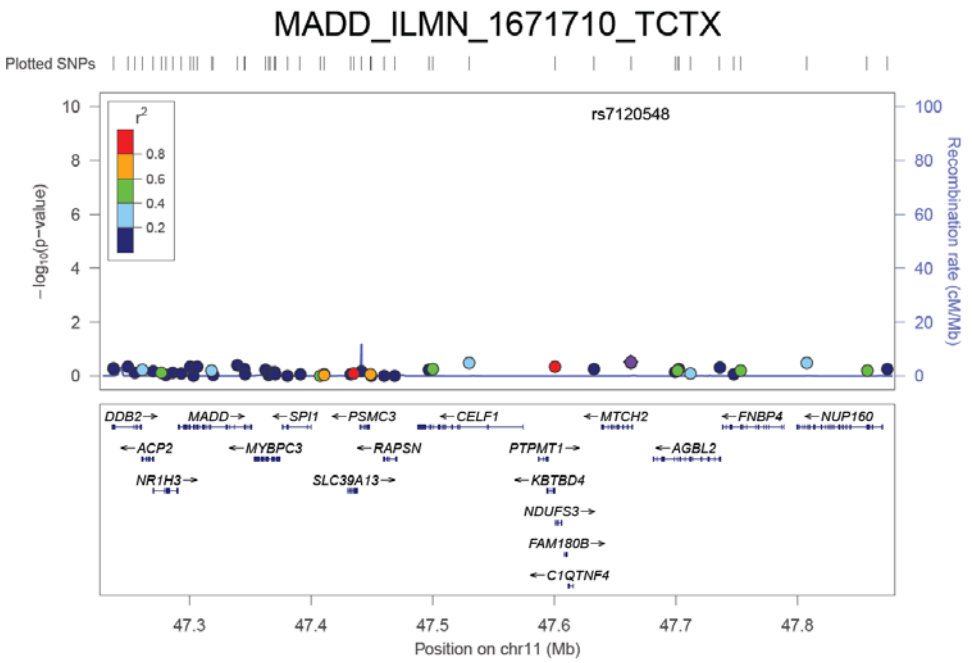
(A)



(B)



(C)


MADD_ILMN_1671710_TCTX
Plotted SNPs
rs7120548
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2→ MADD→ SPI1 PSMC3 CELF1 MTCH2 FNBP4 NUP160
ACP2 MYBPC3 RAPSN PTPMT1→ AGBL2
NR1H3→ SLC39A13→ KBTBD4
NDUFS3→
FAM180B→
C1QTNF4
47.3 47.4 47.5 47.6 47.7 47.8
Position on chr11 (Mb)

(D)

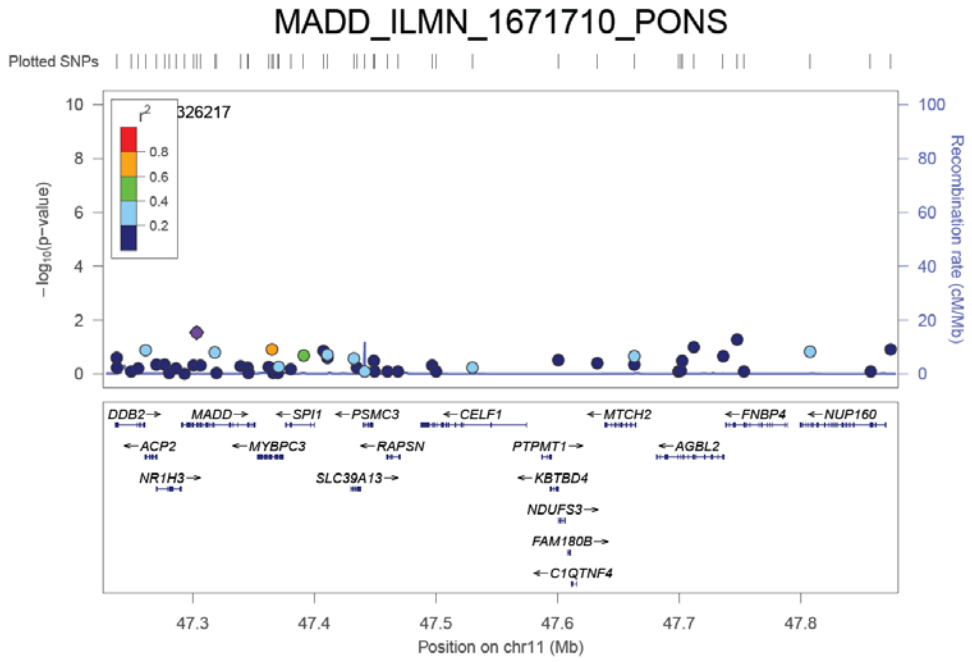
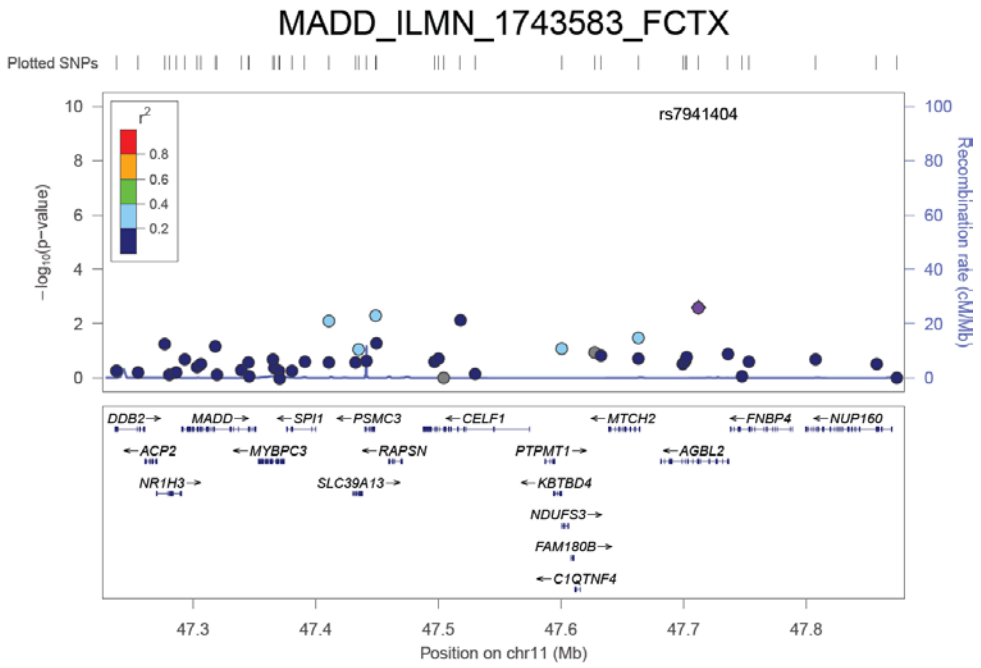
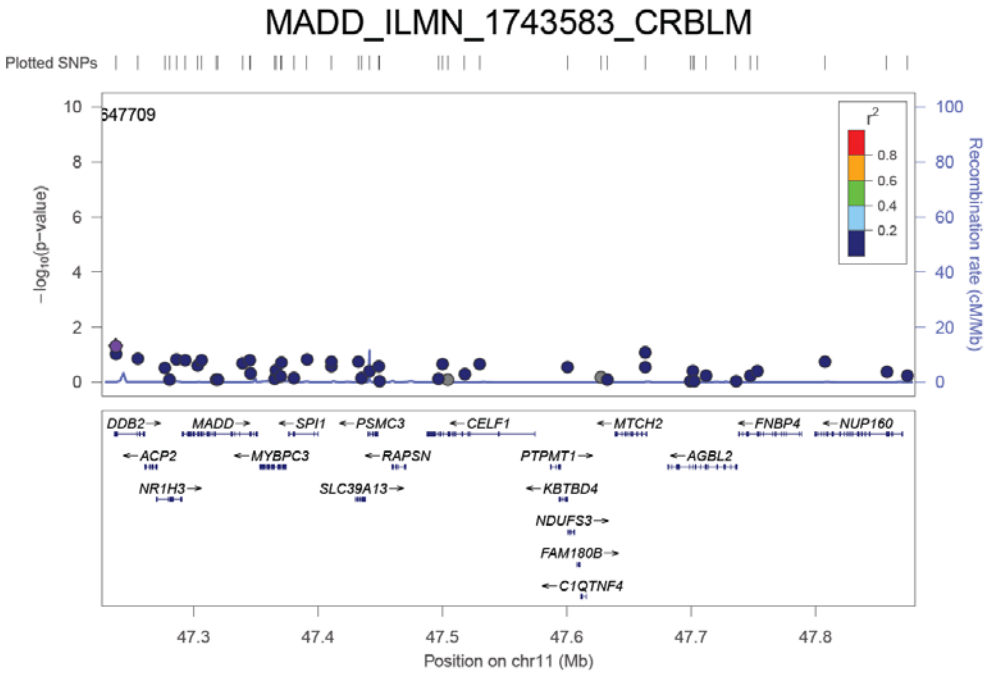

MADD_ILMN_1671710_PONS
Plotted SNPs
326217
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2→ MADD→ SPI1 PSMC3 CELF1 MTCH2 FNBP4 NUP160
ACP2 MYBPC3 RAPSN PTPMT1→ AGBL2
NR1H3→ SLC39A13→ KBTBD4
NDUFS3→
FAM180B→
C1QTNF4
47.3 47.4 47.5 47.6 47.7 47.8
Position on chr11 (Mb)

Figure S11. Cis association for *MADD*-ILMN_1743583 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

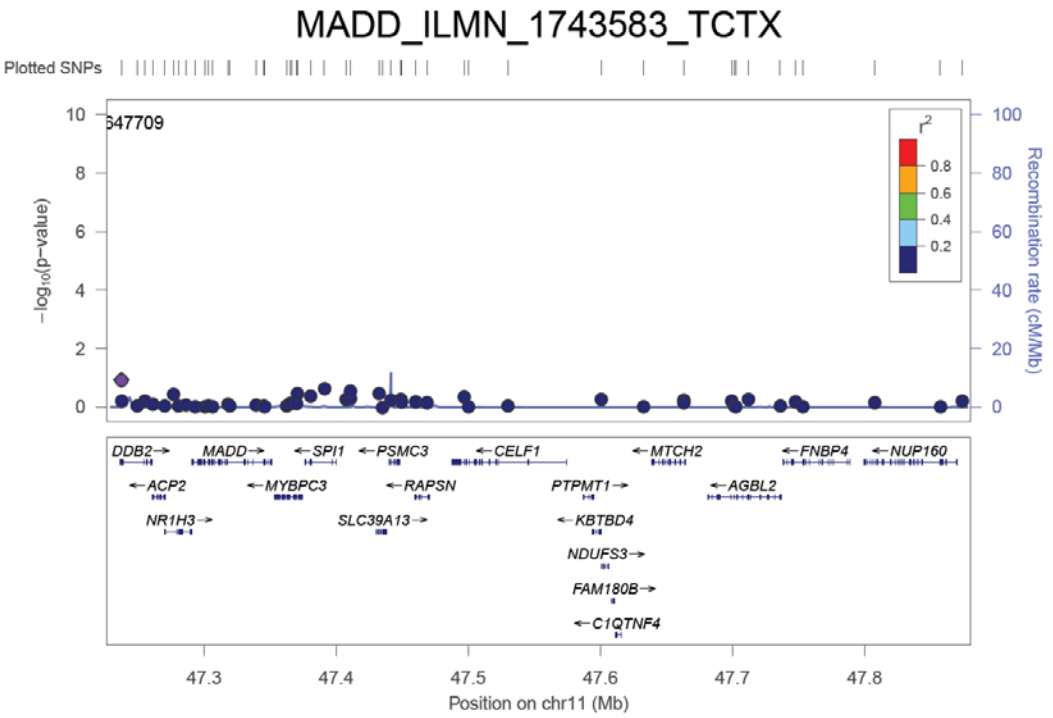
(A)



(B)



(C)


MADD_ILMN_1743583_TCTX
Plotted SNPs
rs47709
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

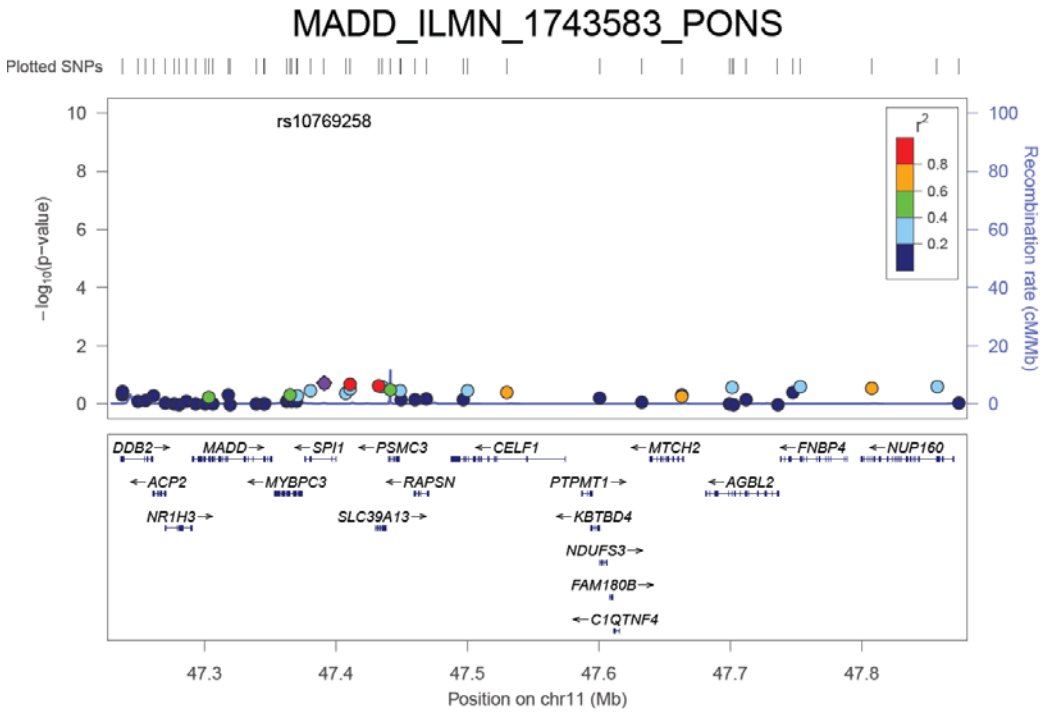
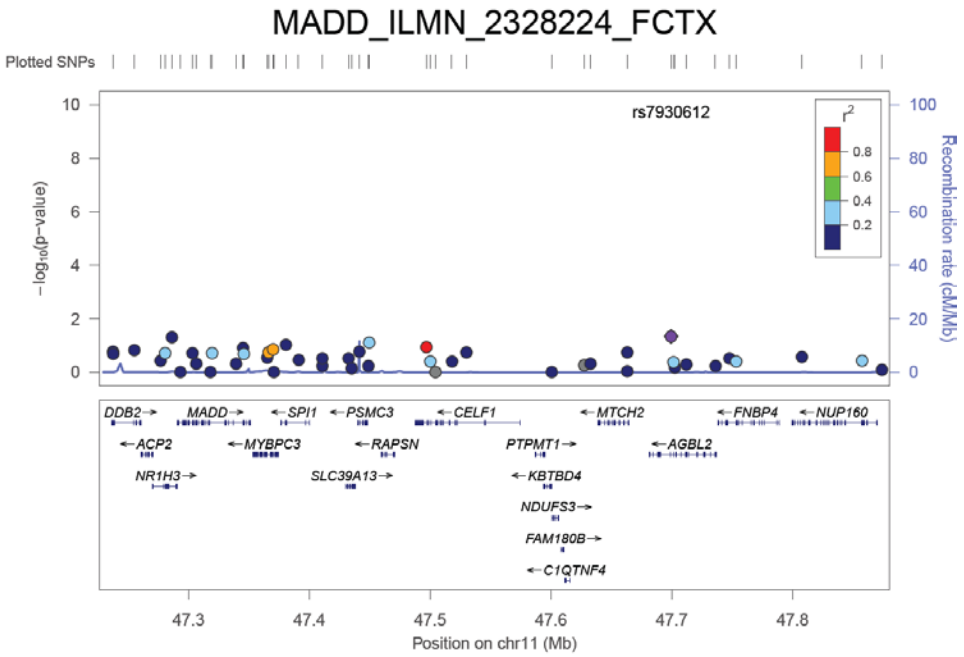

MADD_ILMN_1743583_PONS
Plotted SNPs
rs10769258
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S12. Cis association for *MADD*-ILMN_2328224 expression levels in (A) FCTX and (B) CRBLM.

(A)



(B)

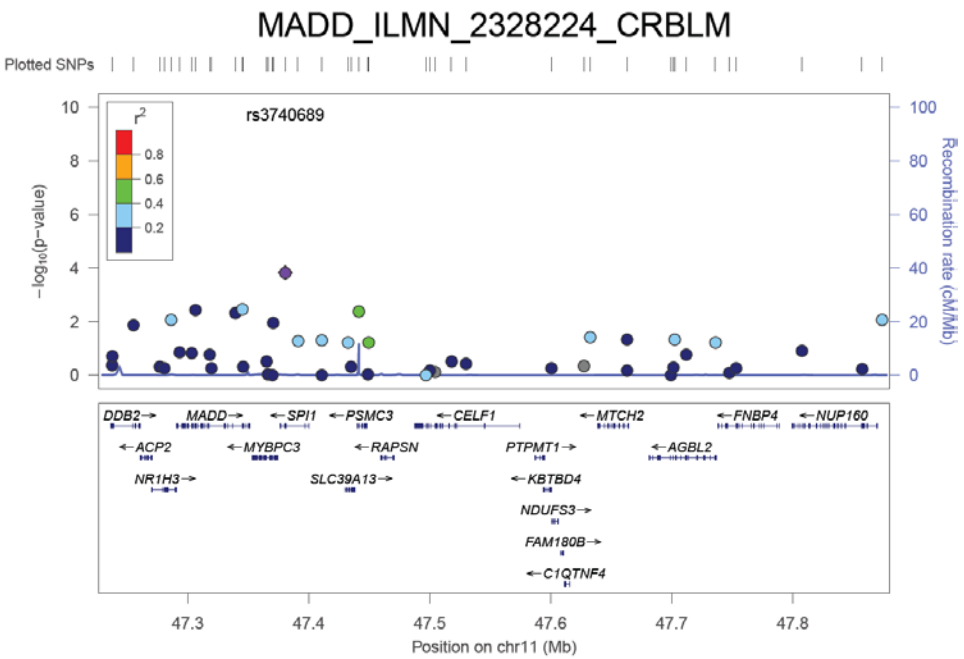
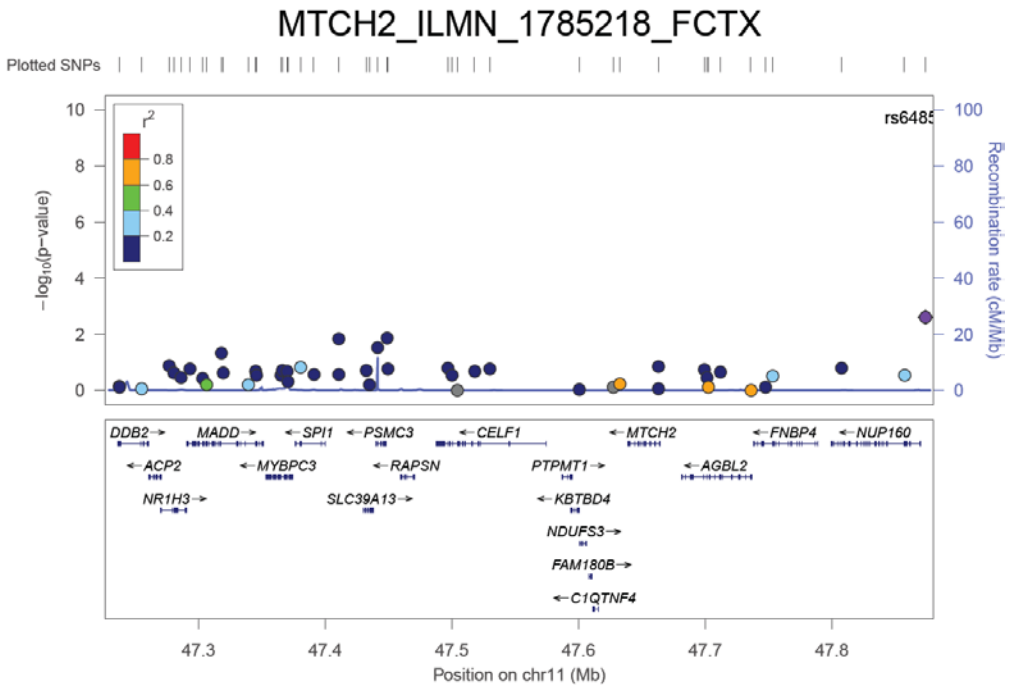
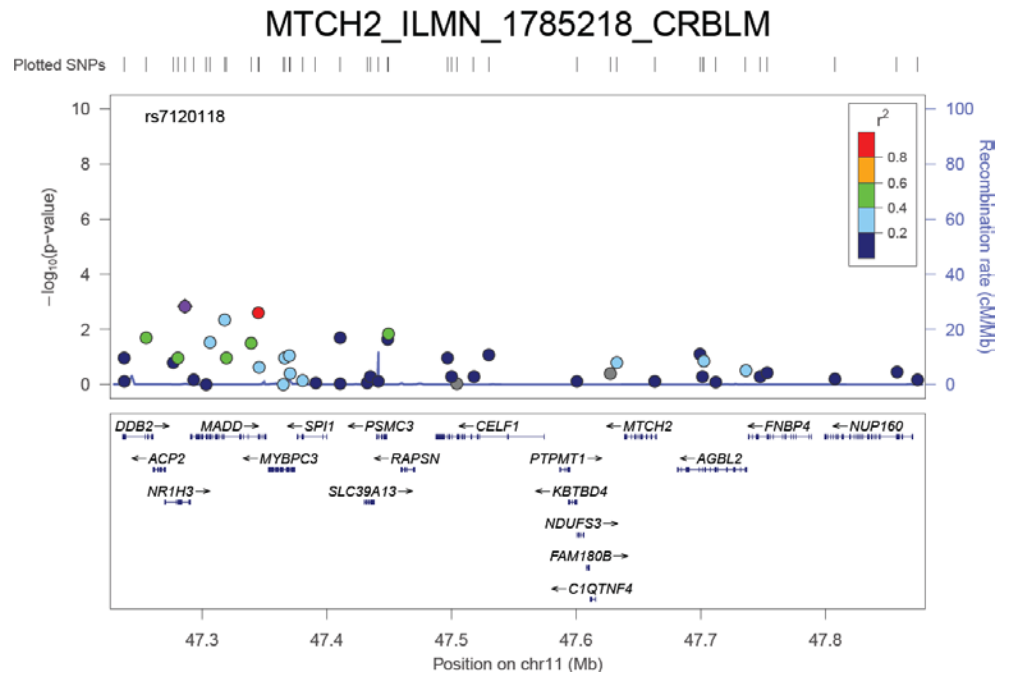


Figure S12. Cis association for *MTCH2*-ILMN_1785218 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

