Aggregating common, rare, and private variants in Alzheimer's disease genes

Prerna Das

Master's Capstone Project

Presented to

Division of Bioinformatics & Computational Biology
Department of Medical Informatics & Clinical Epidemiology
Oregon Health & Science University School of Medicine

in partial fulfillment of the requirements of

Master of Biomedical Informatics

June 2017

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's Capstone Project of

PRERNA DAS

"Aggregating common, rare, and private variants in Alzheimer's disease genes"

Has been approved

Beth Wilmot, Ph.D.
Capstone Advisor
Department of Medical Informatics and Clinical Epidemiology

Acknowledgements

I thank God for His countless blessings and for giving me the opportunity to pursue and finish my Master's degree. I would like to thank my amazing mentor, Prof Beth Wilmot, for her continuous support and encouragement. It has been a thoroughly enriching experience learning and doing Statistical Genetics with Prof Wilmot. I would also like to thank all my instructors at the Oregon Health and Science University and everybody at the Department of Medical Informatics and Clinical Epidemiology. I am also very grateful for the help I received from Ms. Diane Doctor. Finally, I would like to extend my appreciation towards my family - my husband, son and my parents – for their love and for always believing in me.

Table of Contents

List of Tables

List of Figures

Abstract

Alzheimer's disease (AD) is an irreversible neurodegenerative disorder. It is the sixth leading cause of death in the United States. Genome-wide association studies have uncovered nearly 40 common genetic variants (minor allele frequency (MAF) > 5%) which are associated with increased susceptibility to AD. However, the common variants found so far do not completely account for the genetic component of the disease. With the technological advancement of deep sequencing, the focus has shifted to exploring the role of rare (MAF < 0.5%) and private variants. In addition, it has been hypothesized that rare variants are generally functional and highly penetrant with large effect sizes. We investigated whether the aggregation of rare and private variants within a gene region based on linkage disequilibrium (LD) complemented the association between common variants in the gene region and the disease. The whole genome sequencing and phenotype data used in our study come from the Alzheimer's Disease Neuroimaging Initiative (ADNI). We looked at four AD associated genes with different LD structure - *APOE*, *ABCA7*, *CD2AP*, and *CR1*, and found a total of 3016 variants (common, low-frequency, rare and private) across the four genes. All four genes had varying percentages of rare (*APOE*-75.39%, *ABCA7*-47.77%, *CD2AP*-77.34%, *CR1*-70.34%) and private (*APOE*-54.45%, *ABCA7*-43.33%, *CD2AP*-42.74%, *CR1*-52.24%) variants. In order to aggregate the effects all types of variants present in these genes, we used the Sequence kernel association test (SKAT-O) to test the association between the overall burden of variants and the AD phenotype. We found that aggregating all the variants indeed complements the individual common variant – disease association and is an effective strategy to identify the genomic regions harboring potential rare causal variants.

Chapter 1: Introduction

1.1 Alzheimer's Disease and Rare Variants

Alzheimer's disease (AD) is a complex neurodegenerative disorder afflicting approximately 44 million people worldwide (Querfurth & LaFerla, 2010). It is the most predominant form of dementia and is characterized by progressive cognitive decline. It is heritable with a strong genetic risk factor component. Several genome-wide association (GWA) studies have been undertaken to uncover the common variants (Minor Allele Frequency (MAF) > 0.5%) contributing to risk of developing AD. Although these studies have identified more than 40 genetic risk factors for AD, a large percentage of the heritability remains unexplained. This "missing heritability" can be explained by the fact that traditionally GWAS studies have focused mostly on the role of the common variants. It is possible that the low-frequency (0.5%<MAF ≤ 5%), rare (0.1%< MAF≤ 0.5%), very rare (MAF≤ 0.1%), and private variants are the additional variants/risk factors associated with complex diseases (Long et al., 2017). Indeed, recent GWA studies with large number of case-control samples have found significant association of low-frequency and rare variants with other complex traits such as type 2 diabetes and cancers. Rare coding variants have also been found in several Alzheimer's disease genes, such as *PLD3, APP, ADAM10, AKAP9, APOE, SORL1, UNC5C,* and *TREM2* (Lord, Lu, & Cruchaga, 2014). The importance of rare variants can be realized from the fact that many Mendelian disorders are caused by highly penetrant rare variants. It has also been observed that the a large majority of the rare variants – single nucleotide and indels, are functional, resulting in loss of gene function. Hence, rare variants assume importance in

understanding the etiology of a complex disease such as Alzheimer's as well as developing new targets for disease diagnostics and treatment.

1.2 Methods to study the rare variants

Since rare variants are kept at low frequency by purifying selection, very large populations are needed to identify such variants with large effects on clinical traits. Many statistical methods have been developed for testing the association between the sets of rare variants and binary or continuous traits. These methods combine the variants based on a gene or a region, to test for the association with the trait. There are two main types of such collapsing/aggregating tests – burden tests and variance component tests, and a third type, which combines both the burden and the variance component tests. Other strategies to study the rare variant association include, using family samples or isolated populations so as to increase the frequency of the rare variant and use of samples which are phenotypic extremes (Nicolae, 2016).

1.3 Specific Aims

1. Identify all the variants – common, low-frequency, rare, and private variants in AD genes – *CR1, CD2AP, ABCA7,* and *APOE.*
2. Determine association between the known significantly associated variant for each gene and AD status.
3. Aggregate all the variants for a gene/gene-region and determine the association of the region with the AD status.

Chapter 2: Background

3.1 Alzheimer's Disease

Alzheimer's Disease (AD) is the most common cause of dementia in late adult life. Clinical characteristics of AD include, loss of memory, inability to learn new things, inability to do calculations, mental confusion, indifference, depression and delusions. It is estimated that, 5 million or 1 in 9 people over the age of 65 are afflicted by AD, in the United States alone, making it a critical public health issue. $200 billion are spent annually on caring for individuals suffering from dementia.

There are three recognized stages of AD: preclinical, mild cognitive impairment (MCI), and Alzheimer dementia. An AD brain is marked by progressive loss of neurons and synapses in the cerebral cortex and certain subcortical regions on the brain. The neuropathological hallmarks of AD are amyloid plaques and formation of intraneuronal neurofibrillary tangles consisting of hyperphosphorylated tau protein. The amyloid plaques are extracellular precipitations of the β-amyloid (Aβ) peptide derived from the proteolytic cleavage of the amyloid precursor protein (APP). There is progressive deterioration of memory and cognitive functions, causing the patient to lose autonomy, and require full time medical care. No treatment exists for AD. The existing drugs only temporarily relieve the AD symptoms.

2.2 Genetic risk factors for Alzheimer's Disease

AD is a highly heritable (with heritability, that is, the proportion of variation in disease risk attributable to inherited genetic variation, up to 60%) and genetically complex disease . Age is the principal risk factor for AD.

There are two types of AD: early-onset AD (EOAD) and late-onset AD (LOAD). EOAD cases account for 1-2% cases, with symptoms appearing before 65 years of age. Autosomal dominant mutations in three genes – amyloid precursor protein gene (*APP*), presenilin 1 gene (*PSEN1*) and presenilin 2 gene (*PSEN2*), are the genetic risk factors for EOAD.

LOAD is the more common and complex form of the disease and occurs late in life (>65 years). It has a strong generic predisposition with heritability estimate of 60-80%. The genetic component though is complex and heterogeneous, with several gene, different gene mutations possibly interacting with each other, and with the environmental factors. For many years, the *APOE-ε4*, was the only major known genetic risk factor for AD. Other variants that increase the susceptibility to AD have found within/near *CR1, BIN1, CD2AP, EPHA1, CLU, MS4A64, PICALM, ABCA7, CD33, PTK2B, SORL1, SLC24A4-RIN3, DSG2, INPP5D, MEF2C, NME8, ZCWPW1, FERMT2, CASS4, TREM2*, and *UNC5C* (Lambert et al., 2013).

In this work, we focus our attention to variants present in four of the AD genes – *APOE, CR1, CD2AP,* and *ABCA7*. These genes are among the top ten genetic loci strongly associated with AD on the AlzGene database (http://www.alzgene.org).

2.3 Apolipoprotein E (*ApoE)*

The apolipoprotein E (APOE) gene is located on chromosome 19q13.2 and is 3639 base pairs (bp) long. The Apo-E protein is the primary cholesterol transporter in the central nervous system. It is primarily synthesized by astrocytes and microglia in the brain and transports cholesterol to neurons via the Apo-E receptors. There are three polymorphic alleles for the human *APOE* gene – ε2, ε3, and ε4 – with a

worldwide frequency of 8%, 77%, and 15% respectively in the general population . The ε4 allele frequency is approximately 40% in AD patients. These three polymorphic forms are defined by two SNPs, rs429358 and rs7412, in the following manner (Lyall et al., 2014):

rs429358	rs7412	Name
C	T	ε1
T	T	ε2
T	C	ε3
C	C	ε4

GWA studies have determined *APOE*-ε4 to be the strongest risk factor for AD, associated with increased risk for both EOAD and LOAD (Chartier-Harlin et al., 1994). The *APOE*-ε4 allele shows correlation with increased cerebrovascular deposition of Aβ.

Genetic susceptibility variants have also been found in *TOMM40* (translocase of outer mitochondrial membrane 40), which is located adjacent and approximately 2kb upstream to *APOE*. *TOMM40* is in linkage disequilibrium (LD) with *APOE* (Roses et al., 2016).

SNPs and indels conferring susceptibility to AD have also been uncovered in *APOC1* (apolipoprotein C-1). *APOC1* lies approximately 5kb downstream to *APOE* and is also in LD with *APOE* (Zhou et al., 2014). Since, *TOMM40* and *APOC1* are in LD with *APOE*, investigators look the *TOMM40-APOE-APOC1* region as a whole for uncovering susceptibility variants to LOAD.

2.4 Complement receptor 1(*CR1)*

CR1 (also known as *CD35*) is located on chromosome 1 at locus 1q32 and is 144,500 bp long. It is found on the surface of red blood cells in humans and aids in transportation of cellular debris to the liver for degradation. The *CR1* variant - rs6656401, was first implicated as a LOAD susceptibility variant by the GWAS published by Lambert et al. (Lambert et al., 2009). Subsequent GWA studies and meta-analysis have replicated and confirmed the association of *CR1*-rs6656401 with AD phenotype (Zhu et al., 2015). To date, a total of nine *CR1* SNPs have been implicated as being AD risk factors. Six out these, including rs6656401 lie within the intronic region of *CR1*. These do not encode CR1 directly, but potentially regulate the gene expression and therefore influence AD susceptibility. It has been hypothesized that people with AD-risk *CR1* polymorphisms have low levels of CR1 which results in less efficient clearance of Aβ, gradually leading to its aggregation and deposition in the brain.

2.5 CD2-associated protein (*CD2AP)*

CD2AP encodes a scaffolding protein that regulates actin cytoskeleton and is also involved in receptor-mediated endocytosis. It is located on chromosome 6 at locus 6p12 and is 149,474 bp long. LOAD-susceptibility variants in *CD2AP* were first uncovered in 2011 in two meta-analsysis GWA studies using European ancestry subjects (Hollingworth et al., 2011; Naj et al., 2011). Subsequently, Lambert et al found significant association of rs10948363 with AD phenotype (Lambert et al., 2013) by meta-analysis. The role of CD2AP in LOAD pathogenesis is not yet

delineated. It has been suggested that CDA2P regulates the intracellular production of Aβ (Ubelmann et al., 2017).

2.6 ATP-binding cassette transporter A7 (*ABCA7)*

ABCA7 is a member of the A subfamily of ABC transporters, which mainly transport lipids across membranes. It is located on chromosome 19 at locus 19p13 and is 25,469 bp long. It is abundantly expressed in the brain microglial cells. The *ABCA7* gene locus was first identified as an AD-susceptibility locus by Hollingworth et al. through meta-analysis using four GWA data sets (Hollingworth et al., 2011). Multiple susceptibility variants have been found within the *ABCA7* locus. Lambert et at. using a two-stage meta-analysis of GWAS in European ancestry individuals found the significant association of the intronic SNP, rs4147929, with AD (Lambert et al., 2013). It has been suggested that ABCA7 contributes to AD through several pathways such as Aβ accumulation, lipid metabolism, and phagocytosis (Zhao, Yu, Tan, & Tan, 2015).

2.7 Rare Variants in Alzheimer's Disease

Although GWAS with common variants have been successful in identifying several genes associated with AD, these do not completely explain the heritability in AD. Additionally, often the strongest association has been detected within the intronic or intragenic region, making the causal factor unclear. Hence, the focus has shifted to determining all the variants – common, low-frequency, rare and private variants in a locus. Such studies have been aided by technological advances and reducing costs of whole genome sequencing (WGS) and whole exome sequencing

(WES) as well as the 1000 Genomes Project, which aims at cataloging all human variation across multiple populations.

Rare risk AD variants – missense and exonic, have been found in *TREM2* using WGS and imputation into AD case control cohort and in *SORL1* using WES. Using a combination of family based design and WES, rare AD variants have been found in *PLD3*, *UNC5C*, and *AKAP9*. Protective rare AD variants have been found in *APOE* using mining the publicly available sequencing data to identify the rare variants, followed by genotyping those in a large case-control cohort (Lord et al., 2014).