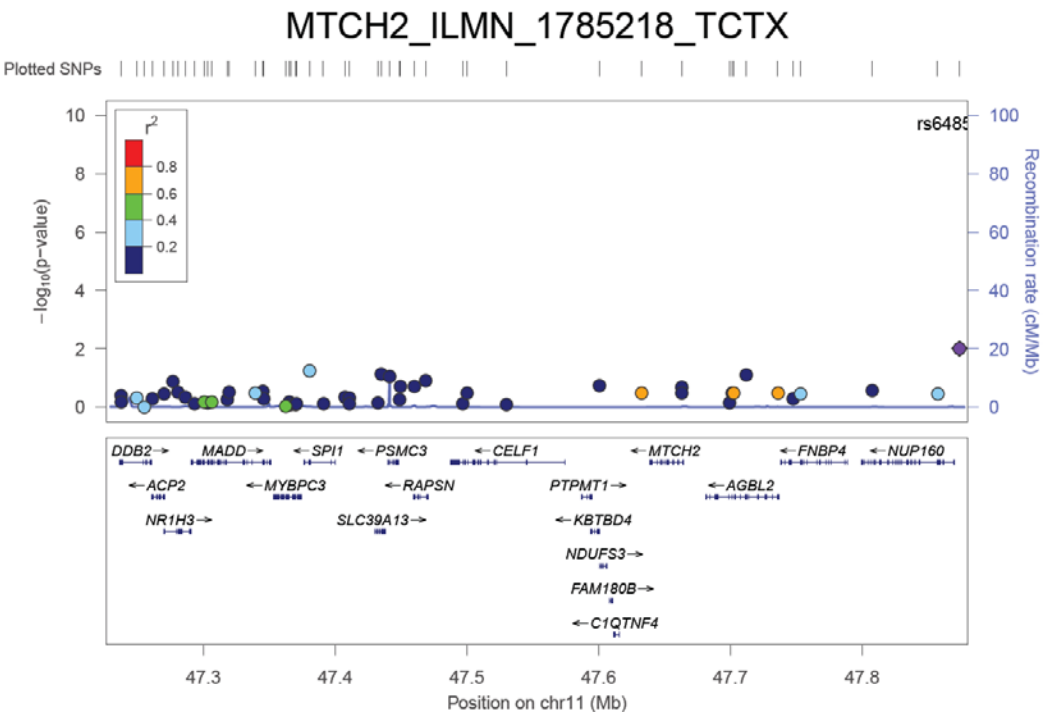
(A)



(B)



(C)


MTCH2_ILMN_1785218_TCTX
Plotted SNPs
rs6485
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
DDB2→ MADD→ ←SPI1 ←PSMC3 ←CELF1 ←MTCH2 ←FNBP4 ←NUP160
←ACP2 ←MYBPC3 ←RAPSN PTPMT1→ ←AGBL2
NR1H3→ SLC39A13→ ←KBTBD4
NDUFS3→
FAM180B→
←C1QTNF4
47.3 47.4 47.5 47.6 47.7 47.8
Position on chr11 (Mb)

(D)

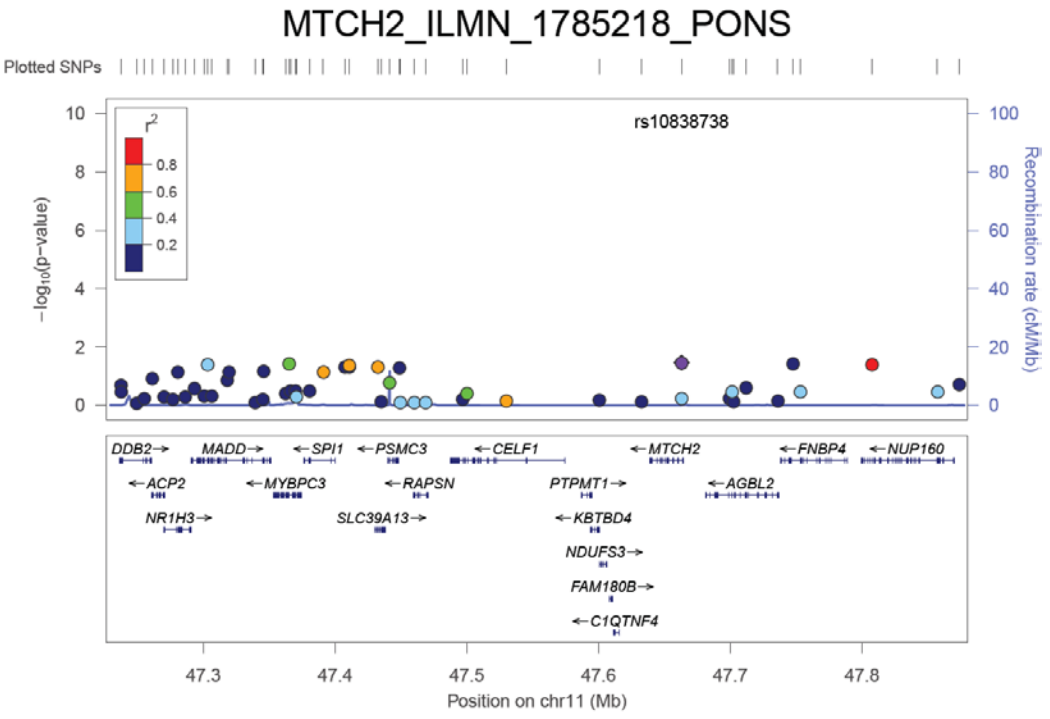
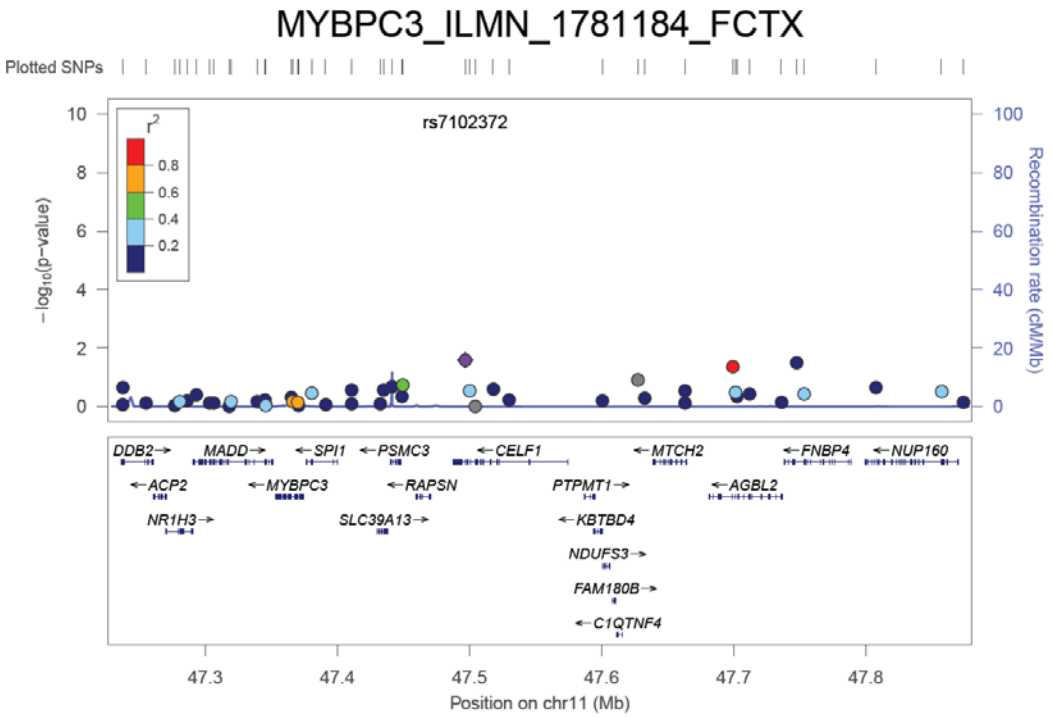

MTCH2_ILMN_1785218_PONS
Plotted SNPs
rs10838738
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
DDB2→ MADD→ ←SPI1 ←PSMC3 ←CELF1 ←MTCH2 ←FNBP4 ←NUP160
←ACP2 ←MYBPC3 ←RAPSN PTPMT1→ ←AGBL2
NR1H3→ SLC39A13→ ←KBTBD4
NDUFS3→
FAM180B→
←C1QTNF4
47.3 47.4 47.5 47.6 47.7 47.8
Position on chr11 (Mb)

Figure S13. Cis association for *MYBPC3*-ILMN_1781184 expression levels in (A) FCTX and (B) CRBLM.

(A)



(B)

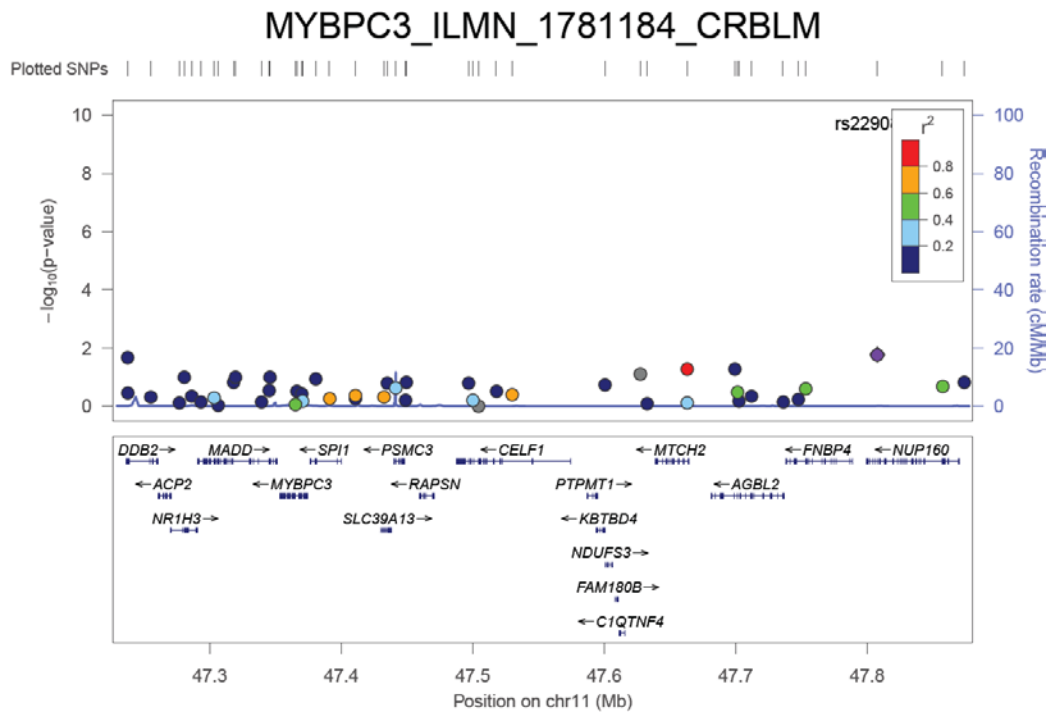
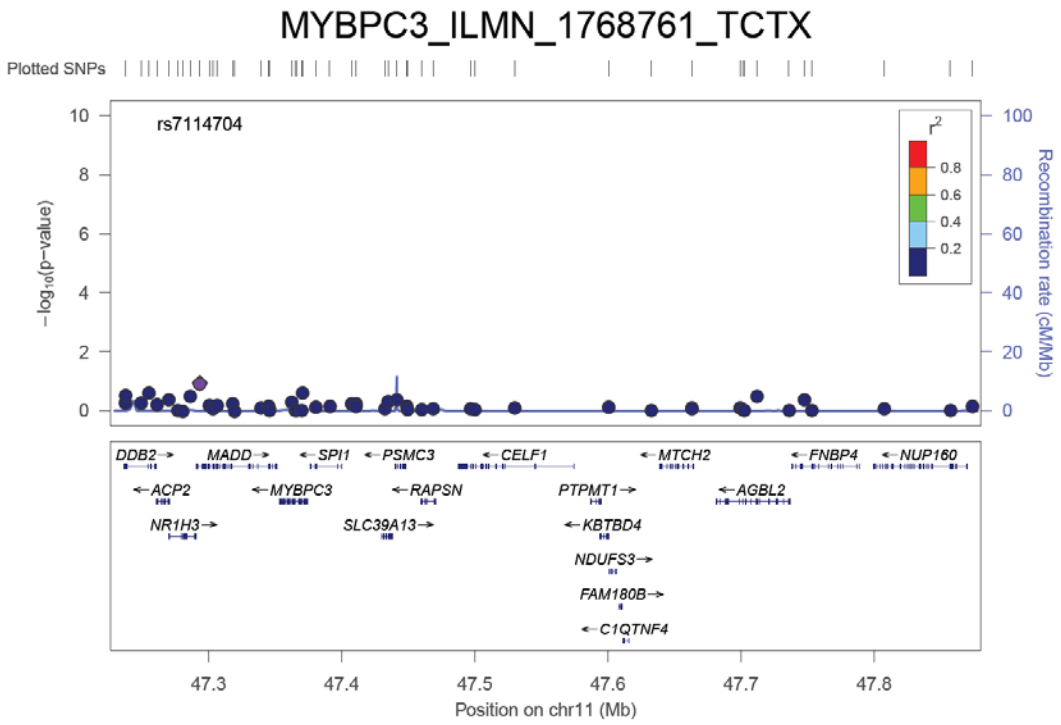


Figure S14. Cis association for *MYBPC3*-ILMN_1768761 expression levels in (A) TCTX and (B) PONS.

(A)



(B)

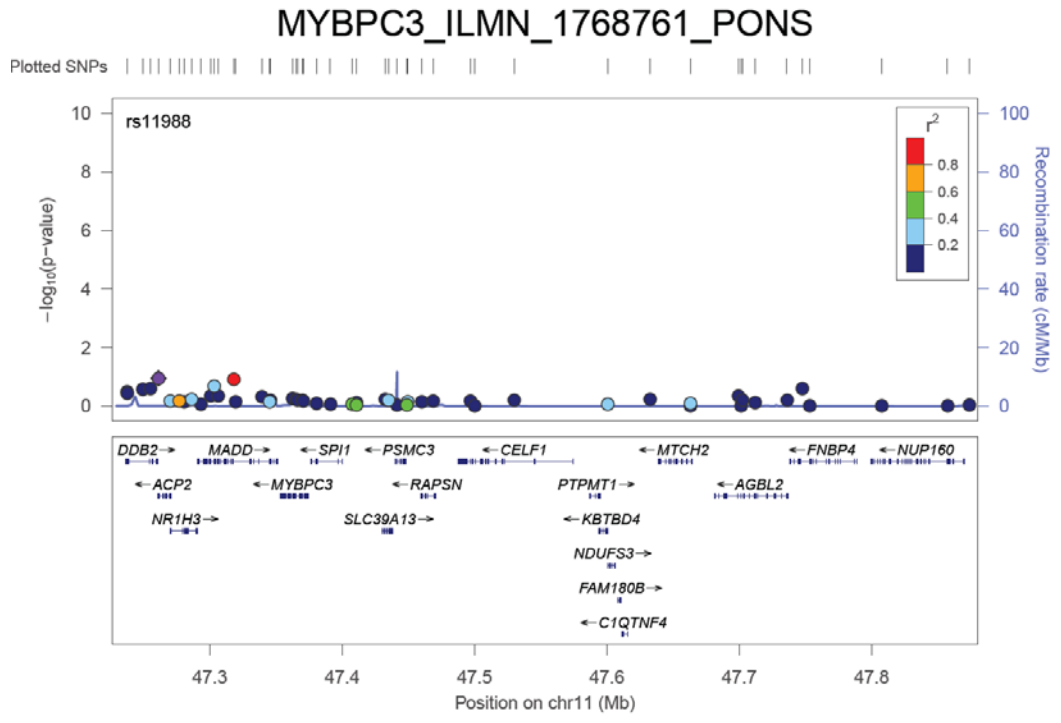
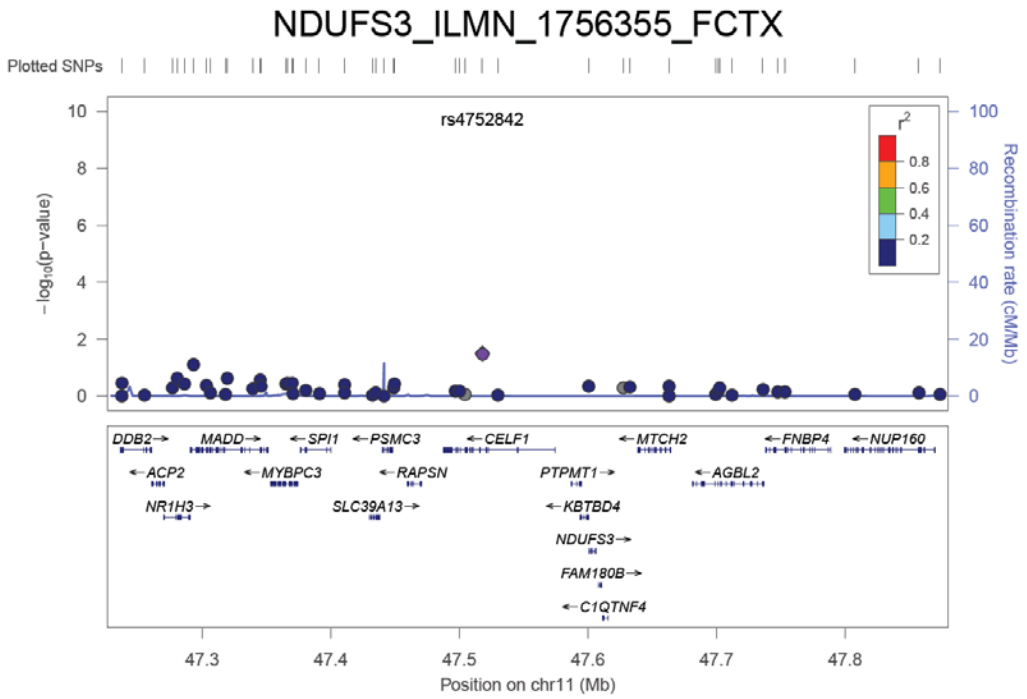
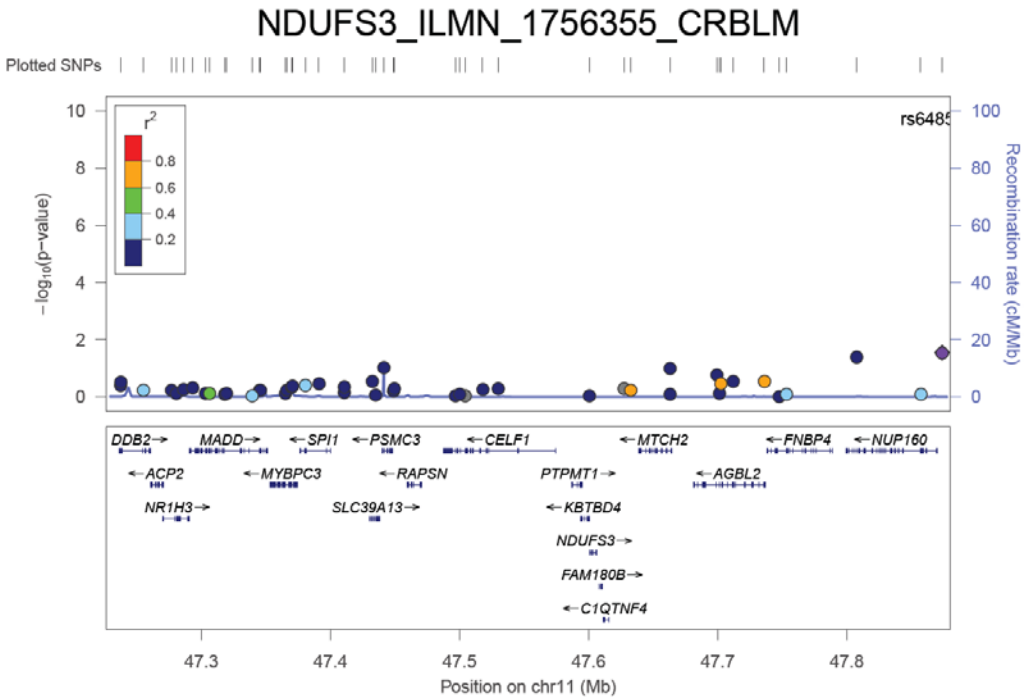


Figure S15. Cis association for *NDUFS3*-ILMN_1756355 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

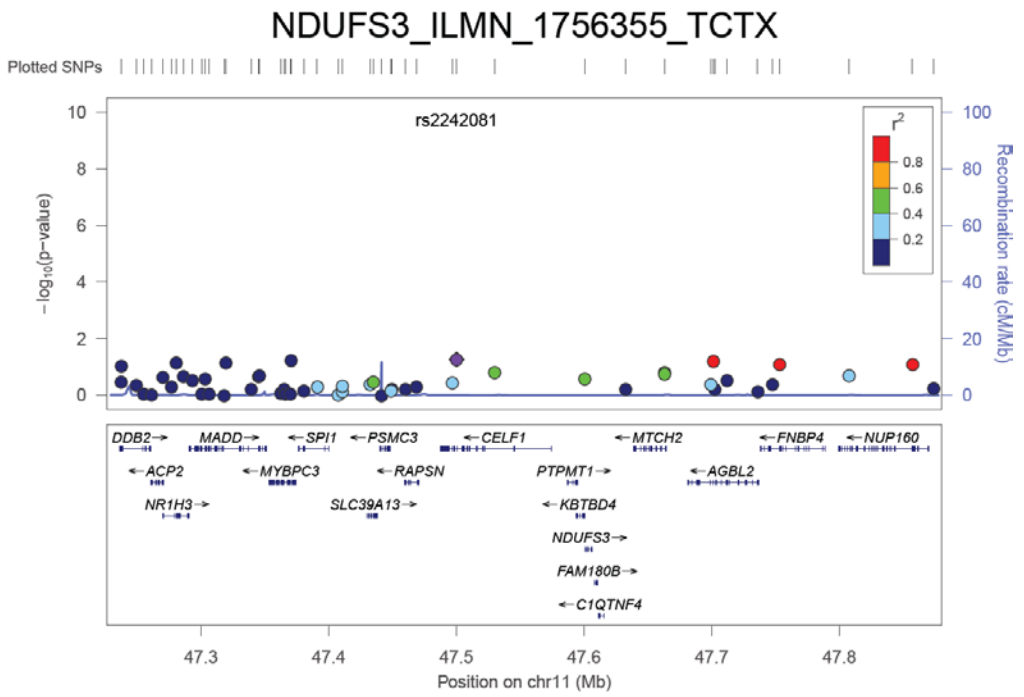
(A)



(B)



(C)


NDUFS3_ILMN_1756355_TCTX
Plotted SNPs
rs2242081
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