The rare variants identified show strong significant association with AD with odds ratio in the range from 0.1 to 5.48. Hence, it becomes important to explore the role of rare variants in AD.

2.8 Optimal Unified Sequence Kernel Association Test (SKAT-O)

Rare variant association studies are usually underpowered, as only a small proportion of samples may carry variant alleles at each locus. Large sample size is needed as even with large effect size, rare variants can only be detected in large samples. Power can also be further affected by the presence of both risk and protective variants. Additionally, the significance level of 5×10^{-8}, accepted for common variants, is not applicable for rare variants as the rare variants are more numerous and less correlated with each other than common variants.

One of the strategies therefore, to study the rare variant association, involves collapsing/aggregating all the observed variants within a sub-region, which may be a gene or an LD region. This method is called a Burden test. It collapses or aggregates multiple rare variants in the region under consideration into a single ‘super’ variant.

The association tests are then performed on this single 'super' variant (Li & Leal, 2008). The simplest way of collapsing the variants is to code 1 for individuals that carry one or more rare variants within the tested region and 0 otherwise. Burden tests are suitable when all the variant in a region are causal and affect the phenotype in the same direction.

The Sequence Kernel Association Test (SKAT) addresses the problem of when the variants with opposite effects are present. Instead of aggregating variants, SKAT aggregates the individual variants score test statistics. This test can be less powerful than the burden tests if a large proportion of the rare variants in the region under consideration are causal and have effects in the same direction.

SKAT-Optimal (SKAT-O) is a hybrid method that combines both the burden test and the SKAT test (Lee et al., 2012). It tackles that problem that both risk, protective, and non-causal variants may be present in a region and that there is no prior knowledge about the directionality of a causal variant. The test statistic for SKAT-O is given by:

$$Q_\rho = \rho Q_{Burden} + (1 - \rho)Q_{SKAT},\ 0 \leq \rho \leq 1, \text{ where}$$

Q_{Burden} is the score test statistic for the burden test, Q_{SKAT} is the test statistic of SKAT, and ρ is a correlation term that determines the relative contribution of either test to the SKAT-O statistic. The value of ρ is determined by performing the test with different values of ρ and choosing the one that gives the minimum p-value. When $\rho=0$, the SKAT-O reduces to SKAT test and when $\rho=1$, it is equivalent to a burden test, and when $0 < \rho < 1$, it achieves the unification of SKAT and burden test.

Chapter 3: Materials and Methods

3.1 Data

The whole-genome sequencing (WGS) and the phenotype data comes from the Alzheimer's Neuroimaging Initiative (ADNI). ADNI is an ongoing multi-site, longitudinal study. The ADNI subjects fall into the following categories – normally aging (CN), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), or AD. ADNI database has clinical, image, genetic, and biomarker data for its subjects. The data is available to registered users for scientific investigation, teaching or planning clinical research studies.

3.2 Data Demographics

The WGS data was available for 808 subjects. All the EMCI, LMCI, and AD subjects were pooled together to form the one group – Case and all the CN subjects formed the other group – Control. **Table 3.1**, shows the descriptive statistics for the Case and the Control group.

Table 3.1 - Descriptive Statistics for ADNI data

	Control (n= 280)	Case (n = 528)
Gender (M/F)	136\144	310\218
Age (Mean ± SD)	74.51 ± 5.57	72.56 ± 7.69

3.3 Genotype Data Cleaning

The WGS data was present in the Variant Call Format (VCF). There was a VCF file for each chromosome. The VCF files for Chromosome 1, Chromosome 6, and Chromosome 19 were subsetted to obtain VCF files corresponding to *CR1*

(Chromosome 1), *CD2AP* (Chromosome 6), *ABCA7* (Chromosome 19) and the whole of *TOMM40-APOE-APOC1* gene region (Chromosome 19).

For each of the above four resulting VCF files, variants were evaluated on read depth (DP), genotype quality (GQ), and genotype (GT), and excluded if they had any sample with either DP=NA or DP < 10, GQ=NA or GQ < 20, or GT=NA (Carson et al., 2014). If any of the know significantly associated variants – rs6656401 in *CR1*, rs10948363 in *CD2AP*, rs4147929 in *ABCA7*, and rs429358 and rs7412 in *APOE* region had any samples which did not meet the filtering criteria, instead of eliminating the variant, the poor quality samples were eliminated. Table 2, depicts the data cleaning process.

3.4 Principal Component Analysis (PCA)

PCA was done to adjust for population stratification. Population stratification refers to allele frequency differences between cases and controls due to differences in ancestry (Price et al., 2006). This can lead to spurious associations between the genotype and the phenotype.

ADNI SNPs on Chromsome 1, Chromosome 6, and Chromsome 19, which had also been genotypes for the HapMap samples were used for the PCA analysis. The self-reported race category was compared with that determined from the PCA analysis. The first principal component was used as a covariate in the null association model in the SKAT-O model (Section 3.9).

Table 3.2 – Genotype Data Cleaning Steps

		CR1	***CD2AP***	***ABCA7***	***TOMM40-APOE-APOC1***
1	Number of samples	808	808	808	808
2	Number of variants	1625	2105	529	480
3	Most significantly associated SNP according to Jean-Charles Lambert et al.	rs6656401	rs10948363	rs4142979	rs429358 and rs7412
4	Is the most significantly associated SNP genotyped	Yes	Yes	Yes	Yes
5	Number of variants out of the total that have samples with DP=NA	82	113	66	12
6	Any sample that has DP=NA, for the most significant associated SNP? If yes, note the sample id(s) and do not filter the SNP	No	No	No	No
7	Number of variants left after filtering out those having samples with DP=NA;	1543	1992	463	468
8	Number of variants out of the variants in 7, with samples having DP < 10	222	229	349	200

9	Any sample that has DP <10, for the most significantly associated SNP? If yes, note the sample_id(s) and do not filter the SNP	No	No	7 samples	rs429358 has 4 samples and rs7412 has 5 samples which have DP < 10
10	Number of variants out of the total left after filtering the variants having samples with DP=NA or DP < 10	1321	1763	115	270
11	Number of variants out of the total that have samples with GQ=NA	190	218	92	28
12	Any sample that has GQ=NA, for the most significant associated SNP?If yes, note the sample_id(s) and do not filter the SNP	No	No	1 sample	No
13	Number of variants left after filtering out those having samples with GQ=NA	1435	1887	438	452
14	Number of variants out of the variants in 13, with samples having GQ < 20	146	184	181	110

15	Any sample that has GQ < 20, for the most significantly associated SNP?If yes, note the sample_id(s) and do not filter the SNP	No	No	140 samples	rs4293458 has 11 samples
16	Number of variants out of the total left after filtering the variants having samples with GQ=NA or GQ < 20	1289	1703	258	343
17	Number of variants that have all samples that with both DP > 10 and GQ > 20, that is variants common from 10 and 16	1223	1640	93	243
18	Number of unique samples to be filtered from 9, 12, and 15	0	0	146 samples	19 samples
19	Any sample that has GT=NA, for the most significantly associated SNP? If yes, filter out those samples and keep the SNP.	No	No	4 samples	3 samples

20	Number of variants out of the variants in 17, that have samples with GT=NA	42	56	33	52
21	**Number of variants left after filtering out variants that have samples with GT=NA**	**1181**	**1584**	**60**	**191**
22	**Number of samples left**	**808**	**808**	**658**	**786**

3.5 Gene Models

Gene Models for *CR1*, *CD2AP*, *ABCA7*, and the *TOMM40-APOE-APOC1*, were obtained using the Bioconductor package, biomaRt, which provides an interface to the Ensembl and other databases such as Uniprot and HapMap {Citation}. The Gene Models were plotted using the Gviz Bioconductor package {Citation}.

3.6 Linkage Disequilibrium (LD) Plots

LD plots were constructed for *CR1*, *CD2AP*, *ABCA7*, and the *TOMM40-APOE-APOC1*, using both the 1000 Genomes markers as well as the ADNI SNPs, using the Haploview software from the Broad Institute (Barrett, Fry, Maller, & Daly, 2005).

3.7 Regional Association Plots

Regional Association Plots display the strength and extent of the association signal relative to genomic position, LD, and recombination pattern and the position of genes in the region. These were plotted for *CR1*, *CD2AP*, *ABCA7*, and *APOE* using the LocusZoom web-based software and the data from the Lambert et al (Lambert et

al., 2013; Pruim et al., 2010). The data from Lambert et al., was available to download from http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php.

3.8 SNP-level Association Test

A logistic regression model, with AD status as the outcome and age, sex, and the first principal component as the covariates was used to determine the association of the SNPs in *CR1*, *CD2AP*, *ABCA7*, and the *TOMM40-APOE-APOC1* with the disease status. The logistic regression model can be represented by,

$$\text{logit}\,(p_i) \sim \beta_0 + \beta_1.\text{Age}_i + \beta_2.\text{Genderi} + \beta 3.\text{Race}_i + \beta 4.\text{Genotype}_i,$$

where, p_i = expected value of phenotype for individual i, given the genotype and the covariates.

3.9 Gene-level Association Test/SKAT-O Analysis

SKAT-O was used to determine the region level association with the AD status. The null model was determined using AD status as the binary outcome, and age, sex, and the first principal component as the covariates. The genotype matrix used in the association testing consisted of all the variants (all SNPs + all Indels), all the rare and low-frequency variants and only the rare variants.

Chapter 4: Results

4.1 Genes/Gene-Region to explore

Lambert et al. have identified 21 different gene loci associated with Alzheimer's. Our aim was to explore the role of LD structure, to determine the region over which to aggregate the variants, for rare variant analysis. The underlying thought here being that the variants in a LD structure are inherited together due to linkage disequilibrium and hence can be thought of as one unit. We therefore, looked at the LD structure (Appendix A) and the Regional Association Plots (Appendix B) for all these genes and chose to study *CR1*, *CD2AP*, *ABCA7*, and *TOMM40-APOE-APOC1*, all of which have very different LD structure (Fig 4.2 a & b).

These genes/gene-regions are known to harbor variants that are associated with Alzheimer's disease. We started by exploring the gene models for these genes and the location of the known most significantly associated variant within these genes. **Table 4.1** gives the gene co-ordinates of the four genes, with their known significant SNP(s), genomic position and location of the SNP(s) with respect to genomic feature. All the known significant SNPs are common SNPs, as is seen from their MAF.

Fig. 4.1 depicts the gene models for the four genes and the known significantly associated SNP. The two known SNPs of the *APOE* gene, are located very close to each other. The *TOMM40* gene lies upstream and the *APOC1* gene lies downstream to the *APOE* gene. Linkage Disequilibrium (LD) structure of the genes was constructed using the both the markers ADNI data (**Fig 4.2a**) as well as the markers from 1000 Genomes (**Fig. 4.2b**). The genes show different LD pattern. *CR1* gene shows long stretches of LD, with the whole region made up of 8 haplotype blocks.

Table 4.1 Genomic Coordinates of the genes and the top hits within those genes.

Gene	Chr	Gene Start Position (bp)	Gene End Position (bp)	Gene Size (bp)	Known Significantly Associated SNP	SNP Position*	SNP Location in gene	MAF**
CR1	1	207669492	207813992	144500	rs6656401	207692049	intron	0.263
CD2AP	6	47445525	47594999	149474	rs10948363	47487762	intron	0.278
ABCA7	19	1040102	1065571	25469	rs4147929	1063443	intron	0.182
APOE	19	45409011	45412650	3639	rs429358 rs7412	45411941 45412079	exon exon	0.177 0.066