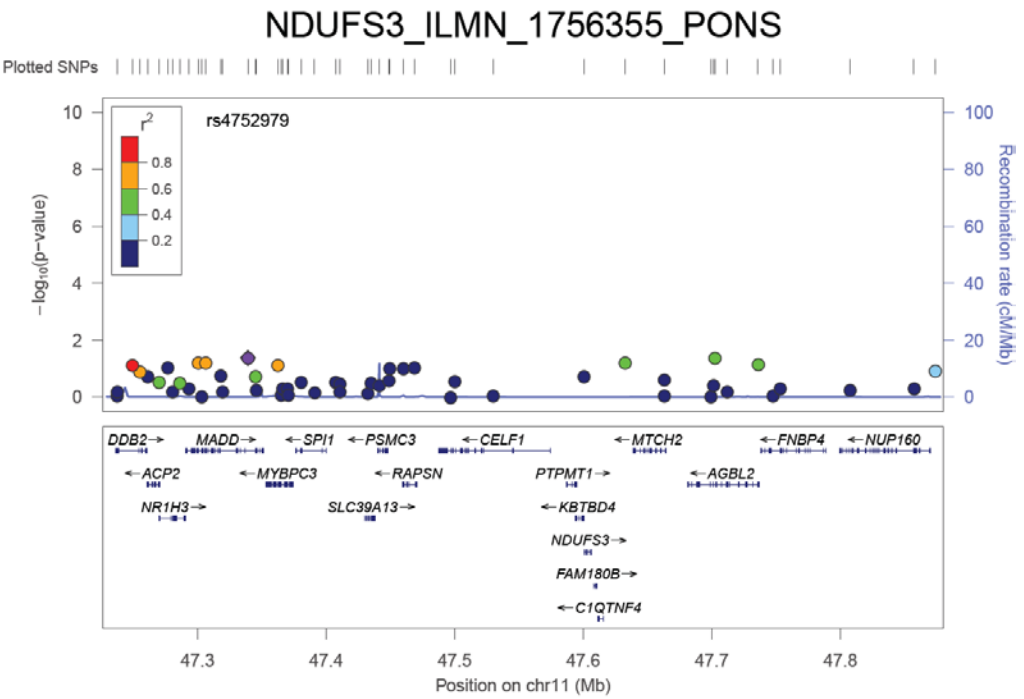
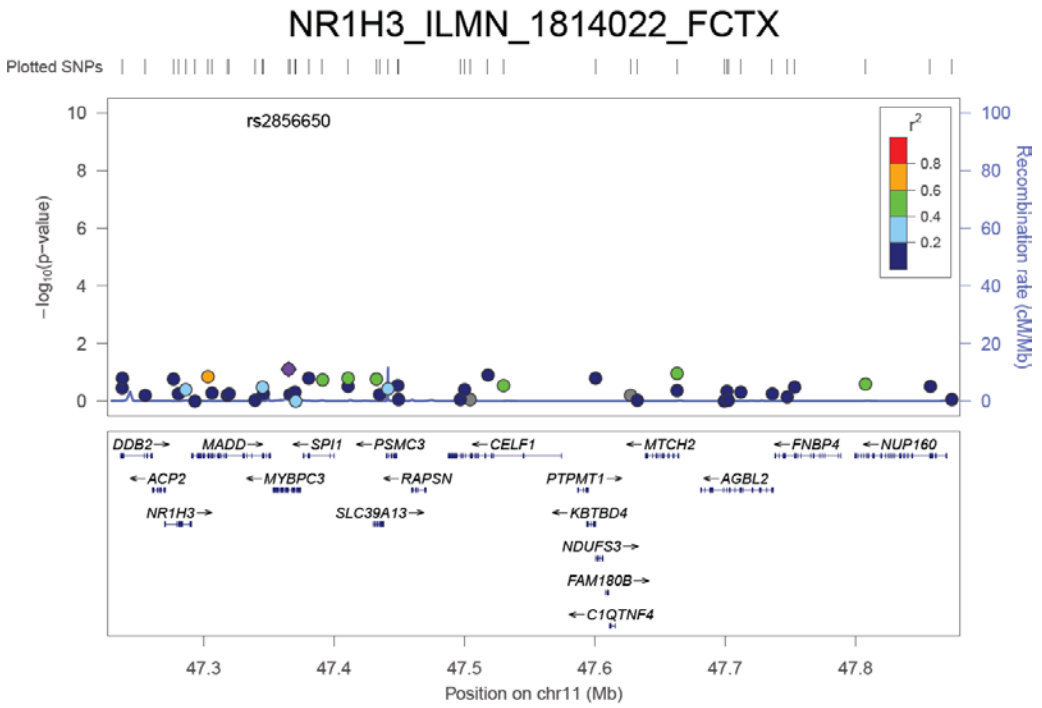
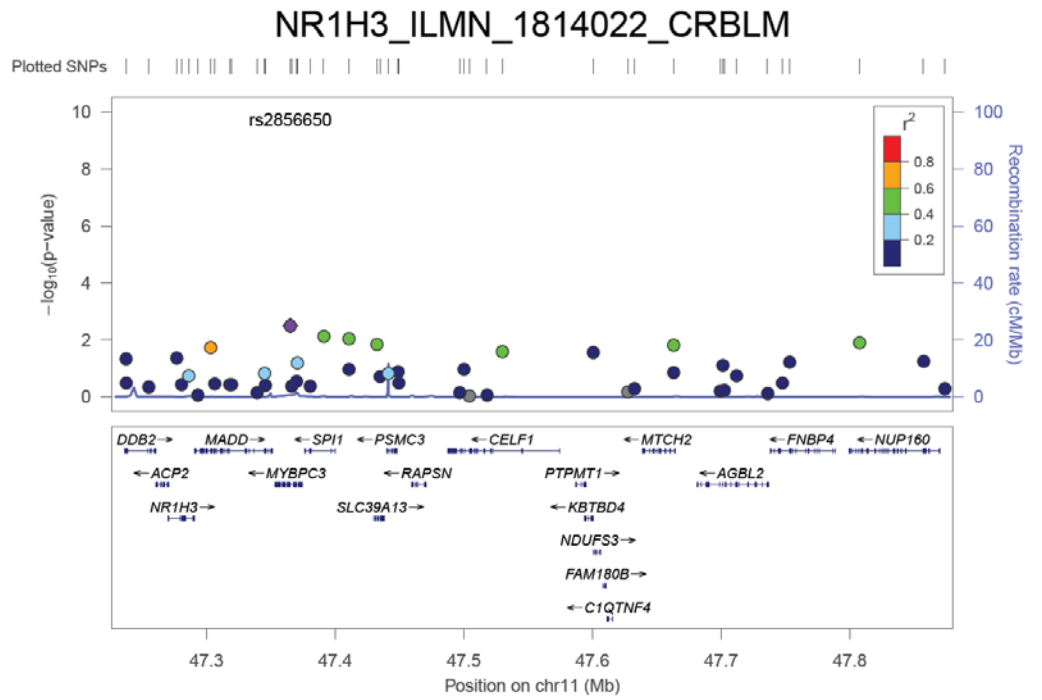

NDUFS3_ILMN_1756355_PONS
Plotted SNPs
rs4752979
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S16. Cis association for *NR1H3*-ILMN_1814022 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

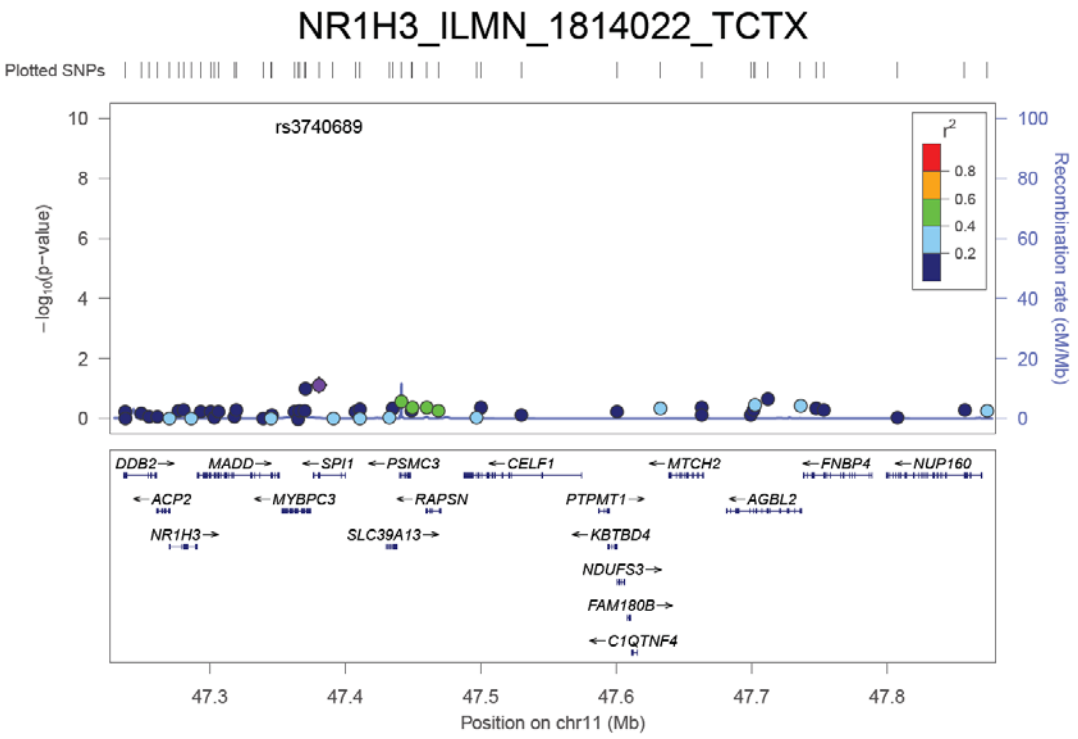
(A)



(B)



(C)


NR1H3_ILMN_1814022_TCTX
Plotted SNPs
rs3740689
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

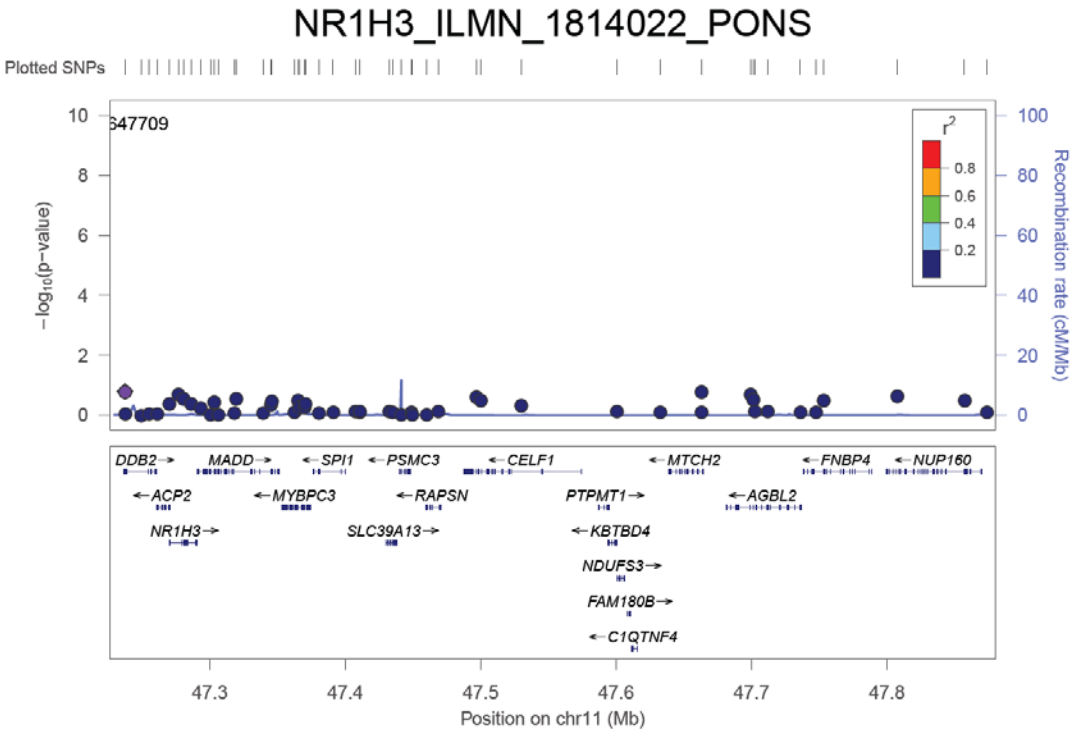
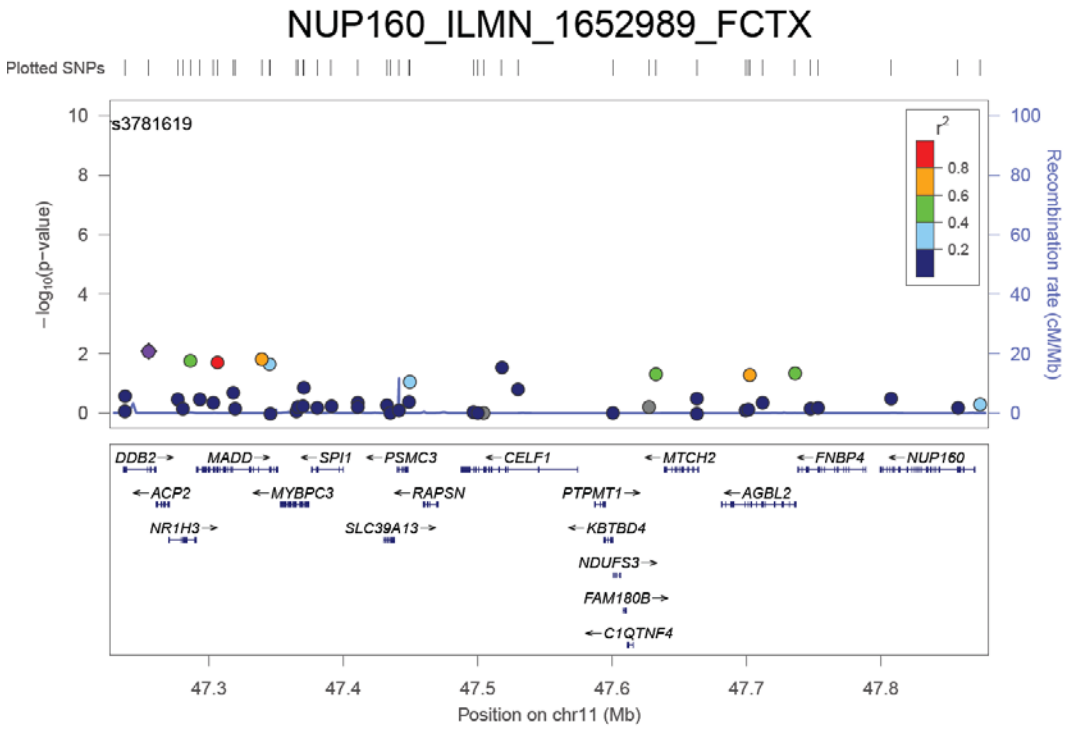
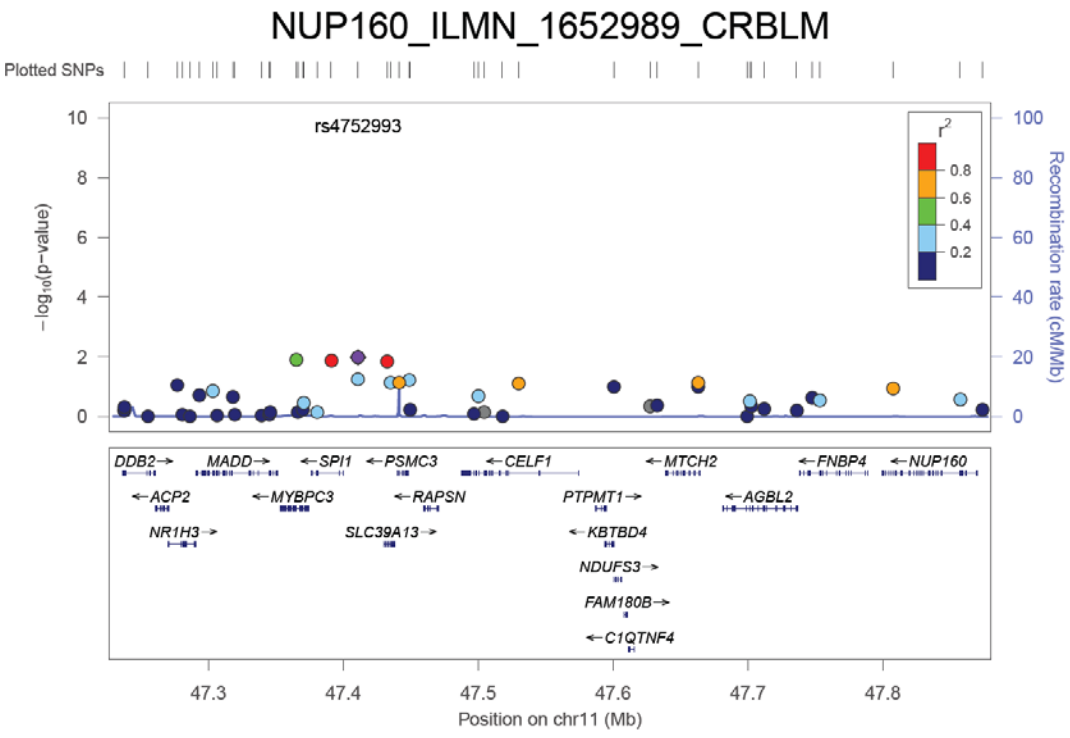

NR1H3_ILMN_1814022_PONS
Plotted SNPs
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S17. Cis association for *NUP160*-ILMN_1652989 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

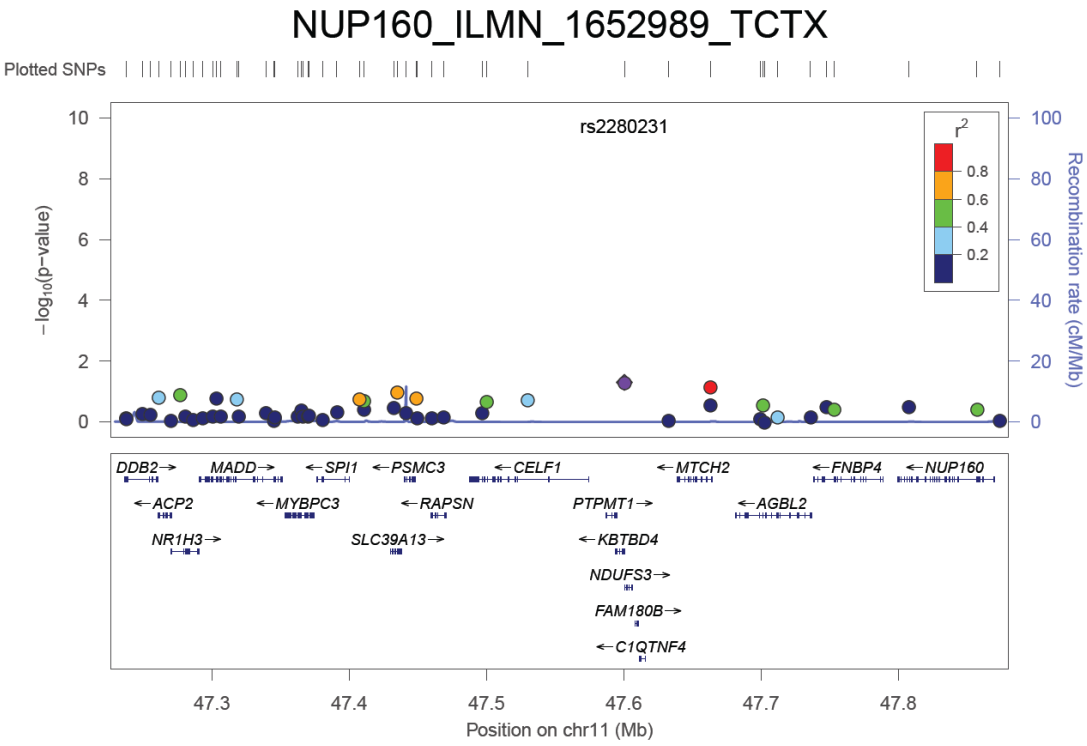
(A)



(B)



(C)


NUP160_ILMN_1652989_TCTX
Plotted SNPs
rs2280231
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2→
MADD→
←SPI1
←PSMC3
←CELF1
←MTCH2
←FNBP4
←NUP160
←ACP2
←MYBPC3
←RAPSN
PTPMT1→
←AGBL2
NR1H3→
SLC39A13→
←KBTBD4
NDUFS3→
FAM180B→
←C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

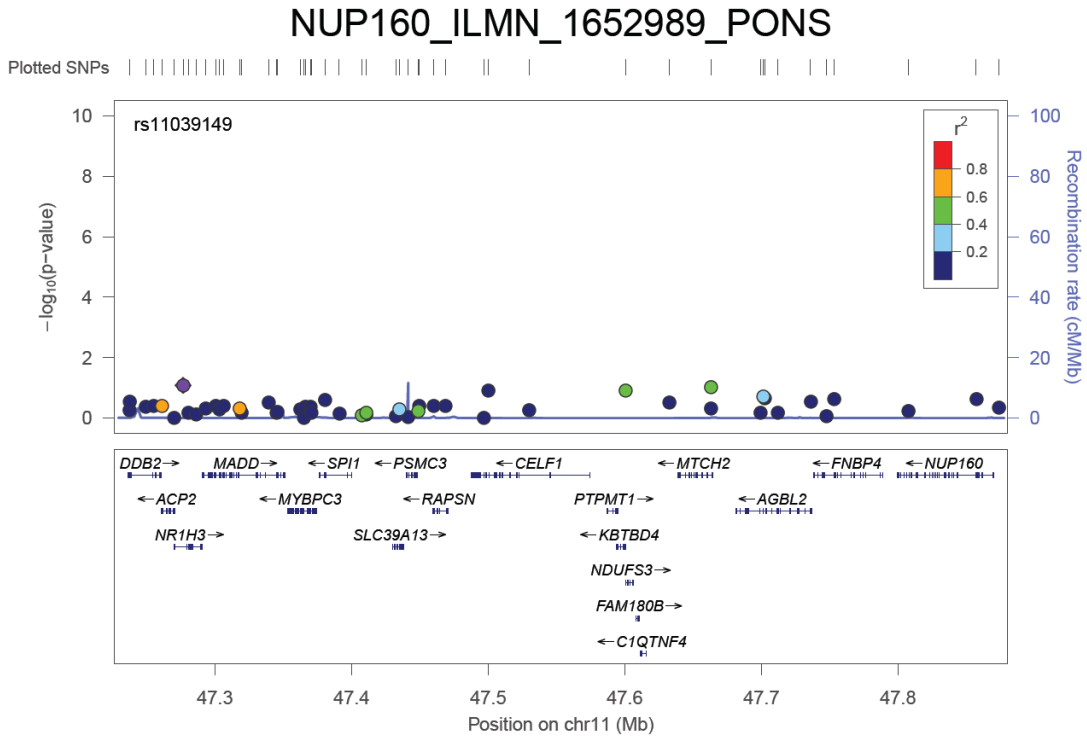
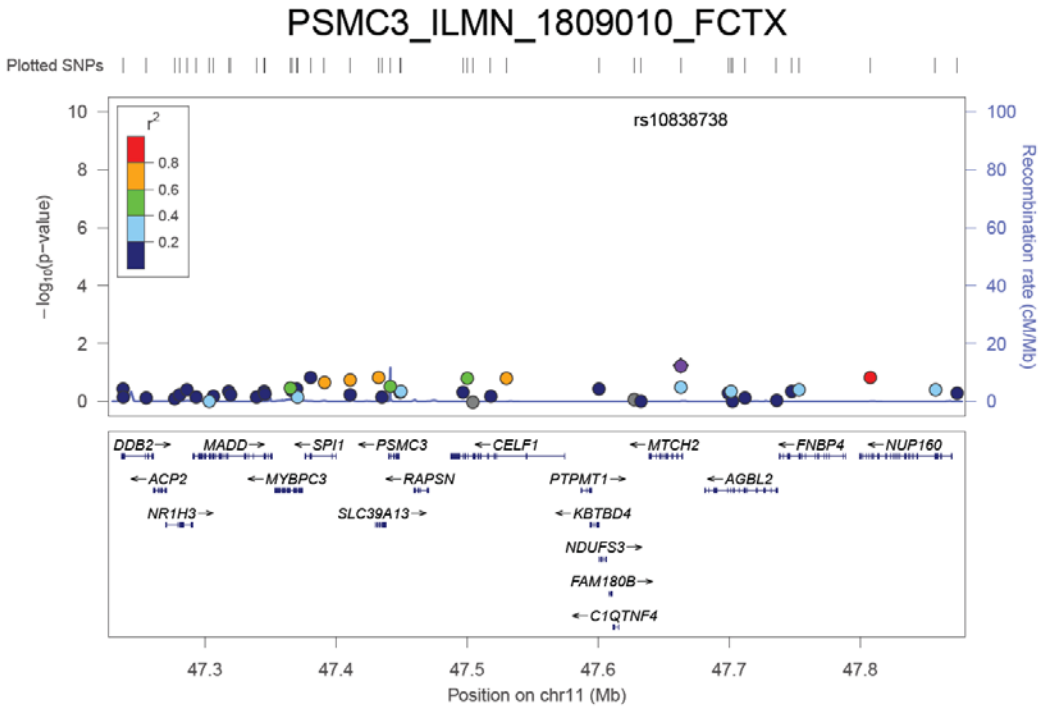
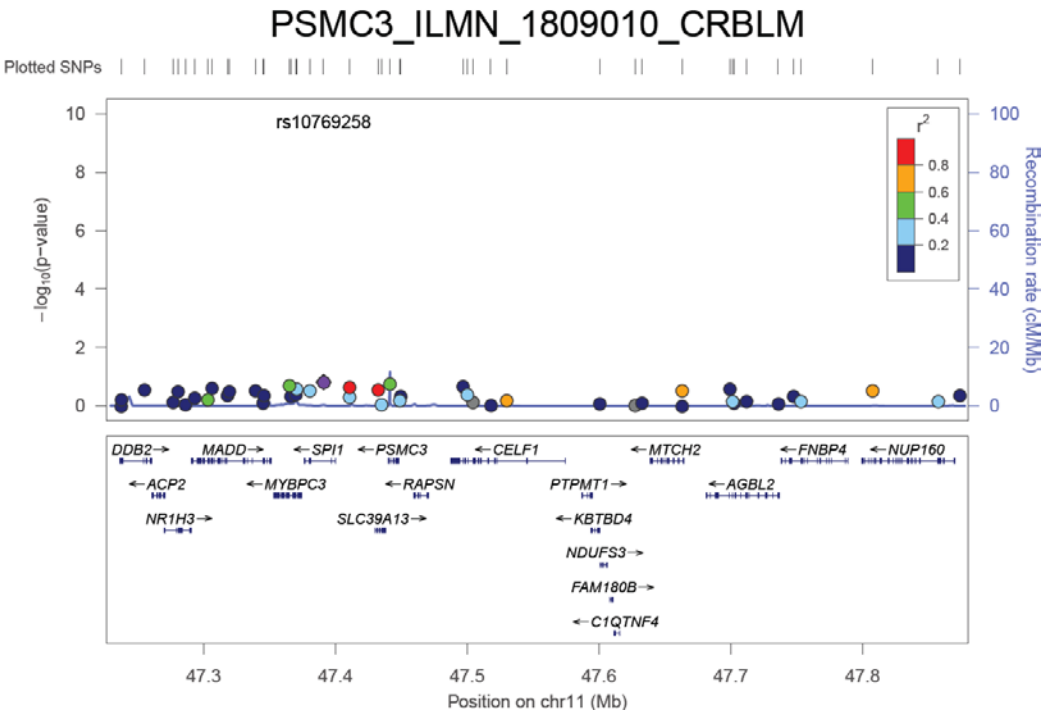