* Build 37, assembly hg19

** From 1000 Genomes

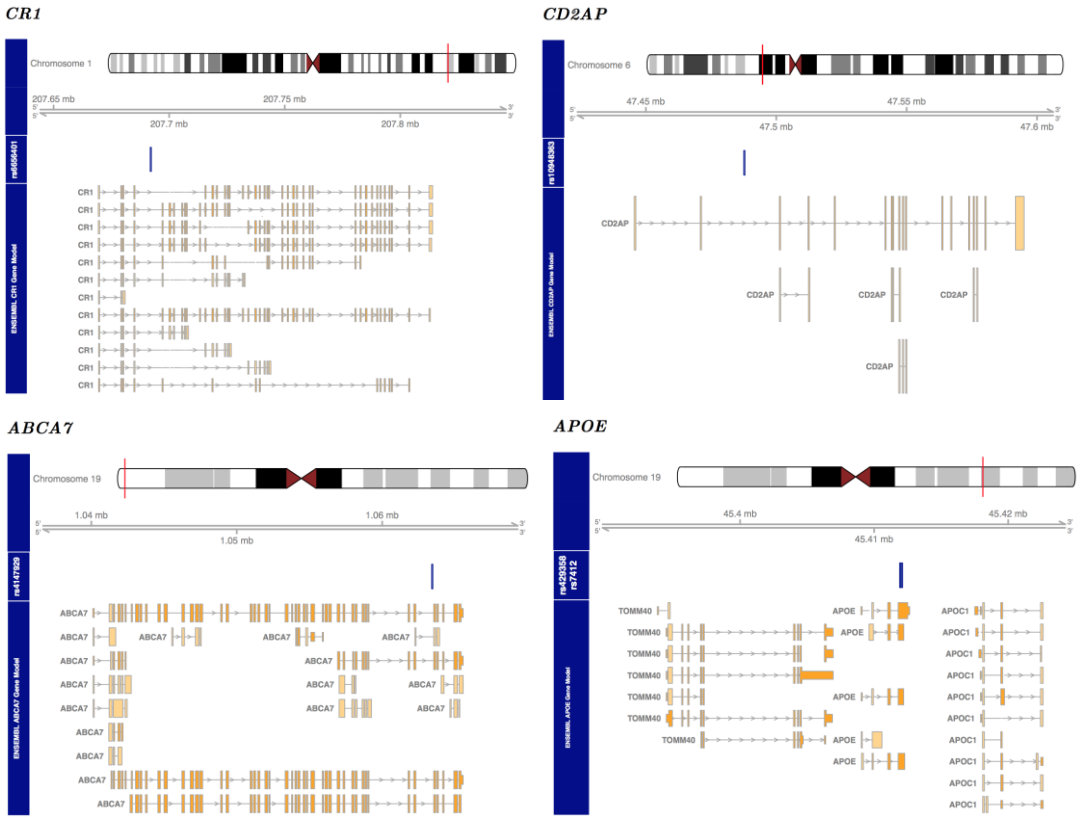


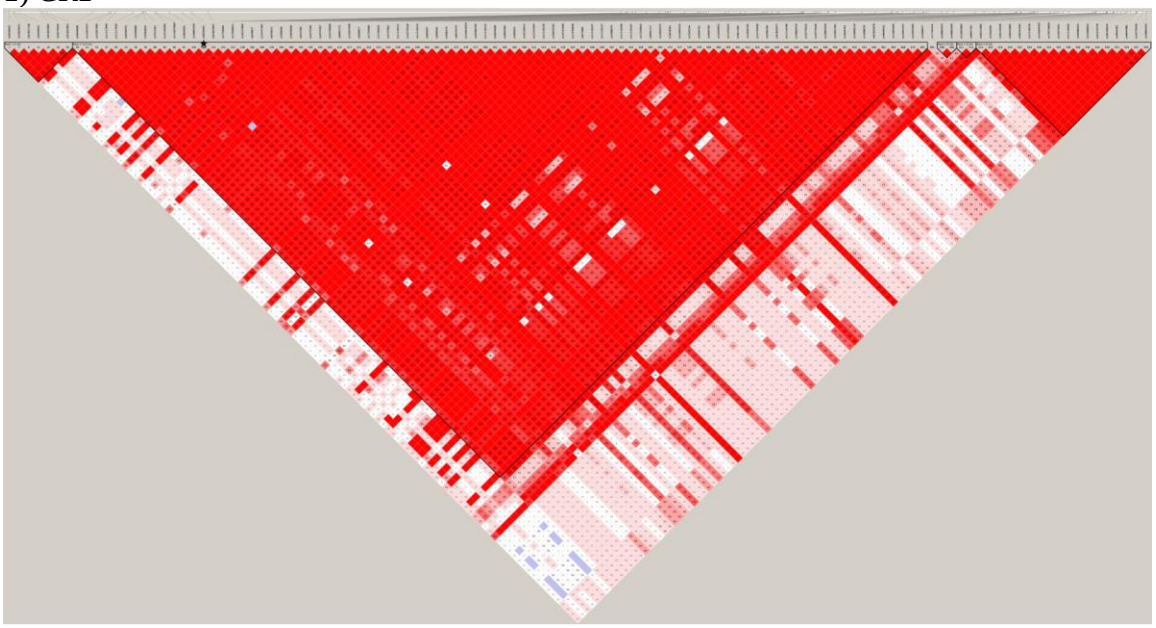
Fig 4.1 - Gene Models for *CR1*, *CD2AP*, *ABCA7*, and the *TOMM40-APOE-APOC1* gene region

The known significantly associated SNP – rs6656401, lies in the biggest haplotype block. The *CD2AP* region is one big LD region, with no recombination spots. The *ABAC7* gene shows several small LD regions interspersed with several probable recombination regions. The *APOE* gene shows LD with its two neighboring genes – *TOMM40* and *APOC1*. Similar LD structure for all the genes was obtained using the 1000 Genomes markers (**Fig. 4.2b**).

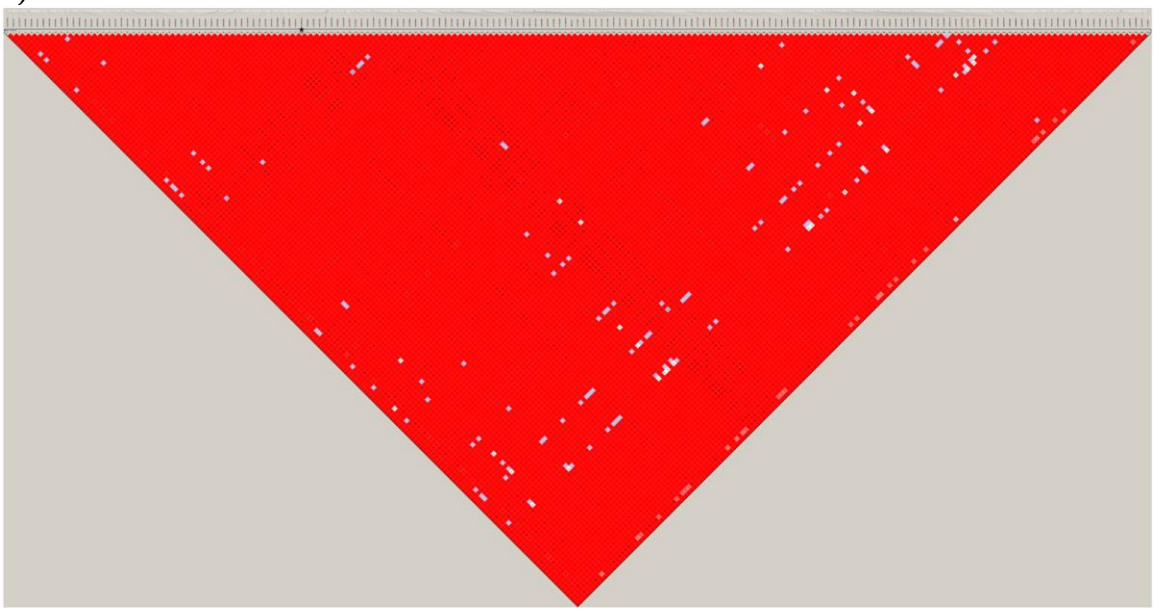
LD ensures that a set of variants are inherited together. The regional association plots for the above genes/gene region (**Fig 4.3**) help in visualizing the varying degrees of association of a single strongly associated SNP, with other variants in a region due to local LD patterns. For each region, there are several other SNPs, that show varying degree of association, with the most significantly associated SNP. Interestingly, rs7412, one of the known significantly associated SNPs for *APOE*, shows as not associated with AD (**Fig 4.3d**) in the Lambert et al data set.

Based on the LD plots and the Regional Association Plots, we decided to aggregate all the variants for the *CD2AP*, *ABCA7*, and the whole of *TOMM40-APOE-APOC1* gene region. For *CR1* and *ABCA7*, we also decided to aggregate the variants in only in the haplotype block that contains the known significant SNP (**Fig 4.2 a and c**).

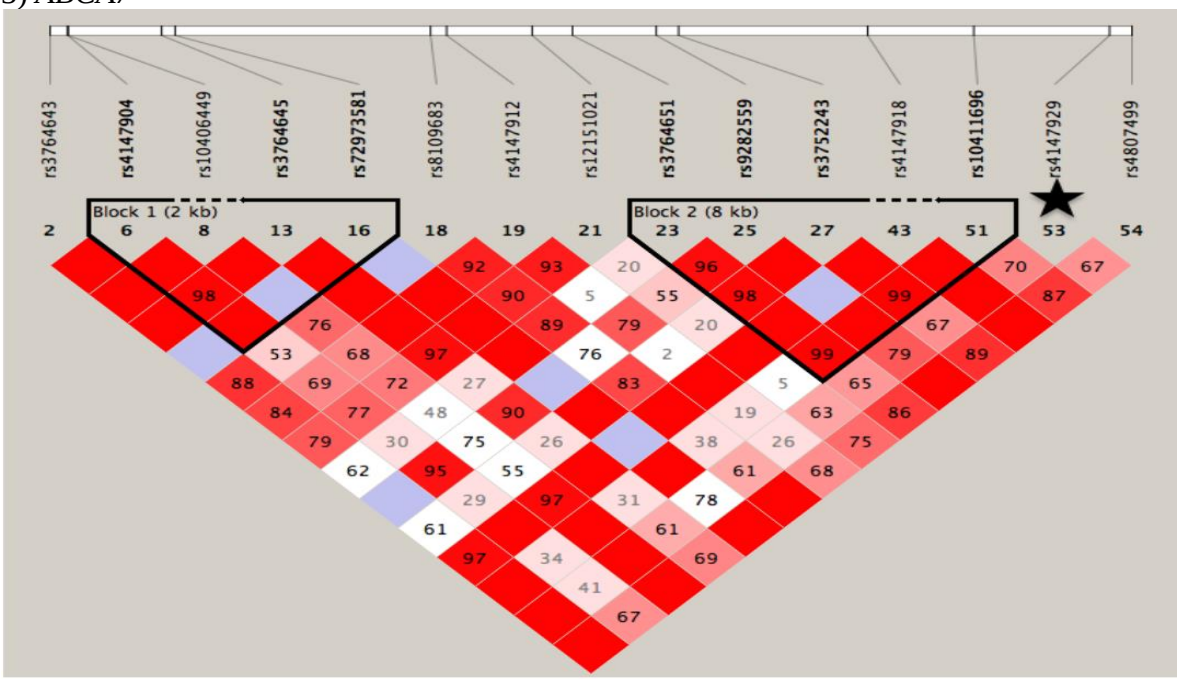
1) *CR1*



2) *CD2AP*



3) *ABCA7*



4) *TOMM40-APOE-APOC1*

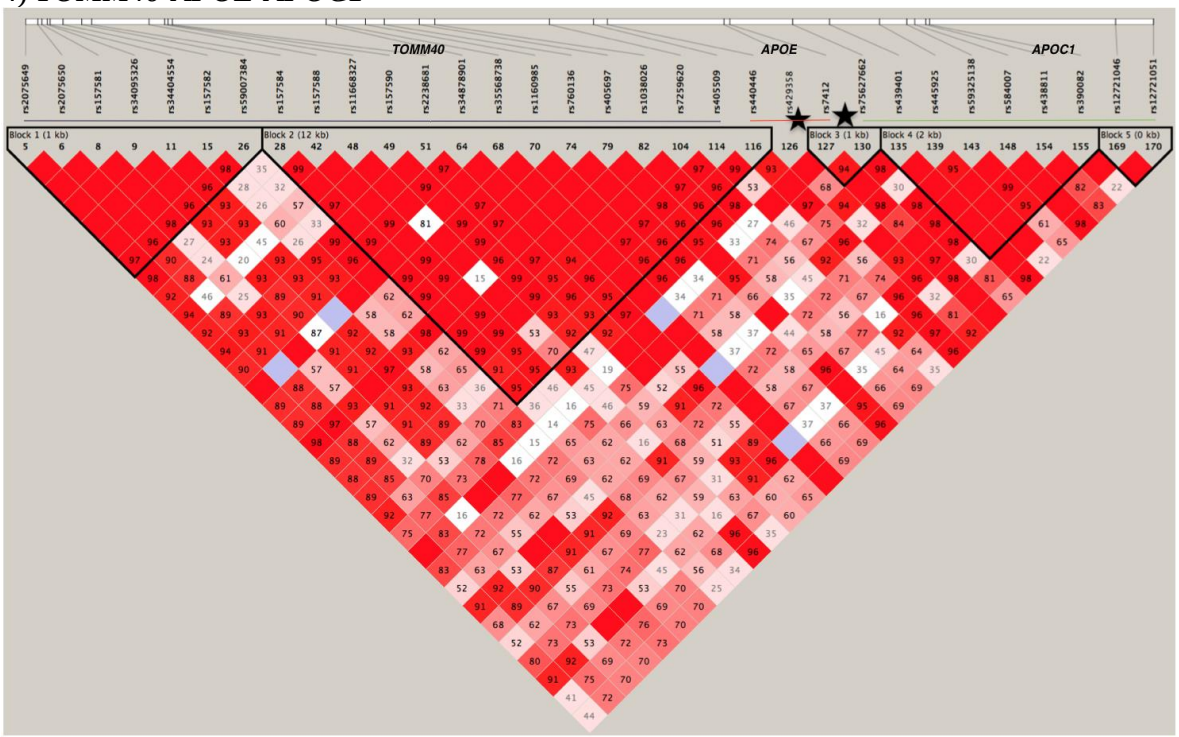
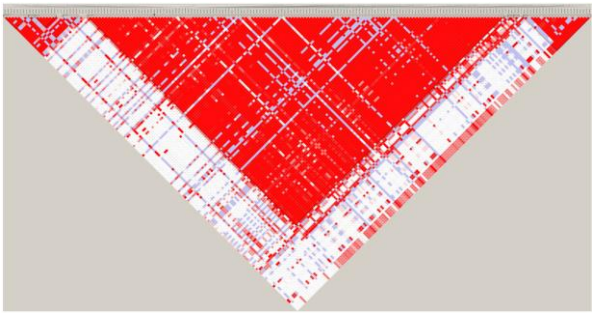
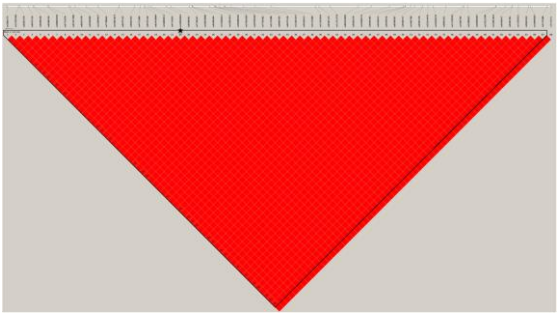


Fig 4.2a - Linkage Disequilibrium Plots for 1) *CR1*, 2) *CD2AP*, 3) *ABCA7*, and the 4)*TOMM40-APOE-APOC1* region constructed using the markers from ADNI data. The known significantly associated SNP(s) is indicated by a star. The markers used for the LD plot have MAF >= 0.05 for *CR1*, *CD2AP* and *TOMM40-APOE-APOC1* and a MAF >= 0.01 for *ABCA7*.