NUP160_ILMN_1652989_PONS
Plotted SNPs
rs11039149
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2→
MADD→
←SPI1
←PSMC3
←CELF1
←MTCH2
←FNBP4
←NUP160
←ACP2
←MYBPC3
←RAPSN
PTPMT1→
←AGBL2
NR1H3→
SLC39A13→
←KBTBD4
NDUFS3→
FAM180B→
←C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S18. Cis association for *PSMC3*-ILMN_1809010 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

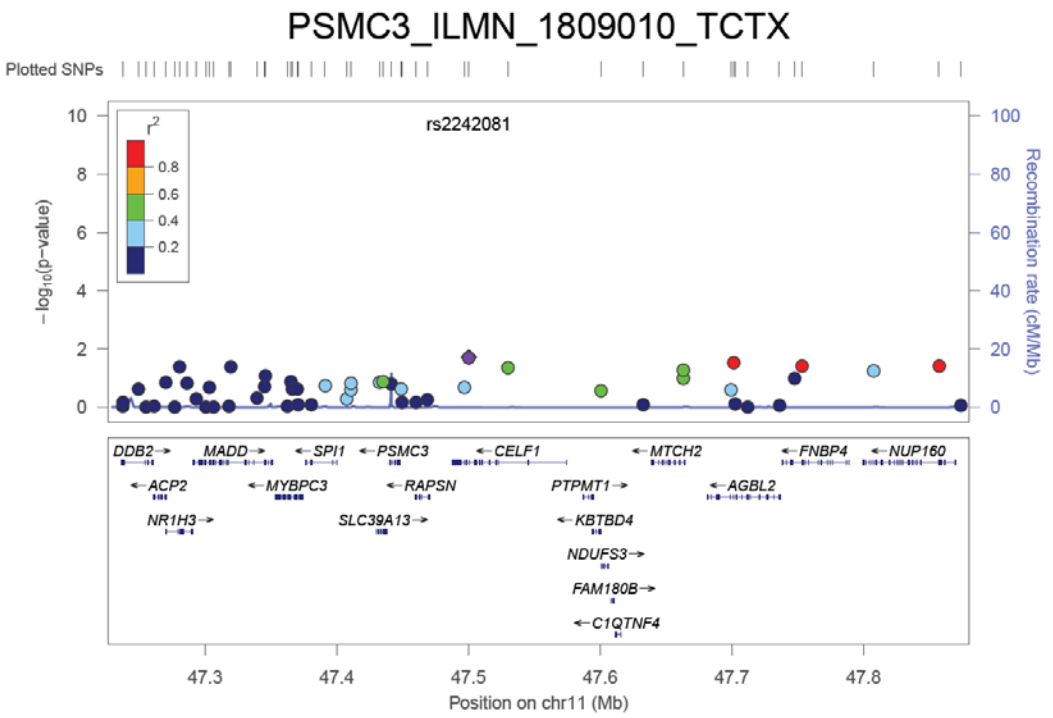
(A)



(B)



(C)


PSMC3_ILMN_1809010_TCTX
Plotted SNPs
rs2242081
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

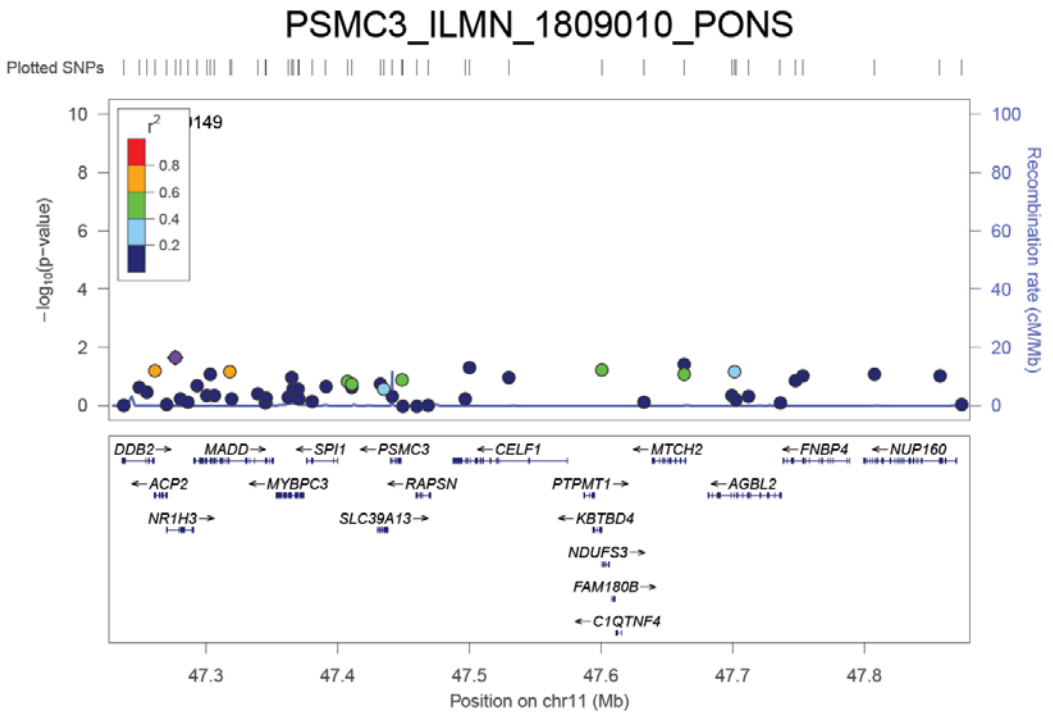
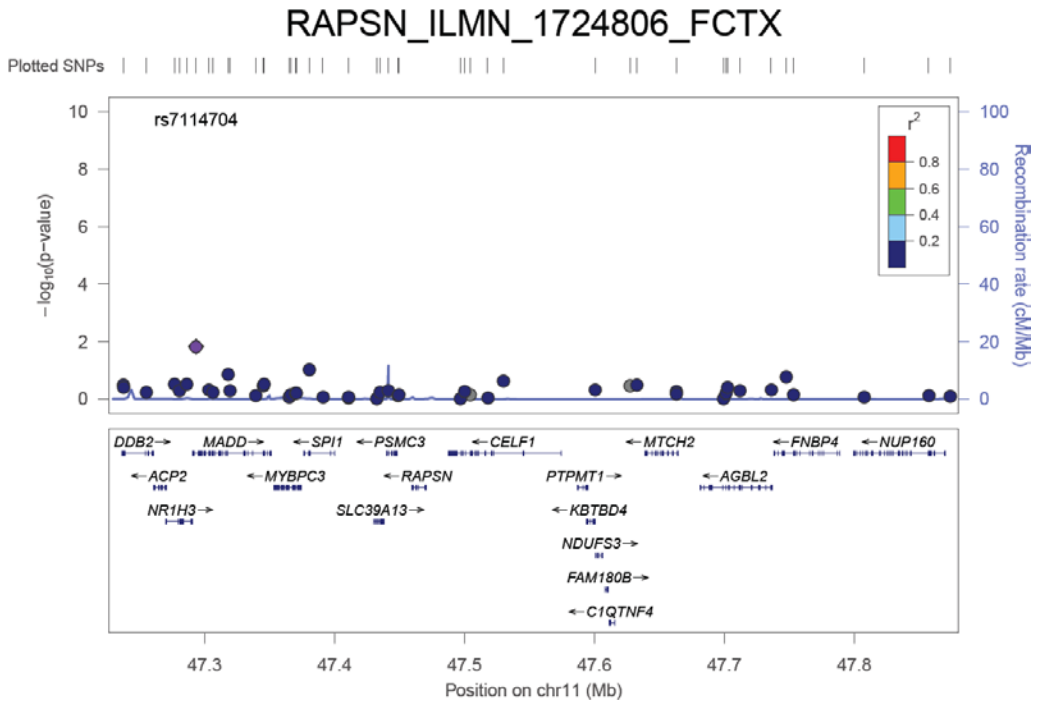
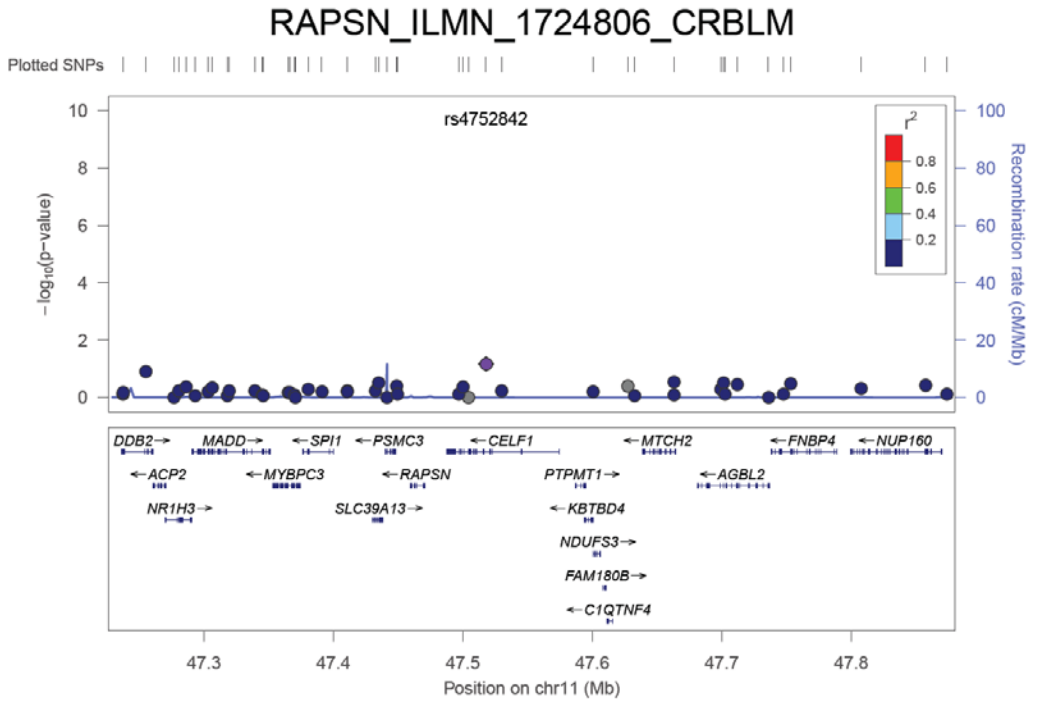

PSMC3_ILMN_1809010_PONS
Plotted SNPs
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S19. Cis association for *RAPSN*-ILMN_17024806 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

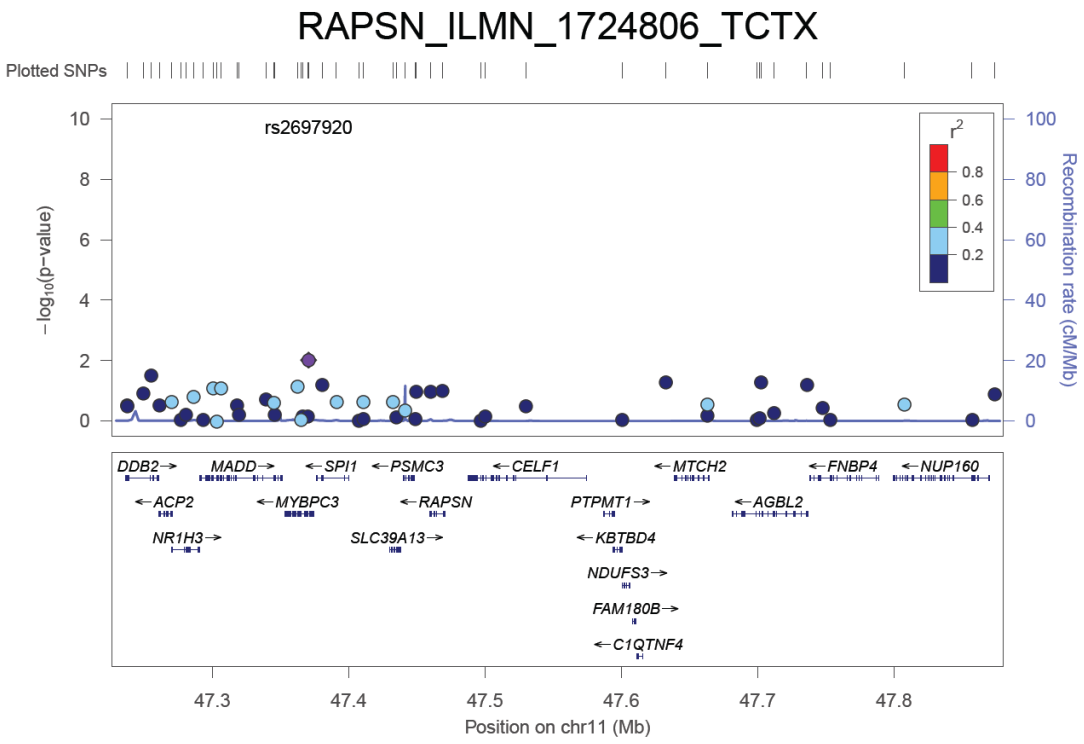
(A)



(B)



(C)


RAPSN_ILMN_1724806_TCTX
Plotted SNPs
rs2697920
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

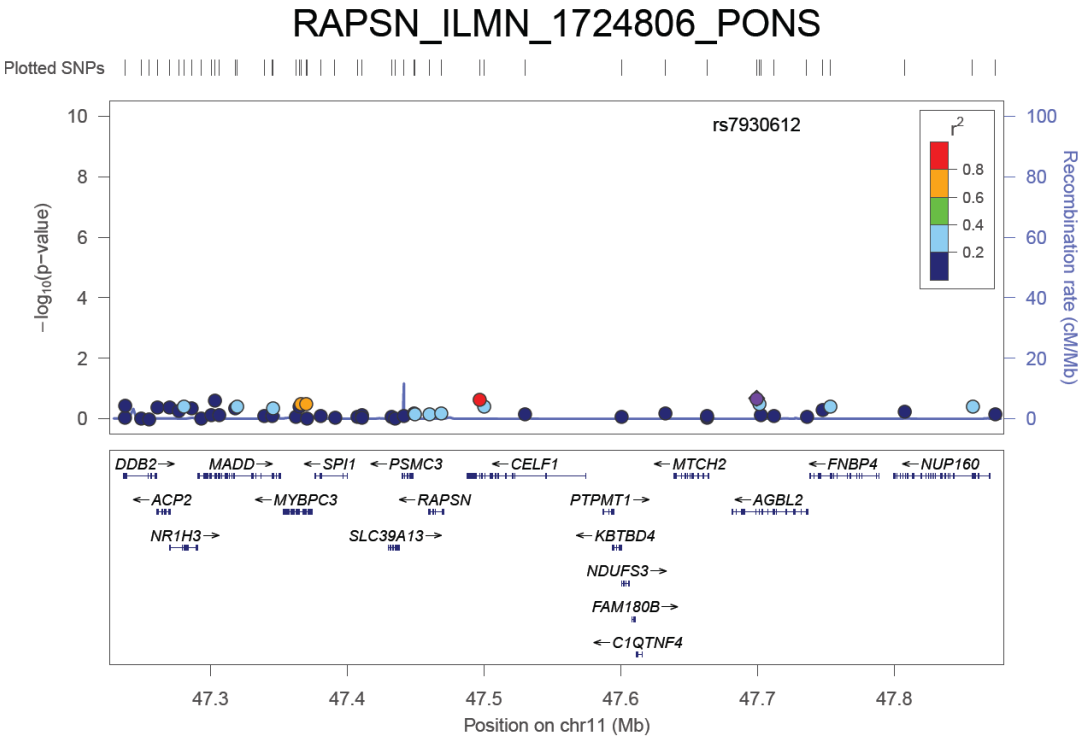
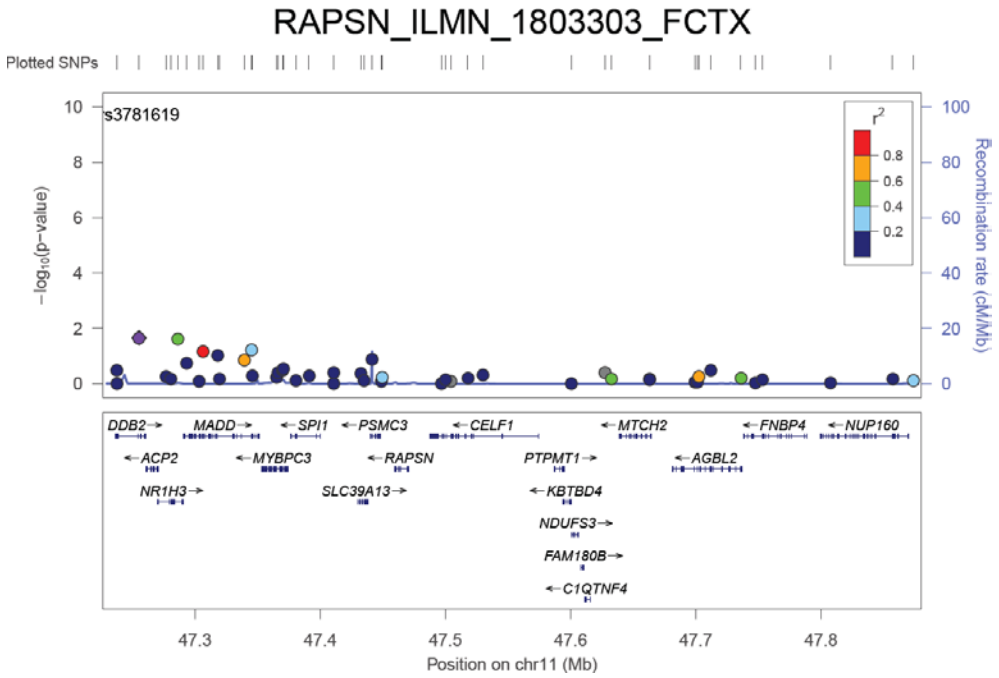
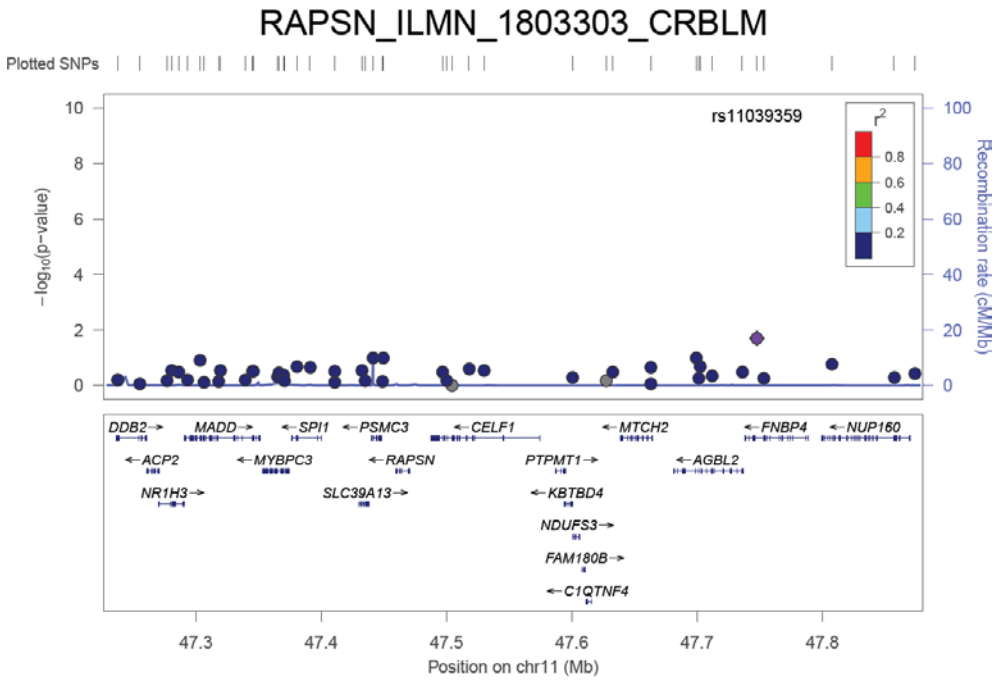

RAPSN_ILMN_1724806_PONS
Plotted SNPs
rs7930612
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S20. Cis association for *RAPSN*-ILMN_1803303 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

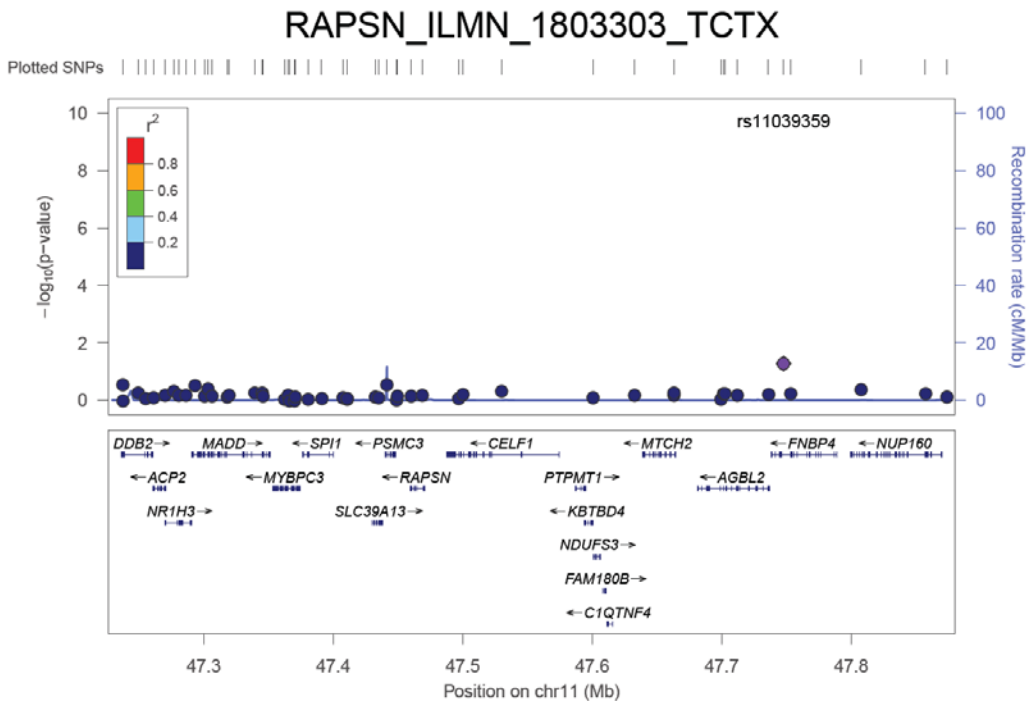
(A)



(B)



(C)


RAPSN_ILMN_1803303_TCTX
Plotted SNPs
rs11039359
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

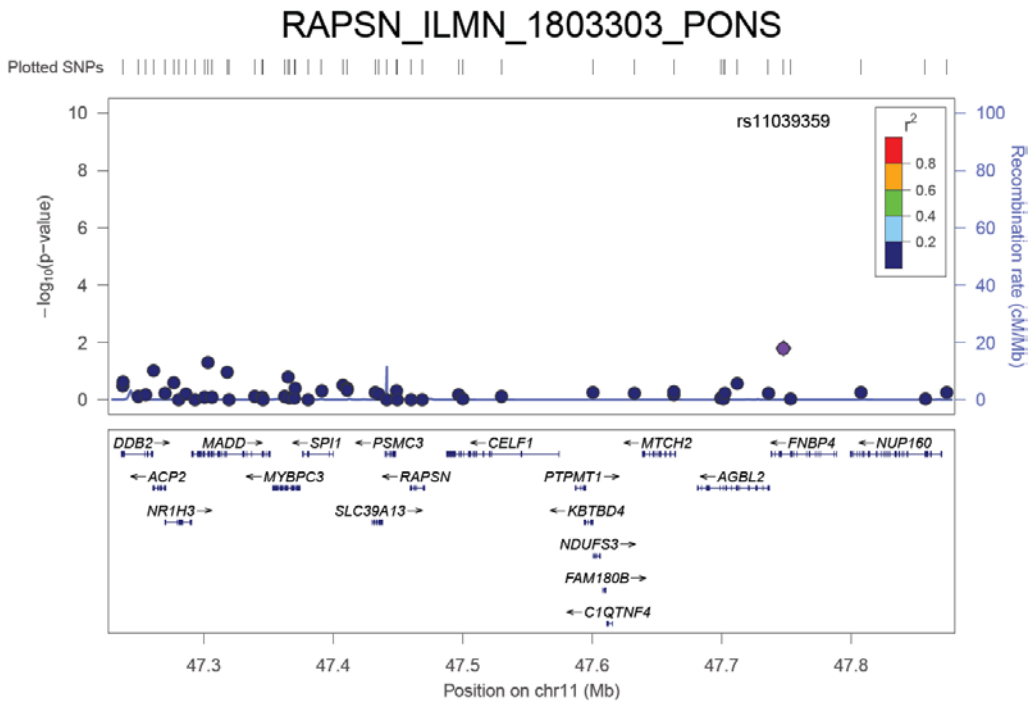
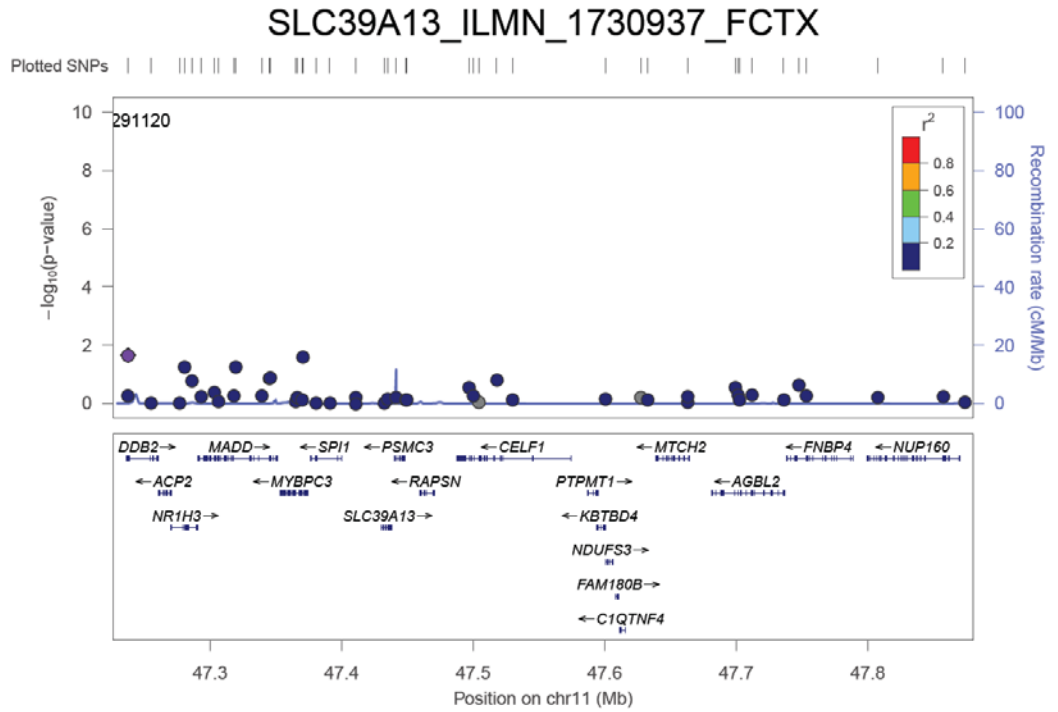
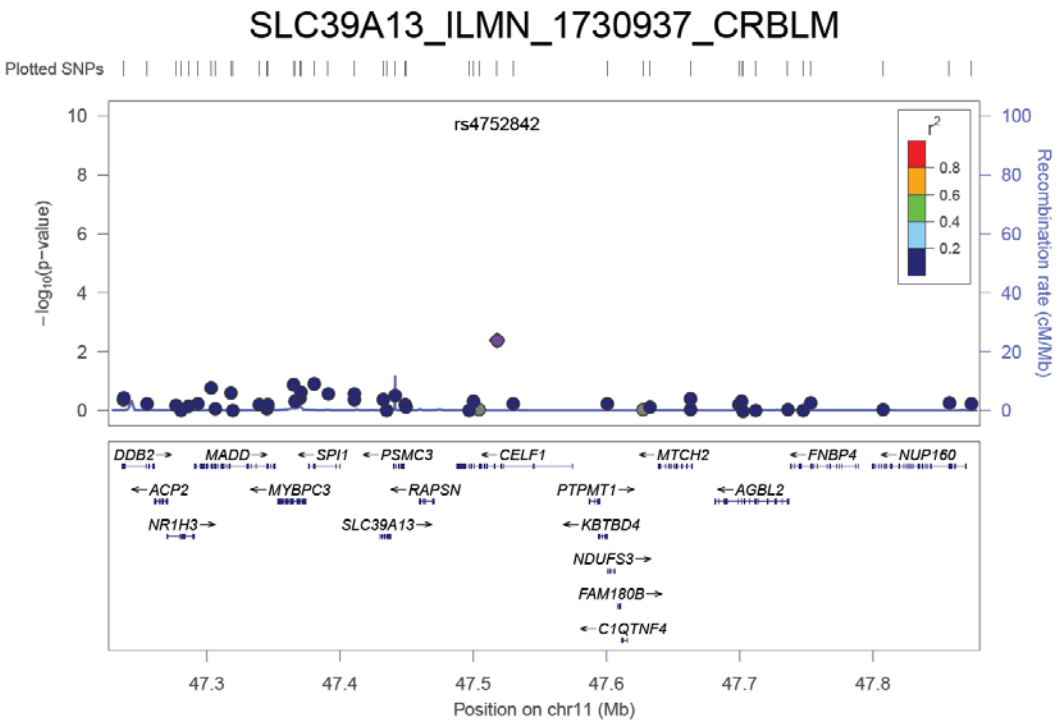

RAPSN_ILMN_1803303_PONS
Plotted SNPs
rs11039359
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S21. Cis association for *RAPSN*-ILMN_1803303 expression levels in (A) FCTX, (B) CRBLM, (C) TCTX, and (D) PONS.

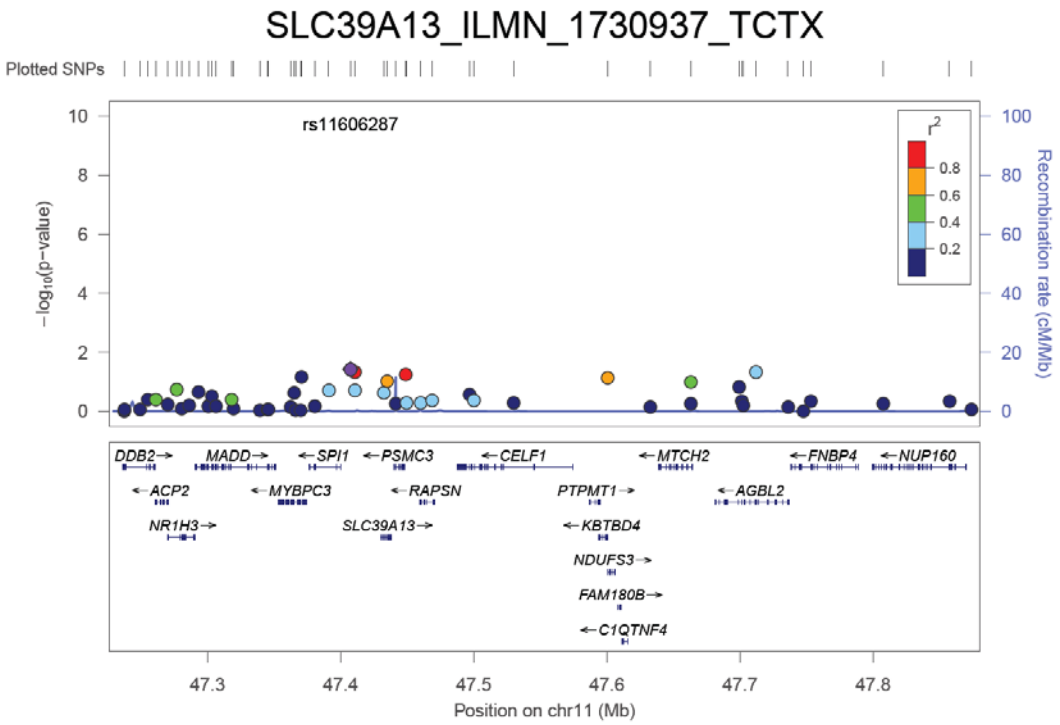
(A)



(B)



(C)


SLC39A13_ILMN_1730937_TCTX
Plotted SNPs
rs11606287
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

(D)

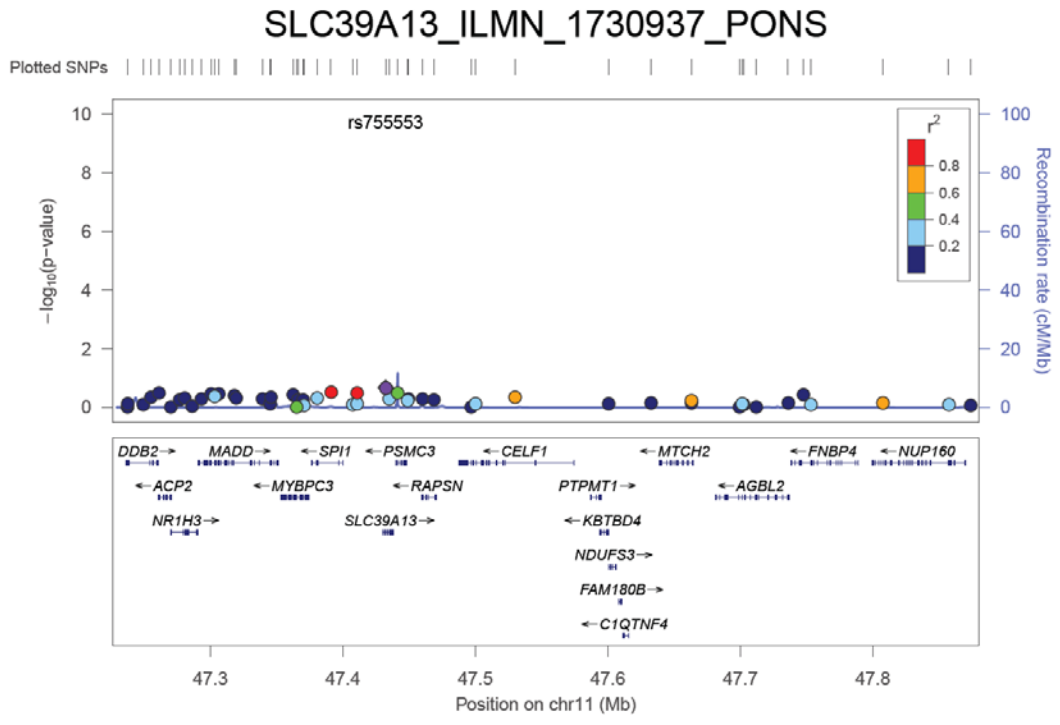
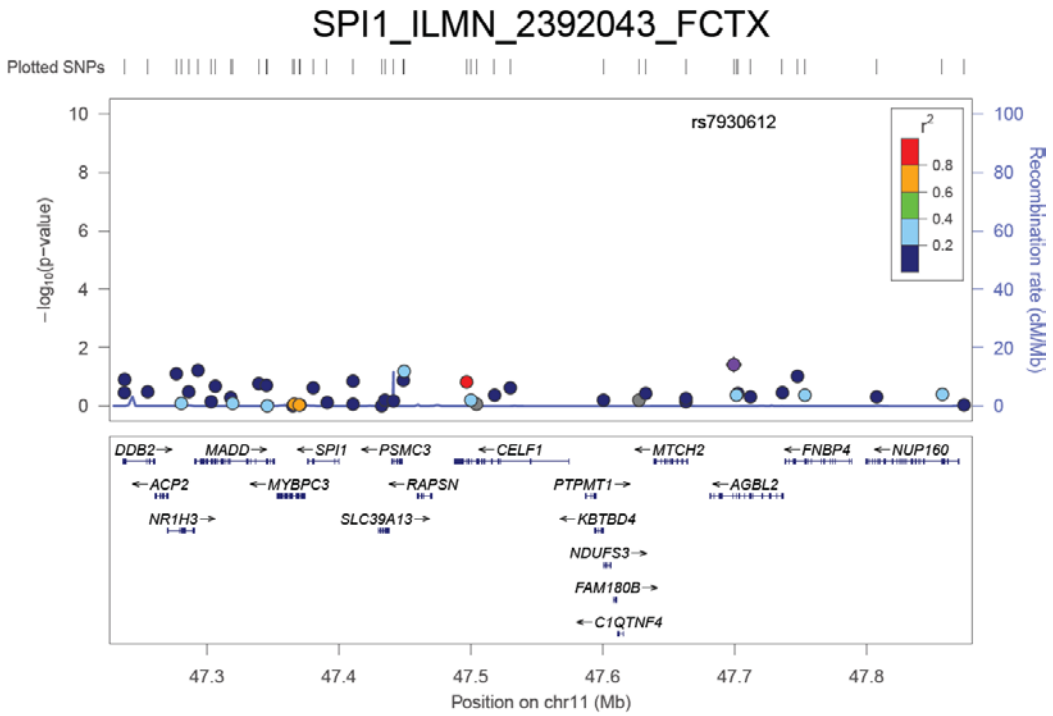
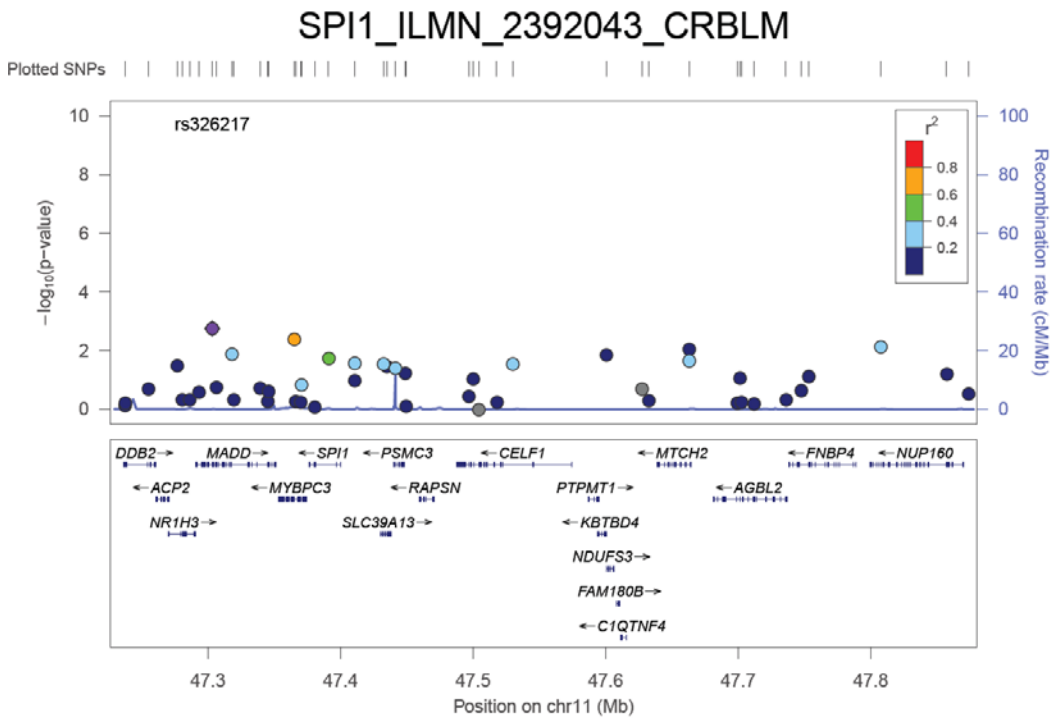

SLC39A13_ILMN_1730937_PONS
Plotted SNPs
rs755553
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DDB2
MADD
SPI1
PSMC3
CELF1
MTCH2
FNBP4
NUP160
ACP2
MYBPC3
RAPSN
PTPMT1
AGBL2
NR1H3
SLC39A13
KBTBD4
NDUFS3
FAM180B
C1QTNF4
47.3
47.4
47.5
47.6
47.7
47.8
Position on chr11 (Mb)

Figure S22. Cis association for *SPI1*-ILMN_2392043 expression levels in (A) FCTX and (B) CRBLM.

(A)



(B)



Chapter 5

Conclusion and future directions

5.1 State of Alzheimer's genetics prior to this work

Alzheimer's disease (AD) is a devastating and irreversible brain disease that affects millions of Americans[25]. AD has become a global public health crisis as the estimated worldwide prevalence is 30 million and is predicted to quadruple by 2050[25]. Currently, there is still no effective treatment for AD. Since many studies have shown a strong genetic component for AD – an estimated heritability of 80%[111], understanding how genetic risk factors affect the disease process will help to identify novel targets for therapeutics.

In the early 1990s, Alzheimer's disease genetic research identified dominantly inherited mutations in three genes (*APP*, *PSEN1*, and *PSEN2*) causing early onset AD[1-4], and one risk gene (*APOE*)[5-7] involved in late-onset AD (LOAD) through genetic linkage studies in families. Until about 2005, most studies examined one or more polymorphisms in a candidate gene to identify novel risk genes due to technical limitations. Even though over 1,000 candidate genes were investigated for AD susceptibility, very few reported genetic association could be replicated, most likely due to false positive associations resulting from population stratification and small sample size (see http://www.alzgene.org). Since 2005, rapid advances in technologies and analytical tools have transformed genetics of complex disorders such as AD[183]. The first advancement was the development of inexpensive and comprehensive genotyping platforms a.k.a "genome-wide association study (GWAS) arrays", which allow simultaneous investigation of millions of single-nucleotide polymorphisms (SNPs) in tens of thousands of individuals. To facilitate the use of GWAS, the HapMap and 1,000 Genomes Project databases catalogued over 20 million SNPs and provided essential resources for SNP queries, genotype phasing, and imputation. Additionally, the use of principal component analysis for matching cases and controls has been routinely used in GWAS to control for population substructure, reducing false-positive findings. Before 2012, several large-scale GWAS in AD were conducted and identified 9 novel loci associated with LOAD: *ATP-binding cassette transporter (ABCA7)*, *bridging integrator protein 1*(*BIN1*), *CD2-associated protein* (*CD2AP*), *sialic acid-binding immunoglobulin (Ig) like lectin* (*CD33*), *clusterin* (*CLU*), *complement receptor 1* (*CR1*), *Ephrin receptor A1* (*EPHA1*), *membrane-spanning 4A gene cluster* (*MS4A4A*/*MS4A4E*/*MS4A6E*), and *phosphatidylinositol –binding clathrin assembly protein* (*PICALM*)[9-11,78,184]. Remarkably, these GWAS have identified robust and replicable genetic variations across independent cohorts ascertained by a variety of methods and individuals diagnosed in innumerable research institutes. However, most common variants identified in GWAS are located in non-coding regions and have very small effects on disease risk (odds ratio [OR] ~ 0.8-1.2). The known susceptibility loci only explain 50% of the heritability for LOAD,

which has an overall heritability of up to 80%[111], indicating that additional variants and genes remain unobserved. Additionally, most of these GWAS susceptibility loci are located in non-coding or gene-dense regions, making it very difficult to identify *bona fide* genes responsible for the association. Linkage disequilibrium (LD) between the GWAS top SNPs with other SNPs also prevents the identification of *bona fide* risk variants for LOAD.