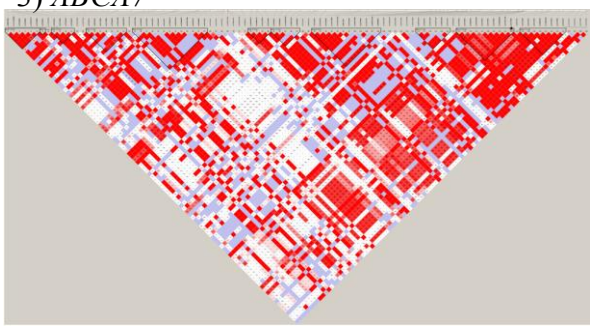
1) *CR1*



2) *CD2AP*



3) *ABCA7*



4) *TOMM40-APOE-APOC1*

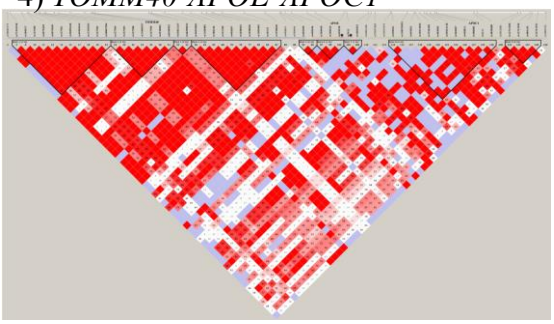


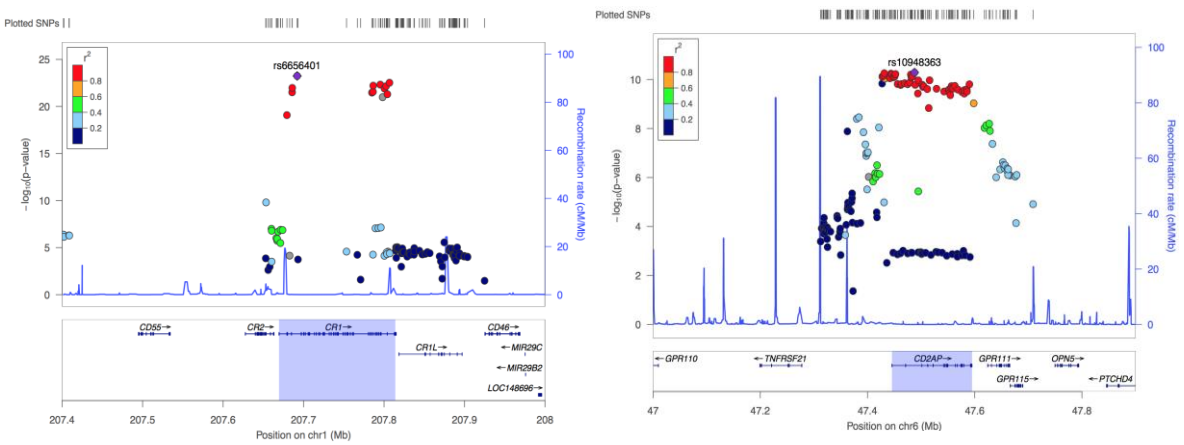
Fig 4.2b - Linkage Disequilibrium Plots for 1) *CR1*, 2) *CD2AP*, 3) *ABCA7*, and the 4)*TOMM40-APOE-APOC1* region constructed using the 1000 Genome markers. The known significantly associated SNP(s) is indicated by a star. The markers used for the LD plot have MAF >= 0.05.

4.2 Variants in the genes/gene region under consideration

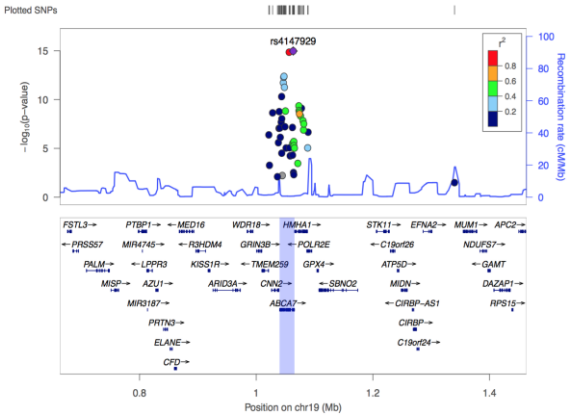
Table 4.2, gives the number of variants and number of samples obtained after data cleaning for each gene/gene region. The variants contain both the SNPs and Indels. The variants were further characterized based on their frequency in the data set. **Table 4.3**, gives the number and type of the variants in each region. **Table 4.4,** looks at how many of the variants are also mentioned in the dbSNP database (Version 144, GRCh37). It is probable that the variants which are not in the dbSNP database, may be the private variants.

a) *CR1*

b) *CD2AP*



c) *ABCA7*



d) *APOE* – rs429358

APOE – rs7412

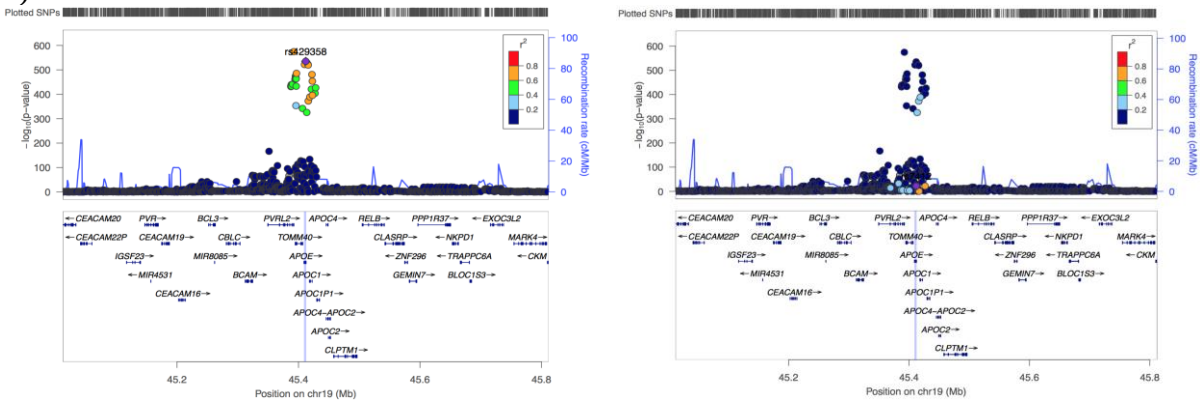


Fig 4.3 - Regional Association Plots for a) *CR1*, b) *CD2AP*, c) *ABCA7*, and the d) *TOMM40-APOE-APOC1* region constructed using the data from Lambert et al and LocusZoom.

Table 4.2 – Final number of variants and number of samples for each gene/gene-region

Gene Locus	*CR1*	*CD2AP*	*ABCA7*	*TOMM40-APOE-APOC1*
Number of Samples	808	808	658	786
Number of Variants	1181	1584	60	191

Table 4.3 - Type of Variants for each gene/gene-region

	Number of Rare Variants	Number of Low Frequency Variants	Number of Common Variants	Total Number of Variants
CR1	937	91	153	1181
CD2AP	1225	49	310	1584
ABCA7	43	4	13	60
TOMM40-APOE-APOC1	144	12	35	191

Table 4.4 - Number of dbSNP144-GRCh37 variants per gene/gene-region

	Number of dbSNP144.GRCh37 variants	Number of NOT in dbSNP144.GRCh37 variants	Total Number of Variants
CR1	564	617	1181
CD2AP	677	907	1584
ABCA7	39	21	60
TOMM40-APOE-APOC1	104	87	191

A plot of the variants versus the minor allele frequency, stratified on the basis of the variant membership in dbSNP144 (**Fig 4.4**), showed that majorly there were more

rare/low-frequency in the non dbSNP group than the dbSNP group. Overall, the total number of rare and low-frequency variants was more than the common variants.

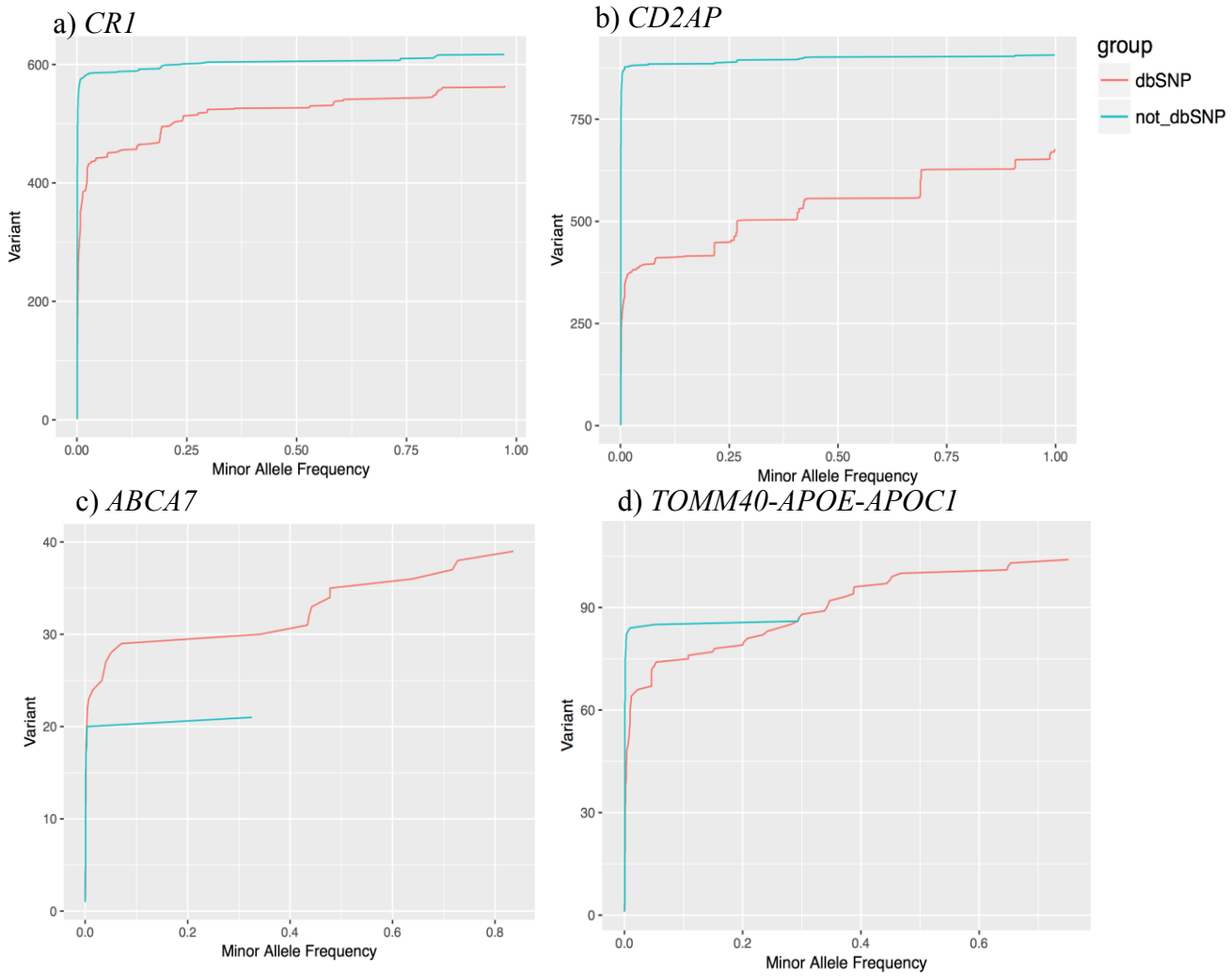


Fig 4.4 - Variant versus Minor Allele Frequency for b) *CD2AP*, c) *ABCA7*, and the d) *TOMM40-APOE-APOC1* region.