When we started this work, several studies had suggested that low-frequency (minor allele frequency [MAF] 1% to 5%) and rare variants (MAF < 1%) could explain part of the missing heritability because GWAS genotyping arrays have largely ignored these portions of the allele frequency spectrum[185,186]. Over the past five years, dramatic developments in next-generation sequencing (NGS) have made high-throughput sequencing of targeted genomic regions of the human genome in many individuals in a single run both cheap and feasible. Moreover, advances in sequencing-based study design and rare-variant association methods enable us to investigate the role of low-frequency and rare variants in complex diseases.

Even though deep-depth whole-genome sequencing (WGS) of a large number of samples remains very expensive, several alternatives including targeted re-sequencing, whole-exome sequencing (WES), and low-depth WGS have been proposed and widely conducted. Previously, our group performed deep re-sequencing of the pathogenic genes (*APP*, *PSEN1*, and *PSEN2*) in LOAD families and uncovered several novel, rare and pathogenic variants, demonstrating that the impact of the variants in these AD-causing genes in LOAD families is stronger than previously estimated[15,16]. In 2013, two independent groups utilized low-depth WGS, WES, and targeted Sanger sequencing to identify a low-frequency variant p.R47H in *triggering receptor expressed on myeloid cells 2* (*TREM2*), that increases AD risk by 2 fold[12,13]. Not long after, our group performed WES in fourteen large LOAD families and identified a low-frequency variant p.V232M in *phospholipase D3* (*PLD3*) which segregated with disease status in 2 independent families[14]. The association of p.V232M with AD risk was replicated after performing follow-up genotyping in seven case-control studies and demonstrated that the minor allele of p.V232M doubles AD risk[14]. A recent study performed targeted sequencing of *CLU*, *CR1*, and *PICALM* in 96 AD samples, but no significant rare variant association was reported[187]. Another study sequenced *TREM2* coding regions in a Belgian population and found additional coding variants in *TREM2*[96]. Although an enrichment of *TREM2* variants in both AD and FTD patients compared to controls was reported, none of the rare variants were individually significant[96]. To our knowledge, no deep re-sequencing studies in GWAS-identified genes (*ABCA7*, *BIN1*, *CD2AP*, *CLU*, *CD33*,

CR1, *EPHA*, *MS4A4A*/*MS4A4E*/*MS4A6E* , and *PICALM*) and sequencing-identified genes (*PLD3* and *TREM2*) have been reported with a larger sample size than this work (N>4,000).

Not long after we started this work, the International Genomics of Alzheimer's Project (IGAP) was founded to perform statistically rigorous and comprehensive GWAS meta-analysis in the hope of finding additional small effect risk loci for LOAD with participation of the ADGC, CHARGE, and GERAD, and EADI consortia[188]. IGAP conducted meta- and gene-wide analyses of 74,046 individuals and identified 23 loci associated with LOAD, of which 13 were novel. These novel signals are close to the following genes: *Cas scaffolding protein family member 4* (*CASS4*), *CUGBP, Elav-like family member 1* (*CELF1*), *Fermitin family member 2* (*FERMT2*). *Major histocompatibility complex class II, DR beta 5-1* (*HLA-DRB5/DRB1*), *Immunoglobulin heavy variable 1-67* (*IGHV1-67*), *Inositol polyphosphate-5-phosphatase* (*INPP5D*), *myocyte enhancer factor 2C* (*MEF2C*), *NME/NM23 family member 8* (*NME8*), *Protein tyrosine kinase 2 beta* (*PTK2B*), *solute carrier family 24* (*SLC24A4*), *sortilin-related receptor L, A repeats containing* (*SORL1*), *Tumor protein p53 inducible nuclear protein 1*(*TP53INP1*), and *zinc finger, and CW type with PWWP domain 1* (*ZCWPW1*). The IGAP study also identified a suggestive association located in the intergenic region between *TREM2* and *TREML2*, but due to the study design (a meta-analysis), it was not possible to determine whether this intergenic signal was independent of p.*R47H* in *TREM2*. These IGAP-identified loci contain multiple genes within the associated region and therefore the suggested genes may not contain the underlying functional variants. Additionally, the most-significant SNPs within these loci were located in non-coding regions, and thus there is no clear functional impact of these variants linked to AD pathogenesis. It is also possible that multiple functional variants reside within the same locus affecting AD risk independently. Endophenotype-based analyses along with conditional analyses can help to identify the potential drivers of the associations, infer the underlying mechanisms associated with AD, and determine whether there are multiple independent genetic variants affecting AD pathogenesis.

In the past few years, our lab has had great success in using CSF biomarkers as endophenotypes to validate AD genetic risk factors[23], to generate hypotheses regarding the mechanism by which AD risk variants contribute to AD development[18], and to find novel variants associated with age at onset and disease progression[17,19]. Several candidate gene studies examined the association between GWAS top hits and CSF biomarkers and found evidence of association for CSF $A\beta_{42}$ levels with variants in *SORL1*, *CLU*, and *MS4A4A*[164,189]. However, another similar study did not find any evidence of association for any GWAS locus[24]. Our group recently performed the largest CSF

GWAS (1,269 individuals) using CSF tau and tau phosphorylated at threonine 181 ($ptau_{181}$) as endophenoytpes and reported four genome-wide significant loci in *APOE*: at 3q28 between *GEMC1* and *OSTN*, at 9q24.2 within *GLIS3*, and 6p21.1 within the *TREM* gene cluster[22]. However, in this paper none of the SNPs within GWAS-identified loci showed genome-wide significant evidence ($p < 5\times10^{-8}$) with CSF tau and $ptau_{181}$ levels, possibly due to a small sample size and an extremely stringent genome-wide significance cutoff (Bonferroni correction)[22]. Despite the work, several questions remained to be answered: (1) whether the top SNPs in novel GWAS genes, including *CASS4*, *CELF1*, *FERMT2*, *HLA-DRB5/DRB1*, *IGHV1-67*, *INPP5D*, *MEF2C*, *NME8, PTK2B, SLC24A4*, *SORL1*, *TP53INP1*, and *ZCWPW1,* identified in the IGAP meta-analysis[8] are also associated with CSF biomarker levels, (2) whether there are multiple independent signals within the same locus affecting AD, (3) whether the intergenic signal located between *TREM2* and *TREML2* drives the association independently of *TREM2* p.R47H, and (4) whether *TREM2* and *TREML2* affect AD pathogenesis distinctly and independently.

Completion of the ENCODE Project has provided valuable information to identify and characterize functional elements in the human genome and facilitate studies of gene expression and the interpretation of non-coding variants linked to human disease[160,190]. Two related studies examined gene expression from laser-capture microdissected non-tangle-bearing neurons in several brain regions[169,171]. These studies also provided valuable gene expression profiles in anatomically and functionally different brain regions and showed that AD is associated with decreased expression of energy metabolism genes in posterior cingulated neurons[169,171]. More recently, gene-expression levels were used as endophenotypes and abundant expression quantitative trait loci (eQTLs) were found in several human primary tissues[106-108,167]. These eQTL analyses[106-108,167] clearly provide a brain region-specific framework for the identification of regulatory variants and genes, which can translate into better understanding of the functional mechanism for genetic variants. However, very little work has been done to investigate cis-eQTL effects on expression levels for the GWAS-identified genes. A recent study has examined *ABCA7*, *BIN1*, *CLU*, *CR1*, *PICALM*, and *MS4A6A*/*MS4A6E* neighboring regions but did not find eQTL evidence[109]. Given that a recent study[142] and our CSF and bioinformatic analyses suggest several SNPs within the *CELF1* region are regulatory variants of nearby genes, much work is required to perform cis-eQTL analyses for these GWAS-identified regions in order to identify potential regulatory variants which can explain the GWAS signals.

Several studies had been done to functionally characterize the *TREM2* gene, an immune phagocytic receptor expressed on brain microglia known to trigger phagocytosis and regulate the inflammatory response. A previous study used microarray and laser microdissection of beta amyloid (Aβ) plaque-associated areas in an animal model of AD and found that *TREM2* is differentially expressed in Aβ plaque-associated versus Aβ plaque-free tissue [91]. A recent study has demonstrated that TREM2 is essential for the microglial response to Aβ deposition while a 50% decrease in TREM2 expression has no effect on Aβ plaque burden[137]. Another study characterized several *TREM2* variants, including, p.C36A, p.Y38C, p.R47H, and p.T66M, and showed that p.T66M but not p.R47H impairs *TREM2* trafficking to the cell surface[191]. Our recent work identified 16 rare coding variants in *TREM2*, 6 of which (p.R52H, p.R136W, p.E151K, p.W191X, p.E202D, and p.H215Q) were not discovered in previous studies[12,13,96]. Additionally, nine variants (p.R52H, p.T66M, p.R136W, p.R136Q, p.H157Y, p.W191X, p.E202D, p.H215Q and p.T223I) were only found in AD cases. However, no work has been done to functionally characterize these novel and potentially functional *TREM2* variants in cell systems.

5.2 Dissertation work

Even though recent sequencing efforts have shown that p.V232M in *PLD3* and p.R47H in *TREM2* increase AD risk by 2 fold[12-14], additional risk variants within *PLD3* and *TREM2* remained undiscovered. Therefore we hypothesized that if *PLD3* and *TREM2* are *bona fide* AD genes, they will carry additional low-frequency and rare variants. We performed deep re-sequencing of exonic and flanking intronic sequence in order to identify novel functional variants and test for association with AD risk. We sequenced 4,387 European Americans (2,363 AD cases and 2,024 controls) and 302 African Americans (130 cases and 172 controls) for *PLD3* and 3,730 European Americans (2,082 AD cases and 1,648 controls) and 336 African Americans (204 cases and 132 controls) for *TREM2*. We demonstrated that rare variants in *PLD3* and *TREM2* are more frequently seen in cases than in controls of European descent (*TREM2*: 6.7% in cases and 2.7% in controls; *PLD3*: 8.0% in cases and 3.1% in controls). Single-variant analyses showed that p.M6R ($p=0.02$; odds ratio [OR]=7.73 [1.09~61]) and p.A442A ($p=3.78\times10^{-7}$; OR=2.21 [1.58~2.8]) in *PLD3* and p.R62H ($p=2.36\times10^{-4}$; OR=2.36 [1.47~3.80]) in *TREM2* are significantly associated with AD risk in addition to p.V232M in *PLD3* and p.R47H in *TREM2*. Gene-based tests demonstrated that *PLD3* ($P_{SKAT\text{-}O}=1.44\times10^{-11}$; OR=2.75 [2.05~3.68]) and *TREM2* ($P_{SKAT\text{-}O}=5.37\times10^{-7}$; OR=2.55 [1.62~3.87]) are genome-wide significantly associated with AD. The associations for *PLD3* and *TREM2* rare variants with AD risk are still highly significant after excluding p.V232M ($P_{SKAT\text{-}O}=1.5\times10^{-8}$; OR=2.58 [1.87~3.57]) and p.R47H ($P_{SKAT\text{-}}$

$_O$=7.72×10^{-5}; OR=2.47 [1.62~3.87]) respectively, which indicates that additional *PLD3* and *TREM2* variants affect AD risk. However, we did not find evidence of association for *TREM2* variants with AD risk in African Americans at either the gene-level or the SNP-level which may be due to our small sample size. Even though *TREM2* p.R47H has bebn found to be associated with AD risk in Caucasian[12,13], Spanish[125], and French[126] populations, no significant association of TREM2 R47H was found in several Asian populations[192-194]. Like Asian populations, the reason why we did not see any significant association in our dataset may be because *TREM2* variants are too rare for any meaningful power in the African Americans. In this work, we also confirmed that at least three *TREM2* transcripts are expressed in human brains, including one encoding a soluble form of TREM2, which suggests that some soluble TREM2 results from alternative splicing rather than cleavage of membrane-associated TREM2.

Since the most-significant SNPs in GWAS-identified susceptibility loci had very small effect size and only explained 50% of the phenotypic heritability for AD (overall heritability of up to 80%), we hypothesized that low-frequency and rare coding variants within GWAS-identified loci have greater effects on AD risk compared to most-significantly associated common variants found in GWAS. We performed pooled-DNA sequencing of exonic and flanking intronic sequences of GWAS-identified genes, including *ABCA7*, *BIN1*, *CD2AP*, *CD33*, *CLU*, *CR1*, *EPHA1*, *MS4A4A*, and *PICALM*, in 3,730 European Americans (2,082 AD cases and 1,648 controls) and 336 African Americans (204 cases and 132 controls). Even though this project has not been completed, we have already validated 90 coding variants in GWAS-identified genes, 66 of which were not annotated in the Exome Variant Server –a database which consists of whole-exome sequencing (WES) data in 2,203 African American and 4,300 European American unrelated individuals. Bioinformatic analyses predict that 56 of the confirmed variants (62.2%) are damaging. Nucleotide conservation analyses suggest that 57 of the validated variants are under evolutionary constraint (a GERP score>2), which implies that a large proportion of rare coding variants in GWAS-identified genes are potentially functional. Together, these findings suggest that deep re-sequencing is an effective strategy to identify additional functional variants in AD-associated genes.

Even though IGAP identified 23 loci associated with AD risk through GWAS meta-analysis and gene-wide analyses, the mechanisms by which most of these loci affect the molecular pathways leading to AD remain unknown. Therefore, we undertook endophenotype-based analyses to determine whether these loci are also associated with CSF $A\beta_{42}$ and $ptau_{181}$ levels. We combined CSF biomarker datasets from several studies (N=2,036)

and performed single-variant and set-based analyses for each locus. In the *APOE* locus, rs769449 is genome-wide significantly associated with both CSF $A\beta_{42}$ and $ptau_{181}$ levels. Furthermore, as previously reported in a smaller dataset, the association between rs769449 and CSF $ptau_{181}$ levels is only partially explained by differences in CSF $A\beta_{42}$ levels. We found evidence of association ($p<0.05$) for the top IGAP SNPs, rs4147929 (*ABCA7)*, rs17125944 (*FERMT2)*, and rs35349669 (*INPP5D*) with CSF $A\beta_{42}$ levels, and for rs6656401 (*CR1)*, rs17125944 (*FERMT2)*, rs190982 (*MEF2C)*, rs10792832 (*PICALM)*, rs28834970 (*PTK2B)*, and rs11218343 (*SORL1)* with CSF $ptau_{181}$ levels. Our locus-specific analyses suggested that after multiple test correction, rs7937331, within the *CELF1* fine-mapping region, is associated with CSF $A\beta_{42}$ levels and AD risk and tags the same signal as the IGAP top SNP, rs10838725. Additionally, rs62003531, located in the intronic region of *FERMT2*, tags the same association as the IGAP top SNP (rs17125944) and is associated with CSF $A\beta_{42}$ levels. The association for the *CELF1* and *FERMT2* fine-mapping regions with CSF $A\beta_{42}$ levels was confirmed in set-based analyses. None of the SNPs within the IGAP-identified AD risk loci except the *APOE* locus are significantly associated with CSF $ptau_{181}$ levels after multiple test corrections. This may be due to lack of statistical power. We also tested whether IGAP top SNPs and CSF top SNPs have any regulatory potential. We demonstrated that the majority of GWAS top SNPs have no significant regulatory potential and are unlikely to be the functional variants for AD risk. However, RegulomeDB predicts that several proxy SNPs in LD with rs7937331 in *SLC39A13* may be cis-acting expression quantitative trait loci (eQTLs) for nearby genes and are located in transcription factor binding sites. Together our results suggest that AD risk variants may not necessarily be associated with CSF biomarker levels, and that GWAS-identified noncoding variants may affect AD risk through regulatory mechanisms.

The IGAP study also identified a suggestive association ($p<10^{-6}$) located in the intergenic region between *TREM2* and *TREML2,* but due to the study design, it was not possible to determine whether this intergenic signal was independent of p.*R47H* in *TREM2*. As a result, we performed comprehensive analyses using WES data, CSF biomarker analyses, and meta-analyses (16,254 cases and 20,052 controls) to demonstrate that the AD risk GWAS association is likely driven by a *TREML2* missense variant p.S144G (rs3747742) and that this association is independent of *TREM2* p.R47H risk for AD. Additionally, we demonstrated that the protective role of *TREML2* in AD is independent of the role of *TREM2* gene as a risk factor for AD. Moreover, advances in sequencing-based study design and rare-variant association methods enable us to investigate the role of low-frequency and rare variants in complex diseases.

Since no work had been done to functionally characterize novel *TREM2* variants *in vitro*, we tested the hypothesis that some of these novel *TREM2* variants may affect TREM2 trafficking to the cell surface thereby interfering with trem2 signaling. To test this hypothesis, we introduced *TREM2* variants into TREM2-DAP12 cDNA constructs using site-directed mutagenesis and measured cell surface expression using an anti-TREM2 antibody. Our flow cytometry analyses suggest that p.T66M and p.R136W variants in *TREM2* robustly impact cell surface expression. However, we did not find any differences in cell surface expression when comparing p.R47H and p.R62H to the wild type (WT). This suggests that p.R47H and p.R62H may affect AD through other mechanisms. These results are in line with a recent publication which shows that *TREM2* p.T66M variant reduces the cell surface expression robustly while *TREM2* p.R47H only slightly affect cell surface expression compared to WT[191].

Finally, we sought to determine whether polymorphisms within the *CELF1* fine-mapping region are eQTLs for nearby genes. We performed cis-eQTL analysis for mRNA expression levels in several brain regions using four publicly available datasets to identify genetic determinants of gene expression in human brains. We found that all of the expression-associated SNPs, including rs7124681, rs10838738, rs2290850, and rs755553, are in tight LD with rs7937331, the top CSF $A\beta_{42}$ SNP in the *CELF1* fine-mapping region. Conditional analyses suggest that there is only one independent driver within this region and the minor allele of the underlying causal variant mediates *C1QTNF4* expression levels. Additionally, we found evidence of differential expression in *C1QTNF4* and *SPI1* transcripts between AD cases and controls in human brains. Overall, these results illustrated that the underlying causal variants and genes may not be the gene originally identified in GWAS studies. Our data suggested the causal variants within the *CELF1* fine-mapping region mediate differences in *C1QTNF4* expression levels and may affect AD risk by affecting Aβ biology. These findings provide additional evidence that genes involved in the inflammatory response play an important role in AD pathogenesis.

5.3 Future directions

While recent deep sequencing studies of AD GWAS genes do not report any positive results[96,187], several recent papers clearly suggest that a very large sample size (N ~ 4,000) is required for the detection of a significant rare-variant association[14,97,195]. Therefore, the sample size of this work (N > 4000), which is close to the sample size of our recently published sequencing papers[14,97], provides enough statistical power to identify significant rare-variant associations. Additionally, our preliminary sequencing efforts have identified a large number of rare coding variants within GWAS-identified variants and over 60% of these variants were predicted to be damaging by more than two

bioinformatic algorithms. This implies that a large proportion of rare coding variants in GWAS-identified genes are potentially functional. In the next few months, our group will finish all the genotyping, perform single-variant and gene-based analyses, conduct bioinformatic predictions, and eventually publish the results. Functional analyses will be required to determine the functional impact of rare variants within AD-associated genes *in vitro* and *in vivo*.

Due to the small sample size (204 cases and 132 controls), we did not find any significant association for *TREM2* variants in African Americans. To better determine whether rare *TREM2* coding variants are involved in AD pathogenesis in African Americans, we are collaborating with researchers at the Mayo Clinic (Mayo), the Indiana University School of Medicine (Indiana), the Emory University, and the Alzheimer's Disease Genetic Consortium (see **Figure 1** for the study design) to increase our current sample size. We have combined and analyzed our genotype data with data from Mayo (183 cases and 351 controls) and Indiana (167 cases and 1,341 controls). With the combined data, we identified thirteen coding variants in *TREM2* (p.R47H, p.R62H, p.T66M, p.D87N, p.T96K, p.T96M, p.A105V, p.P144P, p.E151K, p.H157Y, p.W191X, p.E202D, and p.L211P; **Table 1**), three (p.E151K, p.W191X, and p.L211P; **Table 1**) of which are non-synonymous variants that achieved suggestive association ($p<0.15$) and risk ORs (1.17~3.32) in a cohort of modest sample size (554 AD cases and 1,824 controls). Two (p.W191X and p.L211P) of these 3 non-synonymous variants are in tight LD (D'=1) with the African American (AA) GWAS *TREM2* hit (rs7748513; OR ± SE 1.16±0.05; p=0.001). This AA GWAS *TREM2* hit is an intronic variant and is unlikely to be functional based on RegulomeDB (RegulomeDB score=7). Thus functional studies of these non-synonymous variants can provide insight into the potential mechanism linking these variants to AD etiology.

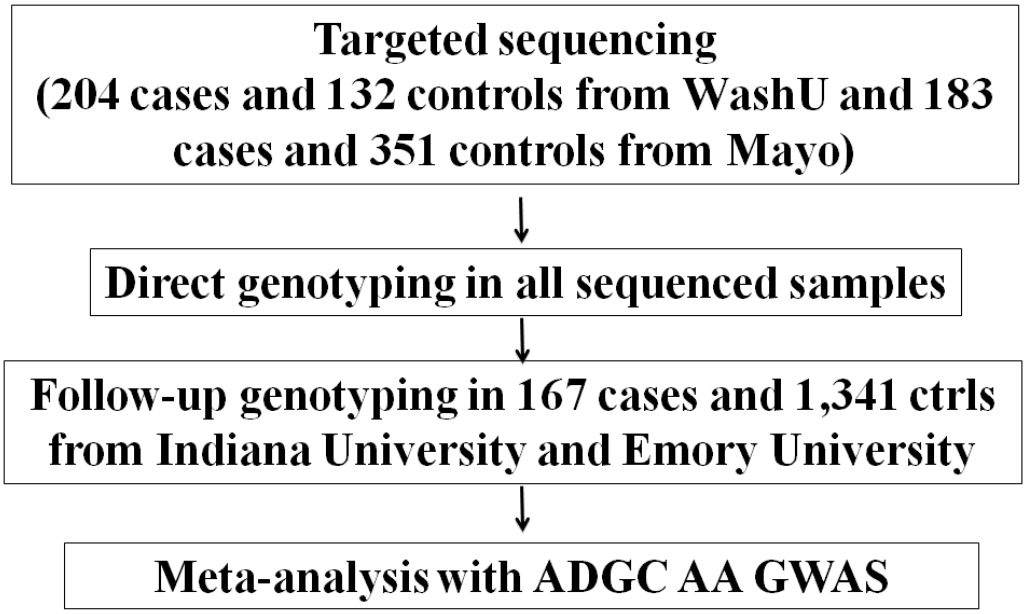


Figure 1: Schematic of the study design for assessing TREM2 association with AD risk in African Americans

Table 1. Rare variants found in African Americans (WashU+Mayo+Indiana)*

			AD Cases		Controls					
Variant	SNP	Position	No. of Cases	No. of Carriers	No. of Controls	No. of Carriers	p	OR (95% CI)	SIFT	PolyPhen
p.R47H	rs75932628	6:41129252	537	1	1733	3	1	1.12 (0.12-10.82)	Tolerated	Damaging
p.R62H	rs143332484	6:41129207	552	3	1821	8	0.72	1.29 (0.34-4.88)	Tolerated	Benign
p.T66M	rs201258663	6:41129195	203	1	132	0	1	NA	Damaging	Damaging
p.D87N	rs142232675	6:41129133	533	1	1682	2	0.55	1.65 (0.15-18.21)	Tolerated	Damaging
p.T96K	rs2234253	6:41129105	381	91	479	112	0.82	1.03 (0.77-1.38)	Damaging	Damaging
p.T96M	NA	6:41129105	160	0	351	1	1	0		
p.A105V	rs145080901	6:41129078	353	1	1336	4	1	1.01 (0.11-9.07)	Damaging	Damaging
p.P144P	NA	6:41127580	156	1	1	0	1	NA	Tolerated	NA
p.E151K	rs79011726	6:41127561	532	3	1685	3	0.14	3.32 (0.67-16.46)	NA	Damaging
p.H157Y	rs2234255	6:41127543	203	1	132	0	1	NA	Damaging	Damaging
p.W191X	rs2234258	6:41126429	515	42	1677	104	0.11	1.37 (0.95-1.97)	NA	NA
p.E202D	NA	6:41126395	204	0	132	1	0.42	0	Damaging	NA
p.L211P	rs2234256	6:41126655	529	130	1679	374	0.15	1.17 (0.95-1.44)	Tolerated	Benign

* Samples from the Washington University (WashU), the Mayo clinic (Mayo), and the Indiana University (Indiana) were combined and analyzed together.

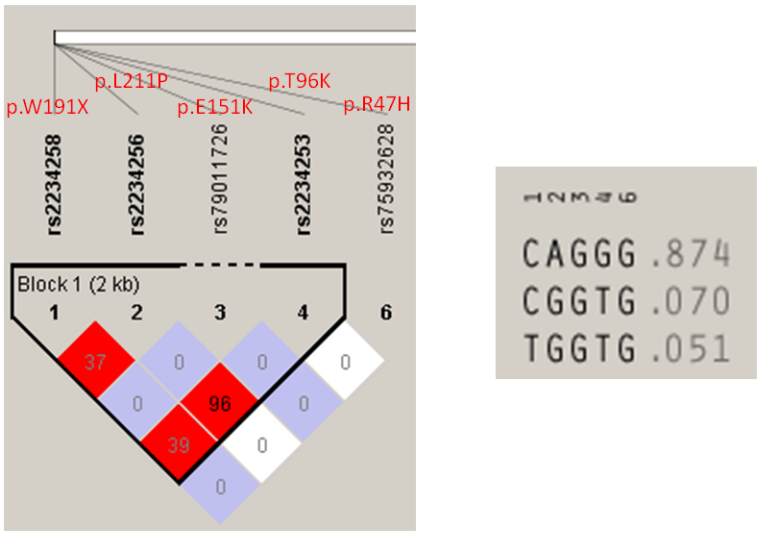


Figure 2: Linkage and haplotype analysis for p.R47H (rs75932628), p.T96K (rs2234253), p.E151K (rs79011726), p.W191X (rs2234258) and p.L211P (rs2234256). P.T96K, p.W191X, and p.L211P are in linkage disequilibrium. D' ranges from 0 to 1 (white to red) between SNPs and numbers indicated R^2 between SNPs.

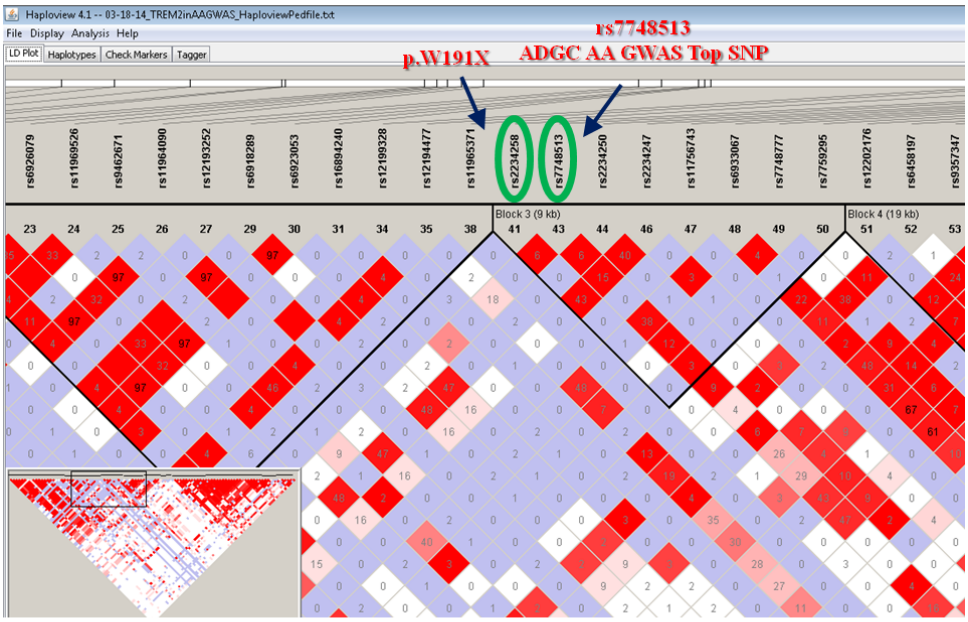


Figure 3: Linkage analysis for p.W191X (rs2234258) and rs7748513. p.W191X is in tight linkage disequilibrium with rs7748513, the top *TREM2* SNP in a recent GWAS (1,968 AD cases and 3,928 controls of African American descent; Reitz C et al. NEJM 2013). D' ranges from 0 to 1 (white to red) between SNPs and numbers indicated R^2 between SNPs.

Several recent papers have shown that loss of a *TREM2* allele alters microglial response to Aβ in a mouse model[137] and that *TREM2* mutations can impair cell surface trafficking and phagocytosis[191], providing support for the hypothesis that AD-associated *TREM2* variant are partial loss-of-function alleles. However, similar functional studies need to be performed for the novel *TREM2* variants identified in recent publications[96,97]. Our preliminary and replication results have shown that T.66M and p.R136W variants in *TREM2* robustly impact cell surface expression, but we did not find any differences in cell surface expression comparing p.R47H and p.R62H to the wild type (WT). This suggests that p.R47H and p.R62H may affect AD through other mechanisms. Stable cell lines should be used to determine the long-term cell surface expression of TREM2 variants. TREM2 trafficking to the cell surface is important for without it, TREM2 cannot bind to its endogenous ligand and thus cannot transduce its intracellular signal properly. Follow-up studies are ongoing to distinguish between expression and trafficking defects by measuring TREM2 and DAP12 RNA and protein levels inside the cell. Additionally, since our functional studies found that *TREM2* p.R136W has very robust effects on cell surface expression, it would be interesting to test whether *TREM2* p.R136Q[21], a different codon change in the same position has a similar functional impact. In other words, we can determine whether the variability in TREM2 cell surface expression is directly correlated to a codon specific, Tryptophan substitution at 136, or whether conservation of this position alone is important in regulating TREM2 cell surface transport.

In the past 5 years, dramatic advances in genotyping, sequencing and analytical areas have driven substantial progress toward understanding the genetic architecture of AD. GWAS have reported more than 20 susceptibility loci for AD[8,10,11,196]. Except *APOE*, common variants within these loci only influence AD risk modestly but they reveal several important disease pathways. Additionally, the use of biomarkers as endophenotypes has been detected additional genetic loci affecting AD risk, tau metabolism, Aβ metabolism, and neuroimaging phenotypes[19,22-24,163,197]. Recently, massively parallel DNA sequencing in small discovery datasets has already led to identification of rare large effect variants in several promising genes for AD[12-14]. Currently, several groups are performing large scale WES or WGS analyses in unrelated AD cases and controls and in families with multiply affected family members in order to identify novel AD genes. WES or WGS data from more than 20,000 individuals are expected to be available before the end of 2014 and subsequently additional AD risk genes are likely to be found. The integration of GWAS, WES/WGS, and other large scale datasets such as transcriptomics and

epigenomics data from human brain tissues will provide better understanding of disease mechanisms, leading to effective targets for therapeutics.

REFERENCES

1. Goate, A. *et al.* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **349**, 704-6 (1991).
2. Levy-Lahad, E. *et al.* Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* **269**, 973-7 (1995).
3. Sherrington, R. *et al.* Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375**, 754-60 (1995).
4. Rogaev, E.I. *et al.* Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* **376**, 775-8 (1995).
5. Coon, K.D. *et al.* A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *The Journal of clinical psychiatry* **68**, 613-8 (2007).
6. Corder, E.H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921-3 (1993).
7. Strittmatter, W.J. *et al.* Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 1977-81 (1993).
8. Lambert, J.C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452-8 (2013).
9. Lambert, J.C. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature genetics* **41**, 1094-9 (2009).
10. Hollingworth, P. *et al.* Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nature genetics* **43**, 429-35 (2011).
11. Naj, A.C. *et al.* Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nature genetics* **43**, 436-41 (2011).
12. Guerreiro, R. *et al.* TREM2 variants in Alzheimer's disease. *N Engl J Med* **368**, 117-27 (2013).
13. Jonsson, T. *et al.* Variant of TREM2 associated with the risk of Alzheimer's disease. *N Engl J Med* **368**, 107-16 (2013).
14. Cruchaga, C. *et al.* Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature* **505**, 550-4 (2014).
15. Cruchaga, C. *et al.* Rare variants in APP, PSEN1 and PSEN2 increase risk for AD in late-onset Alzheimer's disease families. *PloS one* **7**, e31039 (2012).
16. Jin, S.C. *et al.* Pooled-DNA sequencing identifies novel causative variants in PSEN1, GRN and MAPT in a clinical early-onset and familial Alzheimer's disease Ibero-American cohort. *Alzheimer's research & therapy* **4**, 34 (2012).
17. Kauwe, J.S. *et al.* Variation in MAPT is associated with cerebrospinal fluid tau levels in the presence of amyloid-beta deposition. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 8050-4 (2008).
18. Kauwe, J.S. *et al.* Alzheimer's disease risk variants show association with cerebrospinal fluid amyloid beta. *Neurogenetics* **10**, 13-7 (2009).
19. Cruchaga, C. *et al.* SNPs associated with cerebrospinal fluid phospho-tau levels influence rate of decline in Alzheimer's disease. *PLoS genetics* **6**(2010).
20. Kauwe, J.S. *et al.* Extreme cerebrospinal fluid amyloid beta levels identify family with late-onset Alzheimer's disease presenilin 1 mutation. *Annals of neurology* **61**, 446-53 (2007).
21. Cruchaga, C. *et al.* Cerebrospinal fluid APOE levels: an endophenotype for genetic studies for Alzheimer's disease. *Human molecular genetics* **21**, 4558-4571 (2012).
22. Cruchaga, C. *et al.* GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease. *Neuron* **78**, 256-68 (2013).
23. Kauwe, J.S. *et al.* Validating predicted biological effects of Alzheimer's disease associated SNPs using CSF biomarker levels. *J Alzheimers Dis* **21**, 833-42 (2010).
24. Kauwe, J.S. *et al.* Fine mapping of genetic variants in BIN1, CLU, CR1 and PICALM for association with cerebrospinal fluid biomarkers for Alzheimer's disease. *PloS one* **6**, e15918 (2011).
25. Holtzman, D.M., Morris, J.C. & Goate, A.M. Alzheimer's disease: the challenge of the second century. *Sci Transl Med* **3**, 77sr1 (2011).

26. Thies, W., Bleiler, L. & Alzheimer's, A. 2013 Alzheimer's disease facts and figures. *Alzheimers Dement* **9**, 208-45 (2013).
27. Hardy, J. & Selkoe, D.J. Medicine - The amyloid hypothesis of Alzheimer's disease: Progress and problems on the road to therapeutics. *Science* **297**, 353-356 (2002).
28. Fryer, J.D. *et al.* Apolipoprotein E markedly facilitates age-dependent cerebral amyloid angiopathy and spontaneous hemorrhage in amyloid precursor protein transgenic mice. *J Neurosci* **23**, 7889-96 (2003).
29. McGeer, P.L. & McGeer, E.G. The inflammatory response system of brain: implications for therapy of Alzheimer and other neurodegenerative diseases. *Brain Res Brain Res Rev* **21**, 195-218 (1995).
30. Lucin, K.M. & Wyss-Coray, T. Immune activation in brain aging and neurodegeneration: too much or too little? *Neuron* **64**, 110-22 (2009).
31. Tagliavini, F., Giaccone, G., Frangione, B. & Bugiani, O. Preamyloid deposits in the cerebral cortex of patients with Alzheimer's disease and nondemented individuals. *Neurosci Lett* **93**, 191-6 (1988).
32. Giaccone, G. *et al.* Down patients: extracellular preamyloid deposits precede neuritic degeneration and senile plaques. *Neurosci Lett* **97**, 232-8 (1989).
33. Goedert, M., Wischik, C.M., Crowther, R.A., Walker, J.E. & Klug, A. Cloning and sequencing of the cDNA encoding a core protein of the paired helical filament of Alzheimer disease: identification as the microtubule-associated protein tau. *Proc Natl Acad Sci U S A* **85**, 4051-5 (1988).
34. Wischik, C.M. *et al.* Structural characterization of the core of the paired helical filament of Alzheimer disease. *Proc Natl Acad Sci U S A* **85**, 4884-8 (1988).
35. Wischik, C.M. *et al.* Isolation of a fragment of tau derived from the core of the paired helical filament of Alzheimer disease. *Proc Natl Acad Sci U S A* **85**, 4506-10 (1988).
36. Grundke-Iqbal, I. *et al.* Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology. *Proc Natl Acad Sci U S A* **83**, 4913-7 (1986).
37. Priller, C. *et al.* Synapse formation and function is modulated by the amyloid precursor protein. *J Neurosci* **26**, 7212-21 (2006).
38. Turner, P.R., O'Connor, K., Tate, W.P. & Abraham, W.C. Roles of amyloid precursor protein and its fragments in regulating neural activity, plasticity and memory. *Prog Neurobiol* **70**, 1-32 (2003).
39. Duce, J.A. *et al.* Iron-export ferroxidase activity of beta-amyloid precursor protein is inhibited by zinc in Alzheimer's disease. *Cell* **142**, 857-67 (2010).
40. Levy, E. *et al.* Mutation of the Alzheimer's disease amyloid gene in hereditary cerebral hemorrhage, Dutch type. *Science* **248**, 1124-6 (1990).
41. Davis, J. & Van Nostrand, W.E. Enhanced pathologic properties of Dutch-type mutant amyloid beta-protein. *Proc Natl Acad Sci U S A* **93**, 2996-3000 (1996).
42. Grabowski, T.J., Cho, H.S., Vonsattel, J.P., Rebeck, G.W. & Greenberg, S.M. Novel amyloid precursor protein mutation in an Iowa family with dementia and severe cerebral amyloid angiopathy. *Annals of neurology* **49**, 697-705 (2001).
43. Van Nostrand, W.E., Melchor, J.P., Cho, H.S., Greenberg, S.M. & Rebeck, G.W. Pathogenic effects of D23N Iowa mutant amyloid beta -protein. *The Journal of biological chemistry* **276**, 32860-6 (2001).
44. Van Nostrand, W.E. *et al.* Protease nexin-II, a potent antichymotrypsin, shows identity to amyloid beta-protein precursor. *Nature* **341**, 546-9 (1989).
45. Smith, R.P., Higuchi, D.A. & Broze, G.J., Jr. Platelet coagulation factor XIa-inhibitor, a form of Alzheimer amyloid precursor protein. *Science* **248**, 1126-8 (1990).
46. Thinakaran, G. & Koo, E.H. Amyloid precursor protein trafficking, processing, and function. *J Biol Chem* **283**, 29615-9 (2008).
47. Chartier-Harlin, M.C. *et al.* Early-onset Alzheimer's disease caused by mutations at codon 717 of the beta-amyloid precursor protein gene. *Nature* **353**, 844-6 (1991).
48. Murrell, J., Farlow, M., Ghetti, B. & Benson, M.D. A mutation in the amyloid precursor protein associated with hereditary Alzheimer's disease. *Science* **254**, 97-9 (1991).
49. Suzuki, N. *et al.* An increased percentage of long amyloid beta protein secreted by familial amyloid beta protein precursor (beta APP717) mutants. *Science* **264**, 1336-40 (1994).
50. Haass, C., Hung, A.Y., Selkoe, D.J. & Teplow, D.B. Mutations associated with a locus for familial Alzheimer's disease result in alternative processing of amyloid beta-protein precursor. *The Journal of biological chemistry* **269**, 17741-8 (1994).
51. Jarrett, J.T., Berger, E.P. & Lansbury, P.T., Jr. The carboxy terminus of the beta amyloid protein is critical for the seeding of amyloid formation: implications for the pathogenesis of Alzheimer's disease. *Biochemistry* **32**, 4693-7 (1993).

52. Iwatsubo, T. *et al.* Visualization of A beta 42(43) and A beta 40 in senile plaques with end-specific A beta monoclonals: evidence that an initially deposited species is A beta 42(43). *Neuron* **13**, 45-53 (1994).
53. Mullan, M. *et al.* A pathogenic mutation for probable Alzheimer's disease in the APP gene at the N-terminus of beta-amyloid. *Nature genetics* **1**, 345-7 (1992).
54. Rovelet-Lecrux, A. *et al.* APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature genetics* **38**, 24-6 (2006).
55. Nilsberth, C. *et al.* The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced Abeta protofibril formation. *Nature neuroscience* **4**, 887-93 (2001).
56. Cloe, A.L., Orgel, J.P., Sachleben, J.R., Tycko, R. & Meredith, S.C. The Japanese mutant Abeta (DeltaE22-Abeta(1-39)) forms fibrils instantaneously, with low-thioflavin T fluorescence: seeding of wild-type Abeta(1-40) into atypical fibrils by DeltaE22-Abeta(1-39). *Biochemistry* **50**, 2026-39 (2011).
57. Tomiyama, T. *et al.* A new amyloid beta variant favoring oligomerization in Alzheimer's-type dementia. *Annals of neurology* **63**, 377-87 (2008).
58. Di Fede, G. *et al.* A recessive mutation in the APP gene with dominant-negative effect on amyloidogenesis. *Science* **323**, 1473-7 (2009).
59. Jonsson, T. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96-9 (2012).
60. Scheuner, D. *et al.* Secreted amyloid beta-protein similar to that in the senile plaques of Alzheimer's disease is increased in vivo by the presenilin 1 and 2 and APP mutations linked to familial Alzheimer's disease. *Nature medicine* **2**, 864-70 (1996).
61. Duff, K. *et al.* Increased amyloid-beta42(43) in brains of mice expressing mutant presenilin 1. *Nature* **383**, 710-3 (1996).
62. Borchelt, D.R. *et al.* Accelerated amyloid deposition in the brains of transgenic mice coexpressing mutant presenilin 1 and amyloid precursor proteins. *Neuron* **19**, 939-45 (1997).
63. Holcomb, L. *et al.* Accelerated Alzheimer-type phenotype in transgenic mice carrying both mutant amyloid precursor protein and presenilin 1 transgenes. *Nature medicine* **4**, 97-100 (1998).
64. De Strooper, B. *et al.* Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. *Nature* **391**, 387-90 (1998).
65. De Strooper, B. Loss-of-function presenilin mutations in Alzheimer disease. Talking Point on the role of presenilin mutations in Alzheimer disease. *EMBO reports* **8**, 141-6 (2007).
66. Chavez-Gutierrez, L. *et al.* The mechanism of gamma-Secretase dysfunction in familial Alzheimer disease. *EMBO J* **31**, 2261-74 (2012).
67. Saunders, A.M. *et al.* Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**, 1467-72 (1993).
68. Kim, J., Basak, J.M. & Holtzman, D.M. The role of apolipoprotein E in Alzheimer's disease. *Neuron* **63**, 287-303 (2009).
69. Rebeck, G.W., Reiter, J.S., Strickland, D.K. & Hyman, B.T. Apolipoprotein E in sporadic Alzheimer's disease: allelic variation and receptor interactions. *Neuron* **11**, 575-80 (1993).
70. Morris, J.C. *et al.* APOE predicts amyloid-beta but not tau Alzheimer pathology in cognitively normal aging. *Annals of neurology* **67**, 122-31 (2010).
71. Reiman, E.M. *et al.* Fibrillar amyloid-beta burden in cognitively normal people at 3 levels of genetic risk for Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 6820-5 (2009).
72. Bales, K.R. *et al.* Lack of apolipoprotein E dramatically reduces amyloid beta-peptide deposition. *Nature genetics* **17**, 263-4 (1997).
73. Fryer, J.D. *et al.* Human apolipoprotein E4 alters the amyloid-beta 40:42 ratio and promotes the formation of cerebral amyloid angiopathy in an amyloid precursor protein transgenic model. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **25**, 2803-10 (2005).
74. Holtzman, D.M. *et al.* Apolipoprotein E isoform-dependent amyloid deposition and neuritic degeneration in a mouse model of Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 2892-7 (2000).
75. Fagan, A.M. *et al.* Human and murine ApoE markedly alters A beta metabolism before and after plaque formation in a mouse model of Alzheimer's disease. *Neurobiology of disease* **9**, 305-18 (2002).
76. Sharma, M., Kruger, R. & Gasser, T. LRRK2: Understanding the role of common and rare variants in Parkinson's disease. *Mov Disord* **27**, 475 (2012).

77. Rivas, M.A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics* **43**, 1066-73 (2011).
78. Harold, D. *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet* **41**, 1088-93 (2009).
79. Schellenberg, G.D. & Montine, T.J. The genetics and neuropathology of Alzheimer's disease. *Acta neuropathologica* **124**, 305-23 (2012).
80. Cohen, J.C., Boerwinkle, E., Mosley, T.H., Jr. & Hobbs, H.H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *The New England journal of medicine* **354**, 1264-72 (2006).
81. Ahituv, N. *et al.* Medical sequencing at the extremes of human body mass. *American journal of human genetics* **80**, 779-91 (2007).
82. Ji, W. *et al.* Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature genetics* **40**, 592-9 (2008).
83. McDermott, M., Wakelam, M.J. & Morris, A.J. Phospholipase D. *Biochem Cell Biol* **82**, 225-53 (2004).
84. Munck, A., Bohm, C., Seibel, N.M., Hashemol Hosseini, Z. & Hampe, W. Hu-K4 is a ubiquitously expressed type 2 transmembrane protein associated with the endoplasmic reticulum. *FEBS J* **272**, 1718-26 (2005).
85. 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* **164**, 4991-5 (2000).
86. 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* **194**, 1111-22 (2001).
87. Sessa, G. *et al.* Distribution and signaling of TREM2/DAP12, the receptor system mutated in human polycystic lipomembraneous osteodysplasia with sclerosing leukoencephalopathy dementia. *Eur J Neurosci* **20**, 2617-28 (2004).
88. Neumann, H. & Takahashi, K. Essential role of the microglial triggering receptor expressed on myeloid cells-2 (TREM2) for central nervous tissue immune homeostasis. *J Neuroimmunol* **184**, 92-9 (2007).
89. Takahashi, K., Prinz, M., Stagi, M., Chechneva, O. & Neumann, H. TREM2-transduced myeloid precursors mediate nervous tissue debris clearance and facilitate recovery in an animal model of multiple sclerosis. *PLoS Med* **4**, e124 (2007).
90. Turnbull, I.R. *et al.* Cutting edge: TREM-2 attenuates macrophage activation. *J Immunol* **177**, 3520-4 (2006).
91. Frank, S. *et al.* TREM2 is upregulated in amyloid plaque-associated microglia in aged APP23 transgenic mice. *Glia* **56**, 1438-47 (2008).
92. Paloneva, J. *et al.* DAP12/TREM2 deficiency results in impaired osteoclast differentiation and osteoporotic features. *J Exp Med* **198**, 669-75 (2003).
93. Klunemann, H.H. *et al.* The genetic causes of basal ganglia calcification, dementia, and bone cysts: DAP12 and TREM2. *Neurology* **64**, 1502-7 (2005).
94. Soragna, D. *et al.* An Italian family affected by Nasu-Hakola disease with a novel genetic mutation in the TREM2 gene. *J Neurol Neurosurg Psychiatry* **74**, 825-6 (2003).
95. Numasawa, Y. *et al.* Nasu-Hakola disease with a splicing mutation of TREM2 in a Japanese family. *Eur J Neurol* **18**, 1179-83 (2011).
96. Cuyvers, E. *et al.* Investigating the role of rare heterozygous TREM2 variants in Alzheimer's disease and frontotemporal dementia. *Neurobiol Aging* **35**, 726 e11-9 (2014).
97. Jin, S.C. *et al.* Coding variants in TREM2 increase risk for Alzheimer's disease. *Hum Mol Genet* (2014).
98. Reitz, C., Mayeux, R. & Alzheimer's Disease Genetics, C. TREM2 and neurodegenerative disease. *N Engl J Med* **369**, 1564-5 (2013).
99. Klunk, W.E. *et al.* Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. *Annals of neurology* **55**, 306-19 (2004).
100. Clark, C.M. *et al.* Use of florbetapir-PET for imaging beta-amyloid pathology. *JAMA : the journal of the American Medical Association* **305**, 275-83 (2011).
101. Fagan, A.M. *et al.* Inverse relation between in vivo amyloid imaging load and cerebrospinal fluid Abeta42 in humans. *Annals of neurology* **59**, 512-9 (2006).
102. Fagan, A.M. *et al.* Cerebrospinal fluid tau and ptau(181) increase with cortical amyloid deposition in cognitively normal individuals: implications for future clinical trials of Alzheimer's disease. *EMBO molecular medicine* **1**, 371-80 (2009).
103. Jagust, W.J. *et al.* Relationships between biomarkers in aging and dementia. *Neurology* **73**, 1193-9 (2009).