4.3 Principal Components in the ADNI genotype data

Principal Component Analysis (PCA) was done to determine the population structure in the ADNI data (**Fig 4.5**). The first two principal components were plotted and the self-reported race for the ADNI subjects was then compared with the race determined from the PCA. Most of the ADNI samples clustered with the CEU

HapMap samples (the Utah resident with Northern and Western European ancestry). All the samples were included in the final analysis. The population specific differences were adjusted for in the association analysis by using the first principal component (PC1) as the covariate in the association mode.

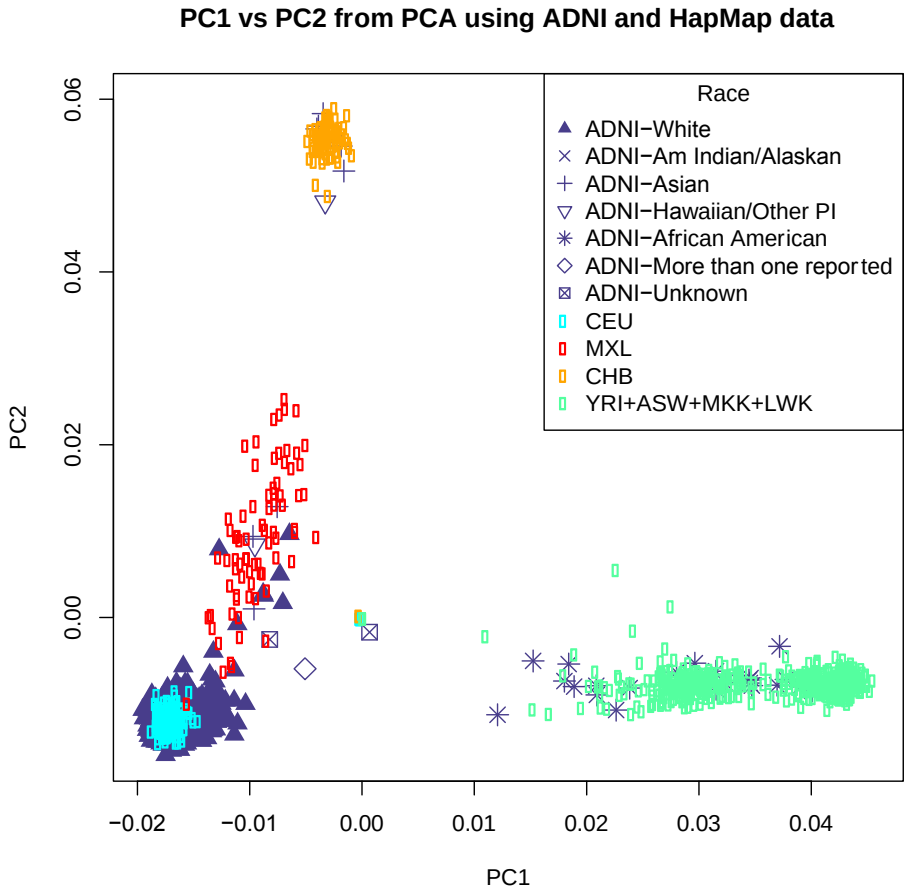


Fig 4.5 – Principal Components Analysis of the ADNI data using HapMap samples.

4.4 SNP-level Association Test

Logistic regression model was used to determine association of the individual variants in the above four genes/gene region with the AD status. Sex, age, and PC1 were used as the covariates. Five SNPs in the *APOE* and *APOC1* gene were observed

to significantly associated (Bonferroni adjusted $\alpha = 1.37 \times 10^{-5}$) with the AD status (**Fig 4.6**). **Table 4.5** gives the p-value of the known significantly associated SNPs.

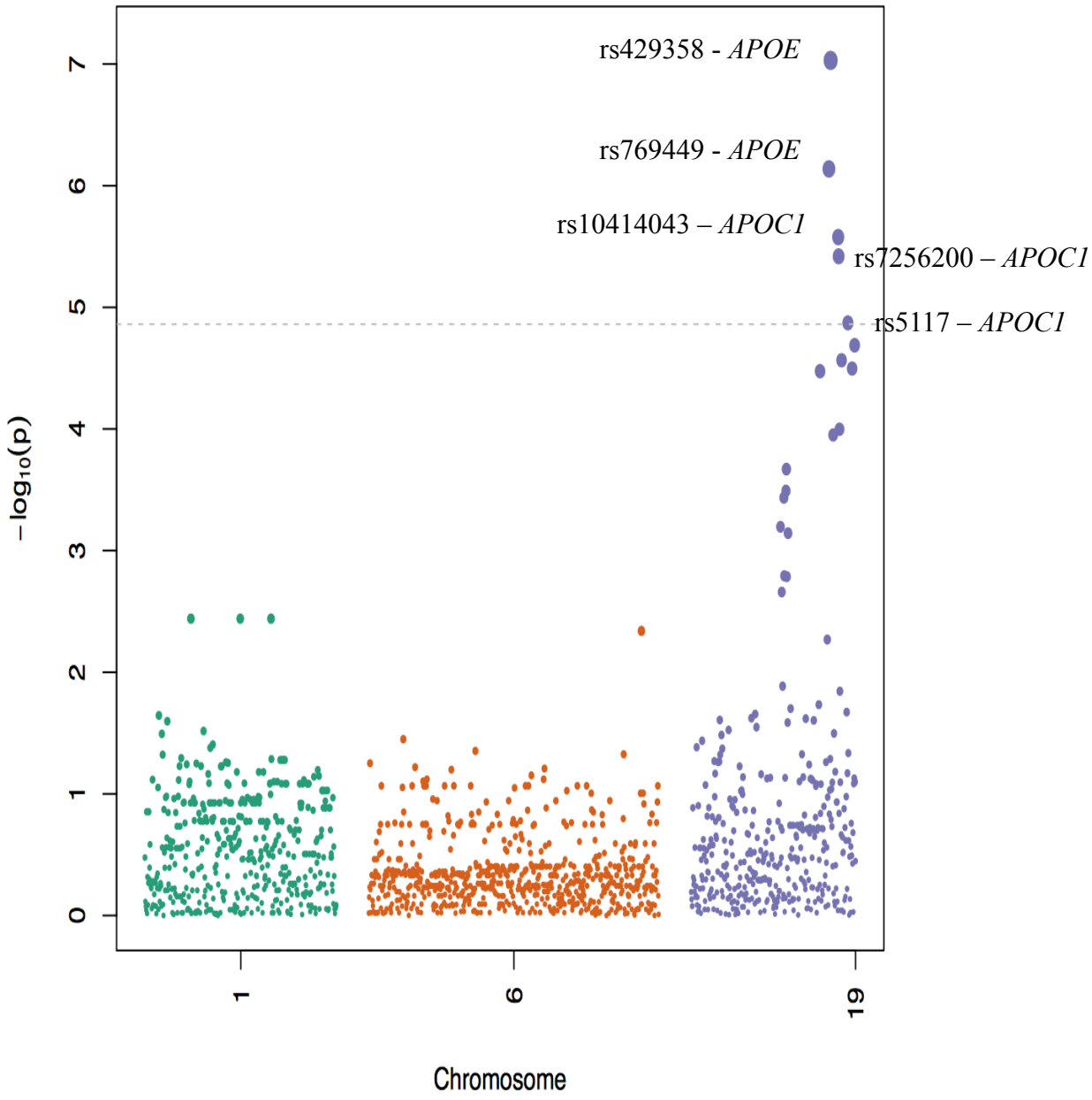


Fig 4.6 – Manhattan Plot showing the SNP-level Association Test

4.5 Region-level Association Test /SKAT-O Analysis

All the variants in the *CR1*, *CD2AP*, *ABCA7*, and *TOMM40-APOE-APOC1*, were aggregated to determine the association of the gene/gene–region as a whole with the AD status using the optimal unified sequence kernel association test or SKAT-O. Different types of variants for a gene region were aggregated to tease out the contribution of different types of variants to the association strength. for the gene-region level association testing.

We tested the association by aggregating all the variants (SNPs + Indels), only all the rare and low-frequency variants and only all the rare variants all over a gene/gene-region. **Table 4.5** gives the p-value obtained. Aggregating the variants over the whole gene/gene region resulted in a stronger association with the AD status for *CR1*, *CD2AP*, and *TOMM40-APOE-APOC1*.

Table 4.5 – Results from Variant Level and Gene Level Association Tests

	Chr	Known Significant SNP	Significant SNP p-value (Logistic Regression Model)	Gene level p-value (All Variants)	Gene level p-value (Rare + Low Frequency Variants)	Gene level p-value (Only Rare Variants)
CR1	1	rs6656401	0.083	0.024	0.029	0.05
CD2AP	6	rs10948363	0.831	0.5	0.78	0.78
ABCA7	19	rs4147929	0.319	0.775	0.51	0.79
TOMM40-APOE-APOC1	19	rs429358 rs7412	9.345231×10^{-8} 0.1625	6.98×10^{-6}	0.011	0.003

In order to test the role of the LD structure, all the variants within the different haplotype blocks were aggregated and the association with AD status was determined. For *CR1*, the haplotype block 2, which contains the known top hit (**Fig**

4.2a) showed the most significant association (p-value = 0.027; **Table 4.6**). Within, this block, the contribution of rare variants was also estimated (p-value = 0.062).

Table 4.6 – Gene level and Haplotype Block p-value

a) *CR1* – All Haplotype Blocks

p-value Whole Gene (1181 variants)	p-value Haplotype Block 1 (77 variants)	p-value Haplotype Block 2 (which has rs6656401) (913 variants)	p-value Haplotype Block 3 (106 variants)
0.024	0.938	**0.027**	0.121

***CR1* – Haplotype Block 2**

	p-value Haplotype Block2 (All 913 variants including rs6656401)	p-value Haplotype Block 2 (811 rare + low frequency variants)
	0.027	0.062

b) *TOMM40-APOE-APOC1* – All Haplotype Blocks

p-value Whole Gene Region (191 variants)	p-value Haplotype Block 1 (23 variants)	p-value Haplotype Block 2 (98 variants)	p-value Haplotype Block 3 (which has rs429358 and rs7412) (6 variants)	p-value Haplotype Block 4 (21 variants)
6.98×10^{-6}	0.0003	0.008	**1.72 x 10-7**	0.006

c) *ABCA7* – All Haplotype Blocks

p-value Whole Gene	p-value Haplotype Block 1 (13 variants)	p-value Haplotype Block 2 (which has rs4147929) (34 variants)
0.775	0.52	0.568

Similarly, for *TOMM40-APOE-APOC1* and *ABCA7* also, all the variants in the different haplotype blocks were aggregated and the association with AD status was determined (**Table 4.6**). For *TOMM40-APOE-APOC1*, as well, the most significant association with AD status was observed with the variants in the haplotype block containing the top hits. However, this block has only a total of 6 variants, and apparently the top hits which are common variants are driving the association. This region overall shows significant association (p-value = 6.98×10^{-6}; **Table 4.5**), with all the haplotype blocks individually also showing strong association (**Table 4.6**). Similar strength of association was observed for both the haplotype blocks of *ABCA7*.

Chapter 5: Conclusions

Significant association at the SNP level was observed for one of the known significantly associated SNP, *APOE*– rs429358. Lack of significant association with the other known top hits may be attributed to the small sample size. This may also due to some variants and samples being lost at the data cleaning step.

APOE is in LD with its neighboring genes, *TOMM40* and *APOC1*. We wanted to see how does aggregating all the variants in a gene/gene-region affect the strength of association relative to the SNP-level association strength. The underlying hypothesis here being that, a known significant SNP, might act as tag for other important/causal variants in a region owing to the LD structure. These other tagged variants might themselves not be able to achieve significance at an individual level, but when aggregated together might add to the strength of the association.

Significant association for the whole *TOMM40-APOE-APOC1* region was observed upon aggregating all the variants (SNPs + Indels) (p-value = 6.98×10^{-6}), aggregating only all rare and low-frequency variants (p-value = 0.011), and aggregating only the rare variants (p-value = 0.003). Since, significant association with the region was observed even when only the rare variants were aggregated, it is highly suggestive of the presence of some rare variant(s) associated with Alzheimer's (Table 4.5). Considering the association strength of the haplotype blocks individually, we found that though the common SNPs drive the association within the haplotype block containing the top hits – rs429358 and rs7412, the whole *TOMM40-APOE-APOC1* region, seems to be significantly associated with the AD phenotype. Since the other haplotype blocks (which also show significant association) are made up of

common, low-frequency, and rare variants, it will be interesting to tease out the contribution of the rare and low-frequency variants only to the signal strength being observed.