104. Sunderland, T. *et al.* Decreased beta-amyloid1-42 and increased tau levels in cerebrospinal fluid of patients with Alzheimer disease. *JAMA* **289**, 2094-103 (2003).
105. Sperling, R.A., Jack, C.R., Jr. & Aisen, P.S. Testing the right target and right drug at the right stage. *Science translational medicine* **3**, 111cm33 (2011).
106. Hernandez, D.G. *et al.* Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiol Dis* **47**, 20-8 (2012).
107. Webster, J.A. *et al.* Genetic control of human brain transcript expression in Alzheimer disease. *Am J Hum Genet* **84**, 445-58 (2009).
108. Gibbs, J.R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* **6**, e1000952 (2010).
109. Holton, P. *et al.* Initial assessment of the pathogenic mechanisms of the recently identified Alzheimer risk Loci. *Ann Hum Genet* **77**, 85-105 (2013).
110. Cohen, J.C. *et al.* Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 1810-5 (2006).
111. Gatz, M. *et al.* Role of genes and environments for explaining Alzheimer disease. *Archives of general psychiatry* **63**, 168-74 (2006).
112. Emison, E.S. *et al.* Differential contributions of rare and common, coding and noncoding Ret mutations to multifactorial Hirschsprung disease liability. *American journal of human genetics* **87**, 60-74 (2010).
113. Zaghloul, N.A. *et al.* Functional analyses of variants reveal a significant role for dominant negative and common alleles in oligogenic Bardet-Biedl syndrome. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 10602-7 (2010).
114. Haller, G. *et al.* Rare missense variants in CHRNB4 are associated with reduced risk of nicotine dependence. *Hum Mol Genet* **21**, 647-55 (2012).
115. Haller, G. *et al.* Rare missense variants in CHRNB3 and CHRNA3 are associated with risk of alcohol and cocaine dependence. *Hum Mol Genet* **23**, 810-9 (2014).
116. Benitez, B.A. *et al.* The PSEN1, p.E318G variant increases the risk of Alzheimer's disease in APOE-epsilon4 carriers. *PLoS Genet* **9**, e1003685 (2013).
117. Morris, J.C. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* **43**, 2412-4 (1993).
118. McKhann, G. *et al.* Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939-44 (1984).
119. Vallania, F., Ramos, E., Cresci, S., Mitra, R.D. & Druley, T.E. Detection of rare genomic variants from pooled sequencing using SPLINTER. *J Vis Exp* (2012).
120. Vallania, F.L. *et al.* High-throughput discovery of rare insertions and deletions in large cohorts. *Genome research* **20**, 1711-8 (2010).
121. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-9 (2006).
122. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-75 (2007).
123. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics* **89**, 82-93 (2011).
124. Chouery, E. *et al.* Mutations in TREM2 lead to pure early-onset dementia without bone cysts. *Hum Mutat* **29**, E194-204 (2008).
125. Benitez, B.A. *et al.* TREM2 is associated with the risk of Alzheimer's disease in Spanish population. *Neurobiol Aging* **34**, 1711 e15-7 (2013).
126. Pottier, C. *et al.* TREM2 R47H variant as a risk factor for early-onset Alzheimer's disease. *J Alzheimers Dis* **35**, 45-9 (2013).
127. Lattante, S. *et al.* TREM2 mutations are rare in a French cohort of patients with frontotemporal dementia. *Neurobiol Aging* **34**, 2443 e1-2 (2013).
128. Guerreiro, R.J. *et al.* Using exome sequencing to reveal mutations in TREM2 presenting as a frontotemporal dementia-like syndrome without bone involvement. *JAMA Neurol* **70**, 78-84 (2013).
129. Rayaprolu, S. *et al.* TREM2 in neurodegeneration: evidence for association of the p.R47H variant with frontotemporal dementia and Parkinson's disease. *Mol Neurodegener* **8**, 19 (2013).
130. Guerreiro, R. & Hardy, J. TREM2 and neurodegenerative disease. *N Engl J Med* **369**, 1569-70 (2013).

131. Bird, T.D. TREM2 and neurodegenerative disease. *N Engl J Med* **369**, 1568 (2013).
132. Rajagopalan, P., Hibar, D.P. & Thompson, P.M. TREM2 and neurodegenerative disease. *N Engl J Med* **369**, 1565-7 (2013).
133. Bertram, L., Parrado, A.R. & Tanzi, R.E. TREM2 and neurodegenerative disease. *N Engl J Med* **369**, 1565 (2013).
134. Jonsson, T. & Stefansson, K. TREM2 and neurodegenerative disease. *N Engl J Med* **369**, 1568-9 (2013).
135. Benitez, B.A., Cruchaga, C. & United States-Spain Parkinson's Disease Research, G. TREM2 and neurodegenerative disease. *N Engl J Med* **369**, 1567-8 (2013).
136. Cady, J. *et al.* TREM2 Variant p.R47H as a Risk Factor for Sporadic Amyotrophic Lateral Sclerosis. *JAMA Neurol* **71**, 449-53 (2014).
137. Ulrich, J.D. *et al.* Altered microglial response to Abeta plaques in APPPS1-21 mice heterozygous for TREM2. *Mol Neurodegener* **9**, 20 (2014).
138. Schmid, C.D. *et al.* Heterogeneous expression of the triggering receptor expressed on myeloid cells-2 on adult murine microglia. *J Neurochem* **83**, 1309-20 (2002).
139. Gattis, J.L. *et al.* The structure of the extracellular domain of triggering receptor expressed on myeloid cells like transcript-1 and evidence for a naturally occurring soluble fragment. *J Biol Chem* **281**, 13396-403 (2006).
140. Piccio, L. *et al.* Identification of soluble TREM-2 in the cerebrospinal fluid and its association with multiple sclerosis and CNS inflammation. *Brain* **131**, 3081-91 (2008).
141. Gomez-Pina, V. *et al.* Metalloproteinases shed TREM-1 ectodomain from lipopolysaccharide-stimulated human monocytes. *J Immunol* **179**, 4065-73 (2007).
142. Rosenthal, S.L., Barmada, M.M., Wang, X., Demirci, F.Y. & Kamboh, M.I. Connecting the Dots: Potential of Data Integration to Identify Regulatory SNPs in Late-Onset Alzheimer's Disease GWAS Findings. *PLoS One* **9**, e95152 (2014).
143. Karch, C.M. *et al.* Expression of novel Alzheimer's disease risk genes in control and Alzheimer's disease brains. *PLoS One* **7**, e50976 (2012).
144. Allen, M. *et al.* Novel late-onset Alzheimer disease loci variants associate with brain gene expression. *Neurology* **79**, 221-8 (2012).
145. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387-9 (2009).
146. Buerger, K. *et al.* CSF phosphorylated tau protein correlates with neocortical neurofibrillary pathology in Alzheimer's disease. *Brain* **129**, 3035-41 (2006).
147. De Meyer, G. *et al.* Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Arch Neurol* **67**, 949-56 (2010).
148. Ford, J.W. & McVicar, D.W. TREM and TREM-like receptors in inflammation and disease. *Curr Opin Immunol* **21**, 38-46 (2009).
149. Koch, W. *et al.* TaqMan systems for genotyping of disease-related polymorphisms present in the gene encoding apolipoprotein E. *Clinical chemistry and laboratory medicine : CCLM / FESCC* **40**, 1123-31 (2002).
150. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
151. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
152. Benitez, B.A. *et al.* Exome-sequencing confirms DNAJC5 mutations as cause of adult neuronal ceroid-lipofuscinosis. *PLoS One* **6**, e26741 (2011).
153. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
154. Dreses-Werringloer, U. *et al.* A polymorphism in CALHM1 influences Ca2+ homeostasis, Abeta levels, and Alzheimer's disease risk. *Cell* **133**, 1149-61 (2008).
155. Group, C.S. Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology* **22**, 316-25 (2003).
156. Heath, S.C. *et al.* Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet* **16**, 1413-29 (2008).
157. Gao, X., Becker, L.C., Becker, D.M., Starmer, J.D. & Province, M.A. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic epidemiology* **34**, 100-5 (2010).
158. Woolf, B. On estimating the relation between blood group and disease. *Ann Hum Genet* **19**, 251-3 (1955).
159. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* **43**, 519-25 (2011).

160. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-7 (2012).
161. Han, M.R., Schellenberg, G.D., Wang, L.S. & Alzheimer's Disease Neuroimaging, I. Genome-wide association reveals genetic effects on human Abeta42 and tau protein levels in cerebrospinal fluids: a case control study. *BMC Neurol* **10**, 90 (2010).
162. Kim, S. *et al.* Genome-wide association study of CSF biomarkers Abeta1-42, t-tau, and p-tau181p in the ADNI cohort. *Neurology* **76**, 69-79 (2011).
163. Shulman, J.M. *et al.* Genetic Susceptibility for Alzheimer Disease Neuritic Plaque Pathology. *JAMA Neurol*, 1-7 (2013).
164. Elias-Sonnenschein, L.S. *et al.* Genetic loci associated with Alzheimer's disease and cerebrospinal fluid biomarkers in a Finnish case-control cohort. *PLoS One* **8**, e59676 (2013).
165. Chapuis, J. *et al.* Increased expression of BIN1 mediates Alzheimer genetic risk by modulating tau pathology. *Mol Psychiatry* **18**, 1225-34 (2013).
166. Shulman, J.M. *et al.* Functional screening in Drosophila identifies Alzheimer's disease susceptibility genes and implicates Tau-mediated mechanisms. *Hum Mol Genet* **23**, 870-7 (2014).
167. Li, Q. *et al.* Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. *Hum Mol Genet* (2014).
168. McCarthy, M.I. & Hirschhorn, J.N. Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet* **17**, R156-65 (2008).
169. Liang, W.S. *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* **105**, 4441-6 (2008).
170. Wood, A.R. *et al.* Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. *Hum Mol Genet* **20**, 4082-92 (2011).
171. Liang, W.S. *et al.* Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiol Genomics* **28**, 311-22 (2007).
172. Chow, M.L. *et al.* Age-dependent brain gene expression and copy number anomalies in autism suggest distinct pathological processes at young versus mature ages. *PLoS Genet* **8**, e1002592 (2012).
173. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-7 (2010).
174. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-265 (2005).
175. Wunderlich, P. *et al.* Sequential proteolytic processing of the triggering receptor expressed on myeloid cells-2 (TREM2) by ectodomain shedding and gamma-secretase dependent intramembranous cleavage. *J Biol Chem* (2013).
176. Zou, F. *et al.* Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLoS Genet* **8**, e1002707 (2012).
177. Fisher, R.A. *Statistical methods for research workers*, xiii p., 1 l., 307, 1 p. (Oliver and Boyd, Edinburgh etc., 1932).
178. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol* **82**, 239-59 (1991).
179. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-8 (2008).
180. Larner, A.J. The cerebellum in Alzheimer's disease. *Dement Geriatr Cogn Disord* **8**, 203-9 (1997).
181. Li, Q. *et al.* Identification of C1qTNF-related protein 4 as a potential cytokine that stimulates the STAT3 and NF-kappaB pathways and promotes cell survival in human cancer cells. *Cancer Lett* **308**, 203-14 (2011).
182. Zhao, Y. *et al.* Regulation of TREM2 expression by an NF-small ka, CyrillicB-sensitive miRNA-34a. *Neuroreport* **24**, 318-23 (2013).
183. Wray, N.R., Goddard, M.E. & Visscher, P.M. Prediction of individual genetic risk of complex disease. *Curr Opin Genet Dev* **18**, 257-63 (2008).
184. Bertram, L. *et al.* Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. *American journal of human genetics* **83**, 623-32 (2008).
185. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
186. Eichler, E.E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446-50 (2010).
187. Lord, J. *et al.* Next generation sequencing of CLU, PICALM and CR1: pitfalls and potential solutions. *Int J Mol Epidemiol Genet* **3**, 262-75 (2012).

188. Holmes, D. Mind the IGAP. *Lancet Neurol* **10**, 502-3 (2011).
189. Alexopoulos, P. *et al.* Impact of SORL1 single nucleotide polymorphisms on Alzheimer's disease cerebrospinal fluid markers. *Dement Geriatr Cogn Disord* **32**, 164-70 (2011).
190. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**, 6131-8 (2014).
191. Kleinberger, G. *et al.* TREM2 mutations implicated in neurodegeneration impair cell surface transport and phagocytosis. *Sci Transl Med* **6**, 243ra86 (2014).
192. Miyashita, A. *et al.* Lack of Genetic Association Between TREM2 and Late-Onset Alzheimer's Disease in a Japanese Population. *J Alzheimers Dis* (2014).
193. Jiao, B. *et al.* Investigation of TREM2, PLD3, and UNC5C variants in patients with Alzheimer's disease from mainland China. *Neurobiol Aging* **35**, 2422 e9-2422 e11 (2014).
194. Ma, J. *et al.* Association study of TREM2 polymorphism rs75932628 with late-onset Alzheimer's disease in Chinese Han population. *Neurol Res*, 1743132814Y0000000376 (2014).
195. Lee, S., Abecasis, G.R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet* **95**, 5-23 (2014).
196. Kamboh, M.I. *et al.* Genome-wide association study of Alzheimer's disease. *Translational psychiatry* **2**, e117 (2012).
197. Biffi, A. *et al.* Genetic variation and neuroimaging measures in Alzheimer disease. *Arch Neurol* **67**, 677-85 (2010).

Sheng Chih Jin, Ph.D.

Washington University in St. Louis
Division of Biology and Biomedical Sciences
660 S. Euclid Ave.
St. Louis, MO 63110

(517) 402-6714
jin810@wustl.edu
U.S. Citizen

EDUCATION

2010-Aug. 2014	Ph.D. Human and Statistical Genetics Washington University in St. Louis Thesis title: Identification of Functional Variants in Alzheimer's Associated Genes
2008	ScM, Biostatistics, Johns Hopkins Bloomberg School of Public Health Thesis title: GWAS of Smoking Behavior: Combining Information from Multiple Phenotypes
2004	B.S., Applied Mathematics, National Chiao Tung University, Taiwan

PROFESSIONAL AND RESEARCH EXPERIENCE

July '08-Aug. '10 — **Biostatistician**
Department of General Internal Medicine (Prof. Nae-Yuh Wang)
Johns Hopkins School of Medicine

- Designed and analyzed clinical research studies ranging from studies in basic science to clinical trials
- Implemented statistical analyses using R, SAS, and Stata, involving mixed-effects models, longitudinal data analysis, and statistical genetics

Jan-Sept '08 — **Research Assistant**
Department of Oncology (Prof. Leslie Cope)
Johns Hopkins School of Medicine

- Analyzed data from several microarray experiments using Bioconductor packages in R to implement statistical analyses, including *t*-tests and Cox proportional hazards regression

Aug '07-June '08 — **Research Assistant**
Biostatistics Branch, Division of Cancer Epidemiology and Genetics,
National Cancer Institute, National Institutes of Health (Dr. N. Chatterjee)

- Performed statistical data analyses for ongoing Cancer Genetic Markers of Susceptibility project pertaining to cigarette smoking
- Conducted genome-wide association studies using novel Omnibus approach
- Performed parametric bootstrapping for selected candidate genes

June-Aug. '07 — **Summer Research Intern**
Biostatistics Branch, Division of Cancer Epidemiology and Genetics,
National Cancer Institute, National Institutes of Health

- Performed statistical data analyses for ongoing Cancer Genetic Markers of Susceptibility project pertaining to prostate cancer
- Conducted genome-wide association studies using polytomous logistic regression based on score tests
- Implemented incidence density sampling in epidemiology studies

Sep '04-Jan '06 — **Mandatory Military Service, Corporal,** R.O.C Army, Taiwan

TEACHING EXPERIENCE

Sept. '11-Jan '12 **Teaching Assistant**, Washington University in St. Louis
- Human Linkage and Association: M21 GEMS 5483

Sept. '07-May '08 **Teaching Assistant**, Johns Hopkins University
- Statistical Reasoning in Public Health I and II: 140.611-612
- Statistical Methods in Public Health III: 140.623
- Introduction to SAS Statistical Package: 140.632

PUBLICATIONS

*** Equal contribution**

- **Jin SC**, Benitez BA, Karch CM, et al. *Coding Variants in TREM2 Increase Risk for Alzheimer's Disease.* Human Molecular Genetics. 2014 Jun; ppi: ddu277.
- Cruchaga C, Karch CM*****, **Jin SC***, et al. *Rare Coding Variants in Phospholipase D3 Gene Confer Risk for Alzheimer's Disease.* Nature. 2014 Jan. 23; 505 (7484):550-4. doi: 10.1038/nature 12825
- Benitez BA*****, **Jin SC***, et al. *Missense Variant in TREML2 Protects Against Alzheimer's Disease.* Neurobiology of Aging. 2013 Dec 21; doi: 10.1016/j.neurobiolaging.2013.12.010.
- Benitez BA, Karch CM, Cai Y, **Jin SC**, et al. *The PSEN1, p.E318G Variant Increase the Risk of Alzheimer's Disease in APOE-ε4 carriers.* PLoS Genetics. 2013 Aug; 9(8): e1003685.
- Cruchaga C, Kauwe JS, Harari O, **Jin SC**, et al. *GWAS of Cerebrospinal Fluid Tau Levels Identifies Risk Variants for Alzheimer's Disease.* Neuron. 2013 Apr 24; 78(2): 256-58.
- Benitez C BA, Cooper B, Pastor P, **Jin SC**, et al. *TREM2 is Associated with the Risk of Alzheimer's Disease in Spanish Population.* Neurobiology of Aging. 2013 Jun; 34(6): 1711.e15-7.
- Patel PJ, Beaty TH, Ruczinski I, Murray JC, Marazita ML, Munger RG, Hetmanski JB, Wu T, Murray T, Rose M, Redett RJ, **Jin SC**, et al. *X-linked Markers in the Duchenne Muscular Dystrophy Gene Associated with Oral Clefts.* European Journal of Oral Sciences. 2013 Jan; 50(1): 96-103.
- Wang H, Zhang T, Wu T, Hetmanski JB, Ruczinski I, Schwender H, Liang KY, Murray T, Fallin MD, Redett RJ, Raymond GV, **Jin SC**, et al. *The FGF and FGFR Gene Family and Risk of Cleft Lip with/without Cleft Palate.* The Cleft Palate-Craniofacial Journal. 2013 Apr; 121(2): 63-8.
- **Jin SC**, Pastor P, et al. *Pooled-DNA Sequencing Identifies Novel Causative Variants in PSEN1,GRN and MAPT, in A Clinical Early-Onset and Familial Alzheimer's Disease Ibero-American Cohort.* Alzheimer's Research & Therapy. 2012 Aug 20; 4(4): 34.
- Beaty TH, Ruczinski I, Murray JC, Marazita ML, Munger RG, Hetmanski JB, Murray T, Redett RA, Fallin MD, Liang KY, Wu T, Murray T, Patel PJ, **Jin SC**, et al. *Evidence for Gene-Environment Interaction in A Genome Wide Study of Isolated, Non-Syndromic Cleft Palate*. Genetic Epidemiology. 2011 Sep; 35(6): 469-78.
- Beaty TH, Murray JC, Marazita ML**,** Munger RG, Ruczinski I, Hetmanski JB, Liang KY, Wu T, Murray T, Fallin MD, Redett RA, Raymond G, Schwender H, **Jin SC**, et al. *A Genome-Wide Association Study of Cleft Lip with/without Cleft Palate Identifies Risk Variants Near MAFB and ABCA4.* Nature Genetics. 2010 May 2. PMID: 20436469
- Caporaso N, Gu F, Chatterjee N**, Jin SC**, et al. *Genome-Wide and Candidate Gene Association Study of Cigarette Smoking Behavior.* PLoS ONE. 2009; 4(2):e4653.

BOOK CHAPTERS

- **Jin SC**, Benitez BA, Deming Y, Cruchaga C. *Pooled-DNA Sequencing for Elucidation of Genomic Risk Factors/Rare Variants Underlying Alzheimer's Disease.* The Systems Biology of Alzheimer's Disease: Methods and Protocols: Springer.

HONORS AND AWARDS

2014 ▪ Finalist, Fourth Annual Hope Center Retreat Poster Session

2014	▪ Howard Hughes Medical Institute Postdoctoral Fellowship, Yale University
2012	▪ Alzheimer's Disease International Conference Travel Fellowship, Alzheimer's Association, 2012
2012	▪ Best Oral Presentation Award, Human Statistical Genetics Program 2012 Retreat
2011-2013	▪ Fellow, Lucille P. Markey Special Emphasis Pathway in Human Pathobiology, Markey Foundation, Washington University School of Medicine
2007-2008	▪ Departmental Scholarship, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
2007	▪ Cancer Research Training Award, National Cancer Institute, NIH
2006	▪ Merica Institute Scholarship
2004	▪ Dean's List for one semester

REFERENCES

Alison Goate, D.Phil.
Samuel & Mae S. Ludwig Professor of Genetics in Psychiatry
Professor of Genetics, Professor of Neurology
Dept. of Psychiatry, B8134
Washington University School of Medicine,
660 S. Euclid Ave., St. Louis, MO 63110
314-362-8691
goatea@psychiatry.wustl.edu

David M. Holtzman, MD
Andrew B. and Gretchen P. Jones Professor and Chairman
Dept. of Neurology, B8111
Washington University School of Medicine,
660 S. Euclid Ave., St. Louis, MO 63110
314-362-7316
holtzman@neuro.wustl.edu

Carlos Cruchaga, Ph.D.
Assistant Professor
Dept. of Psychiatry, B8134
Washington University School of Medicine,
660 S. Euclid Ave., St. Louis, MO 63110
314-286-0546
ccruchaga@wustl.edu