With *CR1* gene, an improvement in strength of association was observed upon aggregating all the variants (p-value = 0.024) as well as on aggregating only the rare and low-frequency variants (p-value = 0.029), compared to association observed at the individual SNP level (p-value = 0.083). We also assessed the role of LD structure in determining the region over which to aggregate the rare variants. Interestingly, with *CR1,* we did observe the most significant contribution from the haplotype block (p-value Haplotype Block 2 = 0.027 compared to p-value = 0.938 (Haplotype Block 1) and p-value = 0.121 (Haplotype Block 3), which contains the top hit. We then evaluated the contribution of the rare and low-frequency variants within this haplotype block. We observed a distinct and well defined contribution from the rare variants to the overall strength of association with the AD phenotype (p-value Haplotype Block 2, all variants = 0.027, p-value Haplotype Block 2, rare and low-frequency variants = 0.062).

The *ABCA7* gene has no/very little LD (Fig. 4.2 a & b). As a result of this, the signals from the variants within this region are independent. The p-value from the most significant SNP, even if all the variants are aggregated over the weak haplotype blocks (Table 4.5 and 4.6).

For *CD2AP*, adding in all the variants results in a lower p-value than that obtained by the top hit alone (Table 4.5). This suggests that all the variants in this region – common, rare, and low-frequency contribute to the signal strength.

These results serve as a proof of concept, the rare causal variants for a phenotype can be identified by aggregating the variants over a genomic region. Aggregating the variants adds to the signal strength and ups the probability of identifying a significant association with a phenotype. Rare and low-frequency variants complement the signal strength association from a common SNP. In addition, the LD structure of a gene/gene region, can be leveraged in order to determine the region over which to aggregate the rare and low-frequency variants.

Once a region having a significant association with a phenotype is determined, there are algorithms which can help to identify a set of potential causal variants in the region. One of these algorithms, works backwards, eliminating one variant at time and looks at the effect of the eliminating the variant. If upon eliminating a variant, the strength of association goes down, then it indicates that the eliminated variant, might be the rare causal variant associated with the phenotype/outcome (Ionita-Laza, Capanu, Rubeis, McCallum, & Buxbaum, 2014).

Further work would involve looking at aggregating variants over several of other AD genes such as *BIN1, EPHA1, CLU, PICALM, SORL1, INPP5D, MEF2C, NME8, CELF1, CASS4,* and *ZCWPW1* based on their LD structure.

References

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*(2), 263–265.

Carson, A. R., Smith, E. N., Matsui, H., Brækkan, S. K., Jepsen, K., Hansen, J.-B., & Frazer, K. A. (2014). Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics*, *15*, 125.

Chartier-Harlin, M. C., Parfitt, M., Legrain, S., Pérez-Tur, J., Brousseau, T., Evans, A., Gourlet, V. (1994). Apolipoprotein E, epsilon 4 allele as a major risk factor for sporadic early and late-onset forms of Alzheimer's disease: analysis of the 19q13.2 chromosomal region. *Human Molecular Genetics*, *3*(4), 569–574.

Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.-C., Carrasquillo, M. M., Williams, J. (2011). Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nature Genetics*, *43*(5), 429–435.

Ionita-Laza, I., Capanu, M., Rubeis, S. D., McCallum, K., & Buxbaum, J. D. (2014). Identification of Rare Causal Variants in Sequence-Based Studies: Methods and Applications to VPS13B, a Gene Involved in Cohen Syndrome and Autism. *PLOS Genetics*, *10*(12), e1004729.

Lambert, J.-C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., Amouyel, P. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature Genetics*, *41*(10), 1094–1099.

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

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Lin, X. (2012). Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *The American Journal of Human Genetics*, *91*(2), 224–237.

Li, B., & Leal, S. M. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *American Journal of Human Genetics*, *83*(3), 311–321.

Long, T., Hicks, M., Yu, H.-C., Biggs, W. H., Kirkness, E. F., Menni, C., Telenti, A. (2017). Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nature Genetics*, *49*(4), 568–578.

Lord, J., Lu, A. J., & Cruchaga, C. (2014). Identification of rare variants in Alzheimer's disease. *Frontiers in Genetics*, *5*.

Lyall, D. M., Harris, S. E., Bastin, M. E., Muñoz Maniega, S., Murray, C., Lutz, M. W., Deary, I. J. (2014). Alzheimer's disease susceptibility genes APOE and TOMM40, and brain white matter integrity in the Lothian Birth Cohort 1936. *Neurobiology of Aging*, *35*(6), 1513.e25-1513.e33.

Naj, A. C., Jun, G., Beecham, G. W., Wang, L.-S., Vardarajan, B. N., Buros, J., Schellenberg, G. D. (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nature Genetics*, *43*(5), 436–441.

Nicolae, D. L. (2016). Association Tests for Rare Variants. *Annual Review of Genomics and Human Genetics*, *17*(1), 117–130.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909.

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Willer, C. J. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, *26*(18), 2336–2337.

Querfurth, H. W., & LaFerla, F. M. (2010). Alzheimer's Disease. *New England Journal of Medicine*, *362*(4), 329–344.

Roses, A., Sundseth, S., Saunders, A., Gottschalk, W., Burns, D., & Lutz, M. (2016). Understanding the genetics of APOE and TOMM40 and role of mitochondrial structure and function in clinical pharmacology of Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *12*(6), 687–694.

Ubelmann, F., Burrinha, T., Salavessa, L., Gomes, R., Ferreira, C., Moreno, N., & Almeida, C. G. (2017). Bin1 and CD2AP polarise the endocytic generation of beta-amyloid. *EMBO Reports*, *18*(1), 102–122.

Zhao, Q.-F., Yu, J.-T., Tan, M.-S., & Tan, L. (2015). ABCA7 in Alzheimer's Disease. *Molecular Neurobiology*, *51*(3), 1008–1016.

Zhou, Q., Zhao, F., Lv, Z., Zheng, C., Zheng, W., Sun, L., Yang, Z. (2014). Association between APOC1 Polymorphism and Alzheimer's Disease: A Case-Control Study and Meta-Analysis. *PLoS ONE*, *9*(1).

Zhu, X.-C., Yu, J.-T., Jiang, T., Wang, P., Cao, L., & Tan, L. (2015). CR1 in Alzheimer's Disease. *Molecular Neurobiology*, *51*(2), 753–765.

Appendix

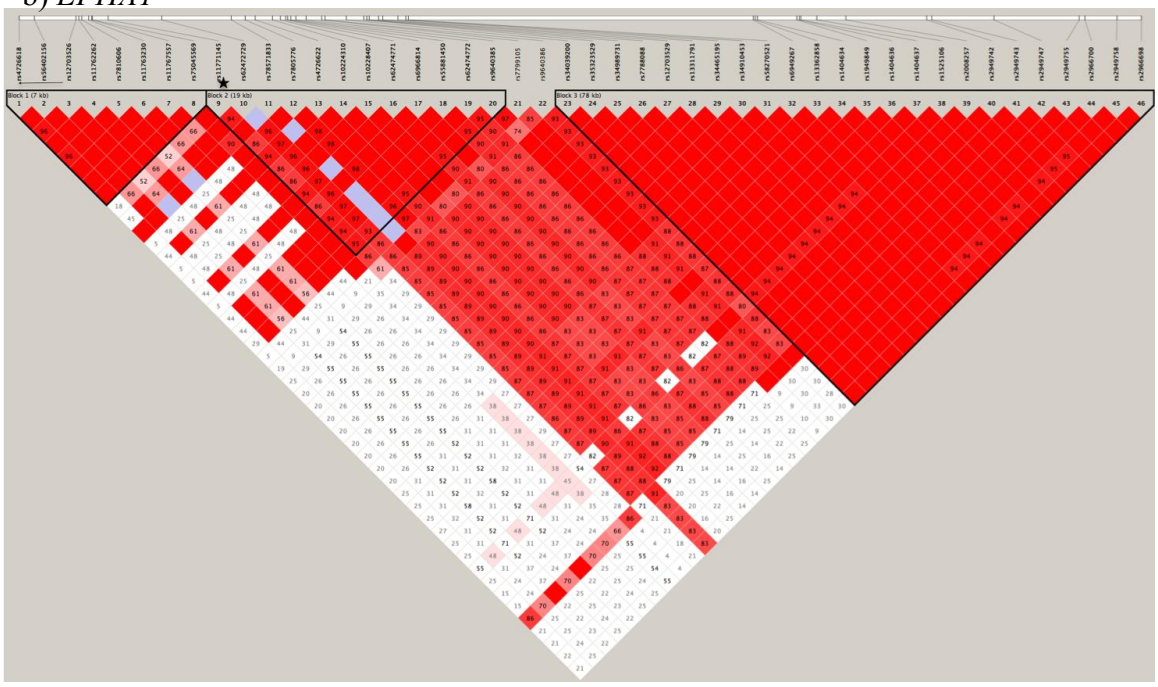
Appendix A: Linkage Disequilibrium (LD) Structure of AD associated genes.

Appendix B: Regional Association Plots for AD associated genes.

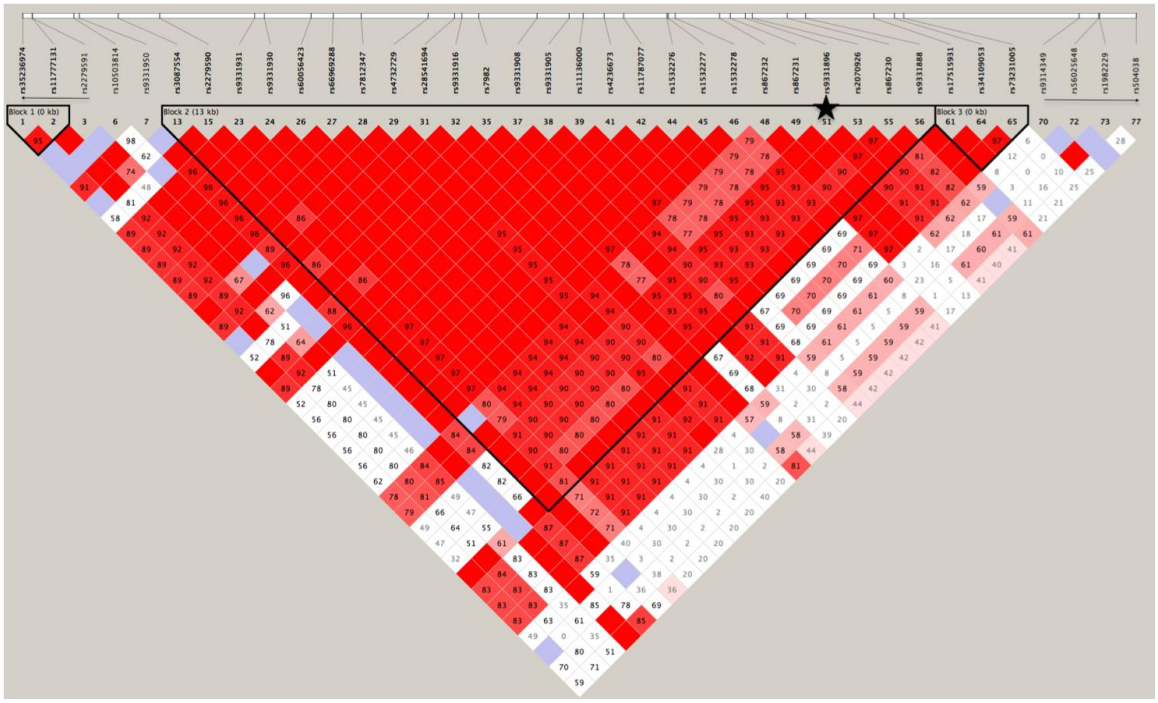
Appendix A

a) *BIN1*

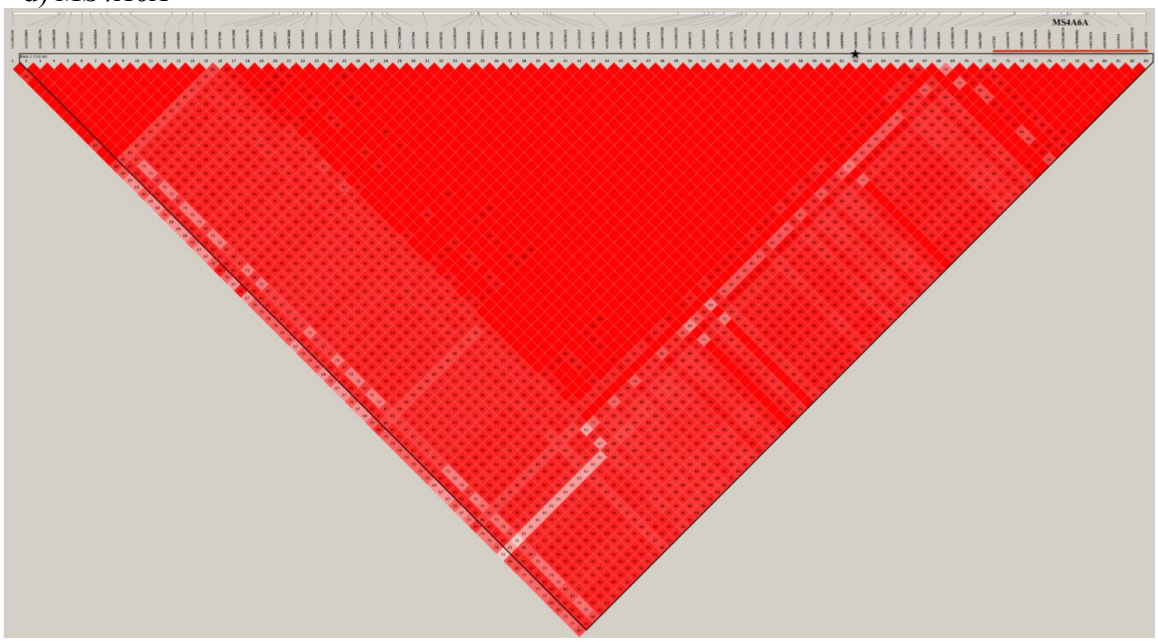
b) *EPHA1*



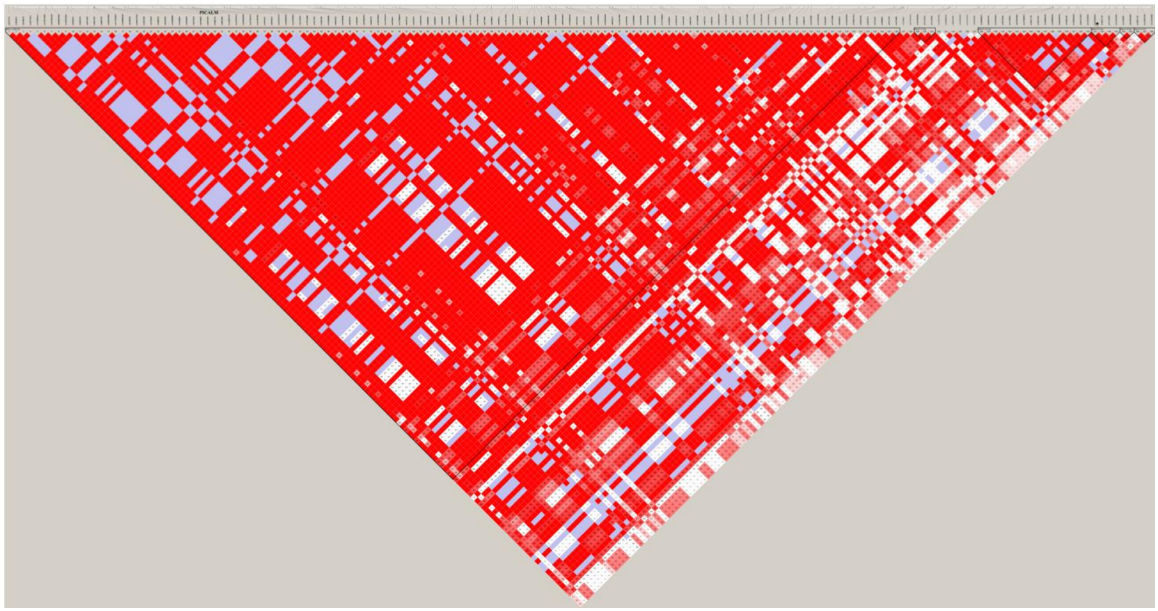
c) *CLU*



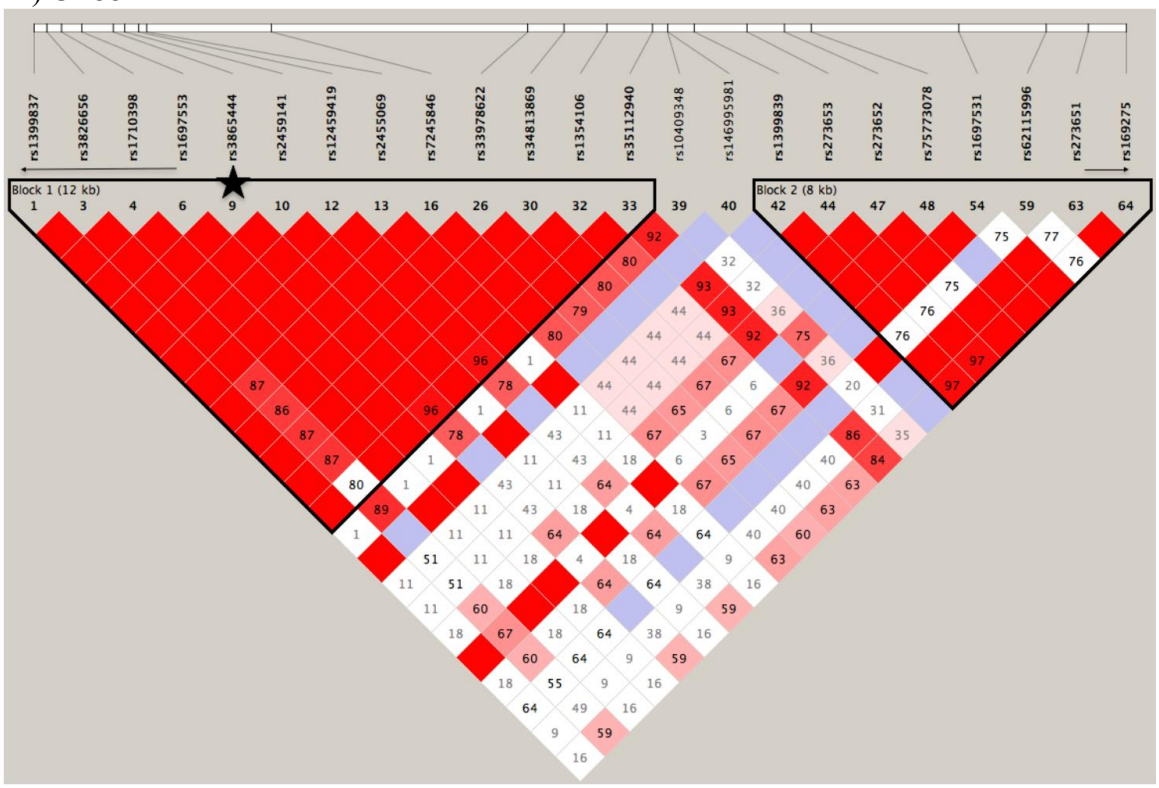
d) *MS4A6A*



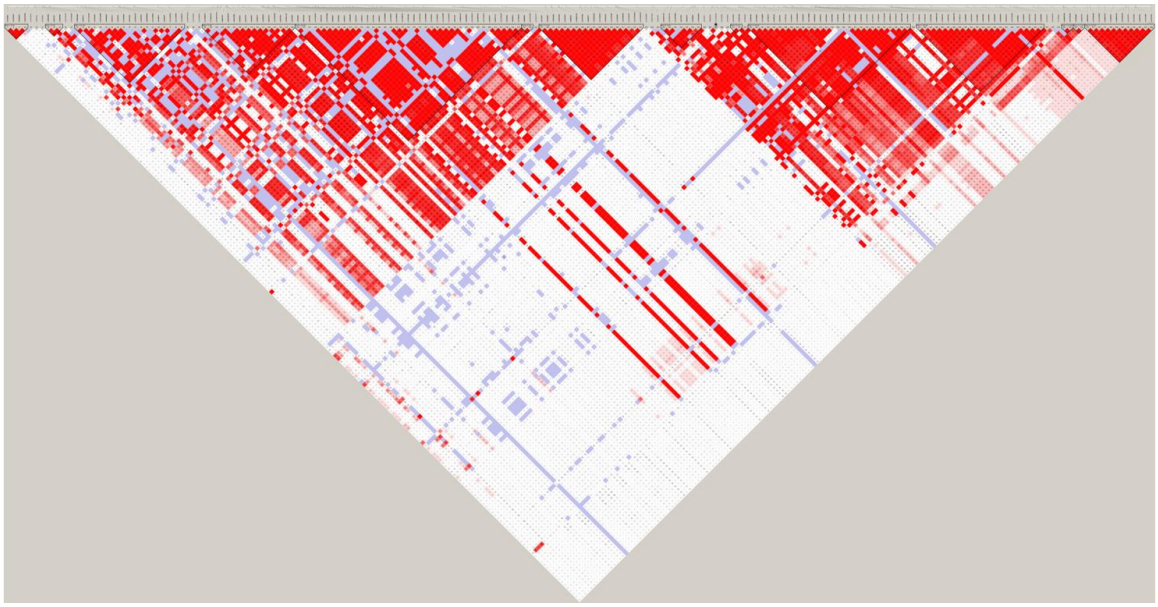
e) *PICALM*



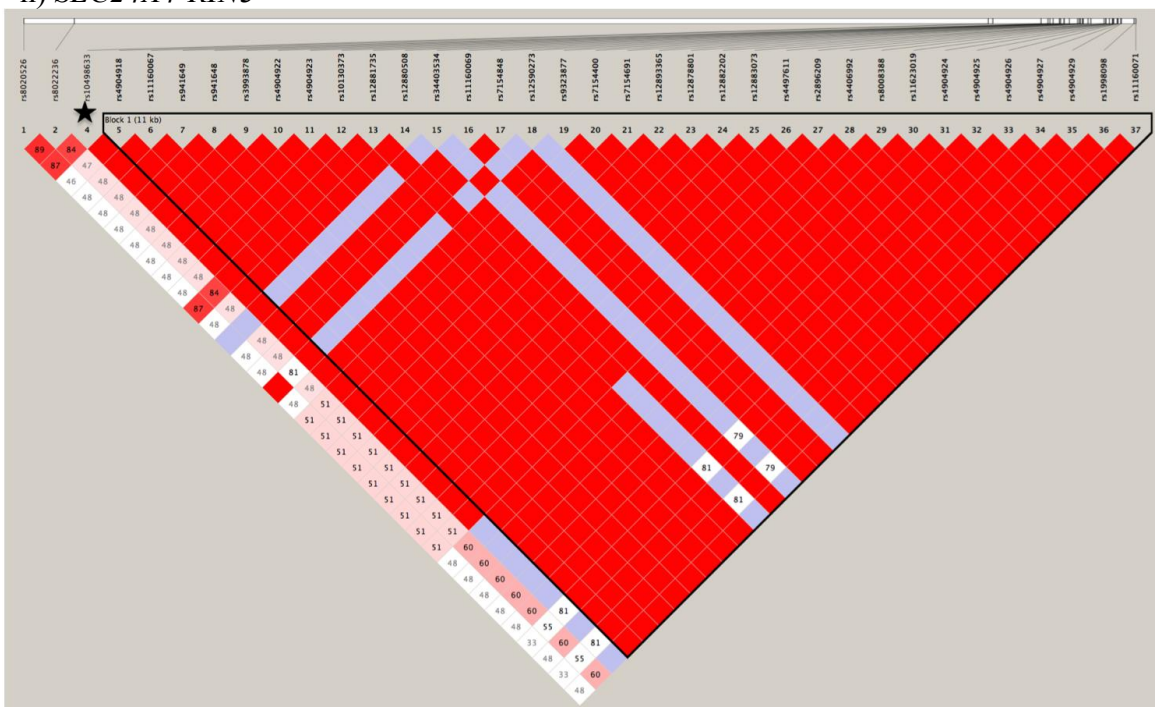
f) *CD33*



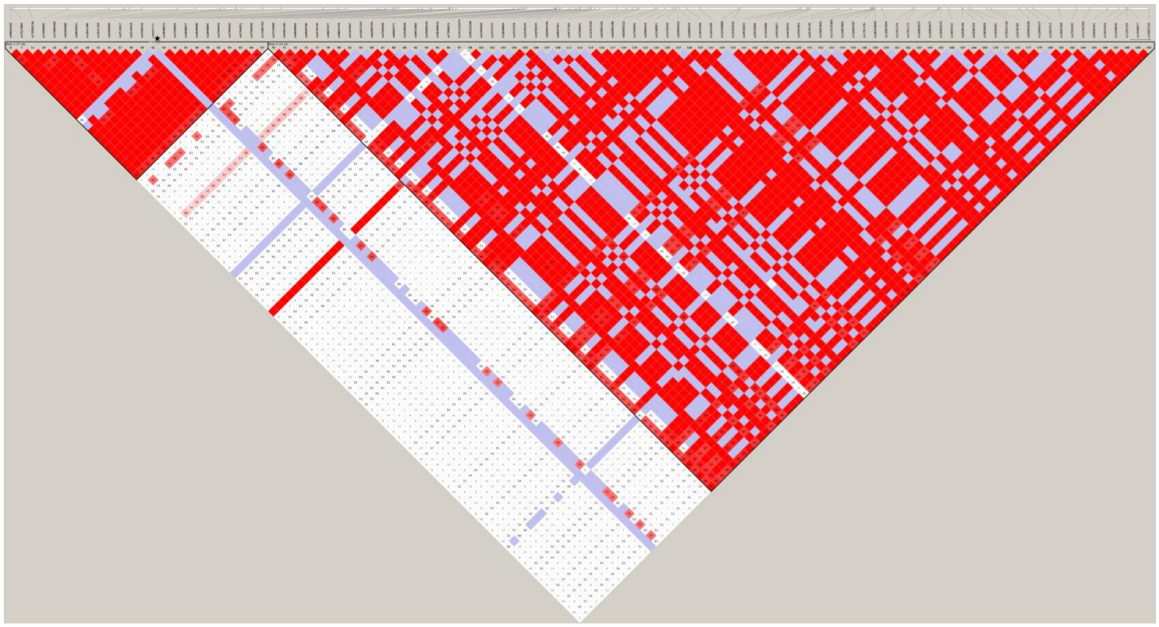
g) *SORL1*



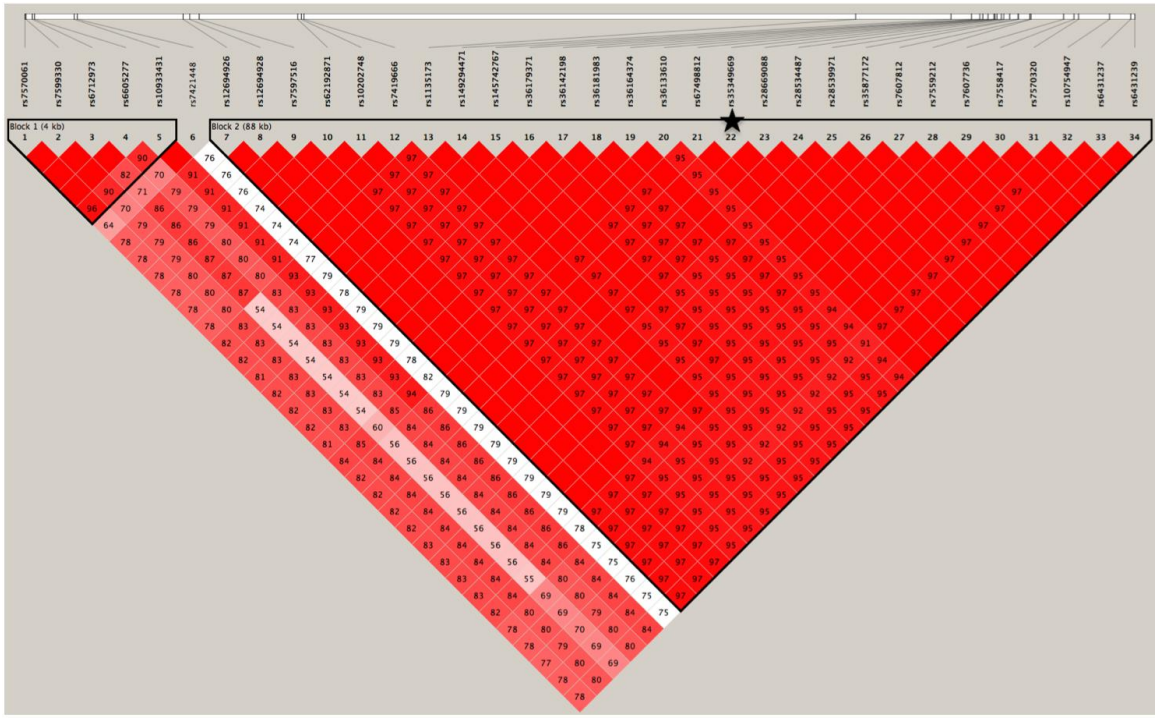
h) *SLC24A4-RIN3*



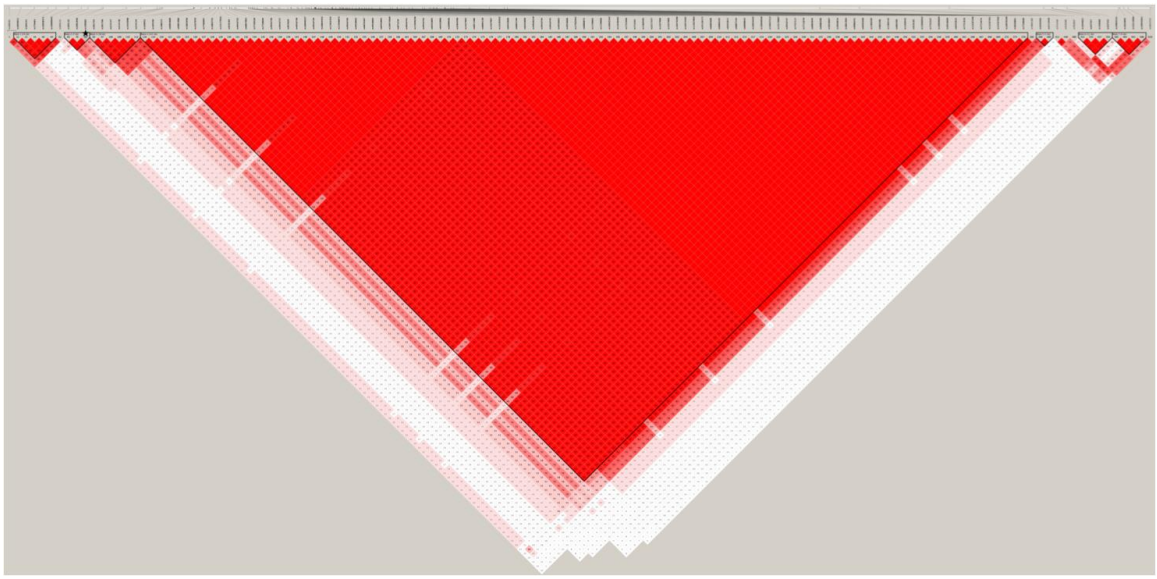
i) *DSG2*



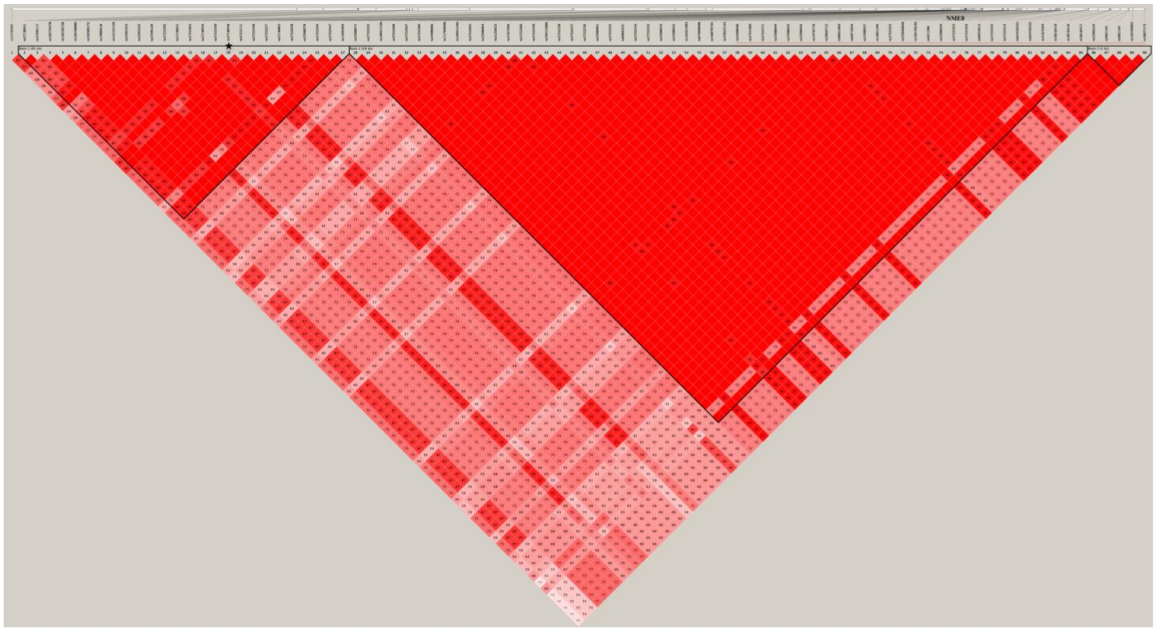
j) *INPP5D*



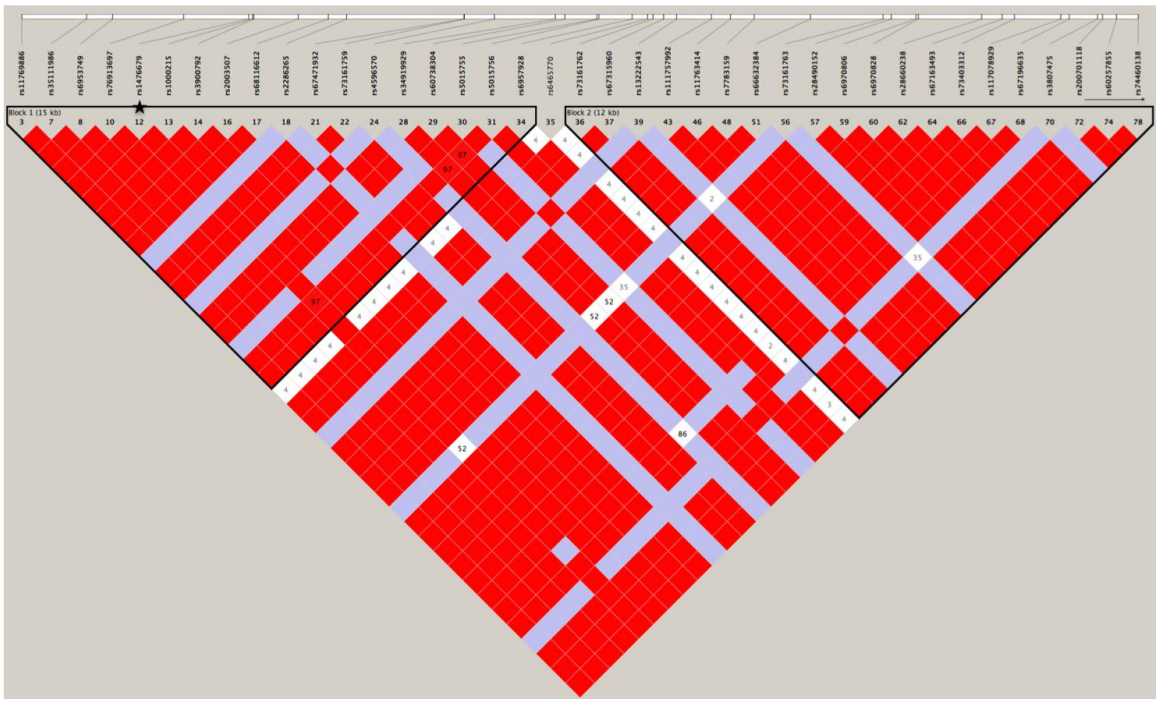
k) *MEF2C*



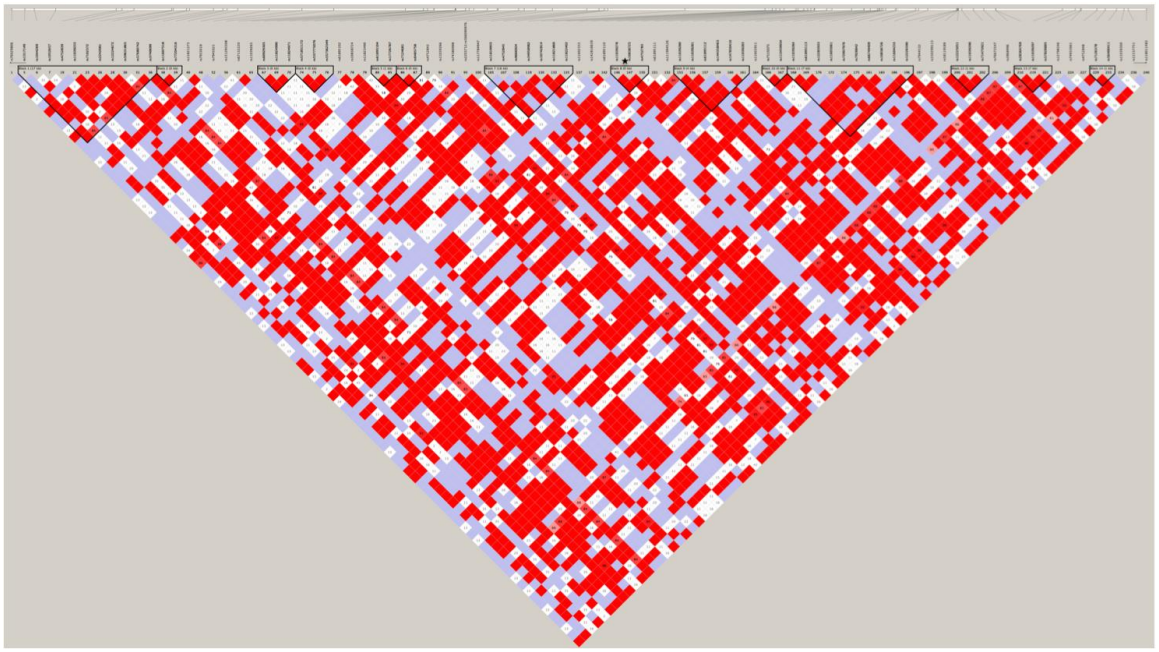
l) *NME8*



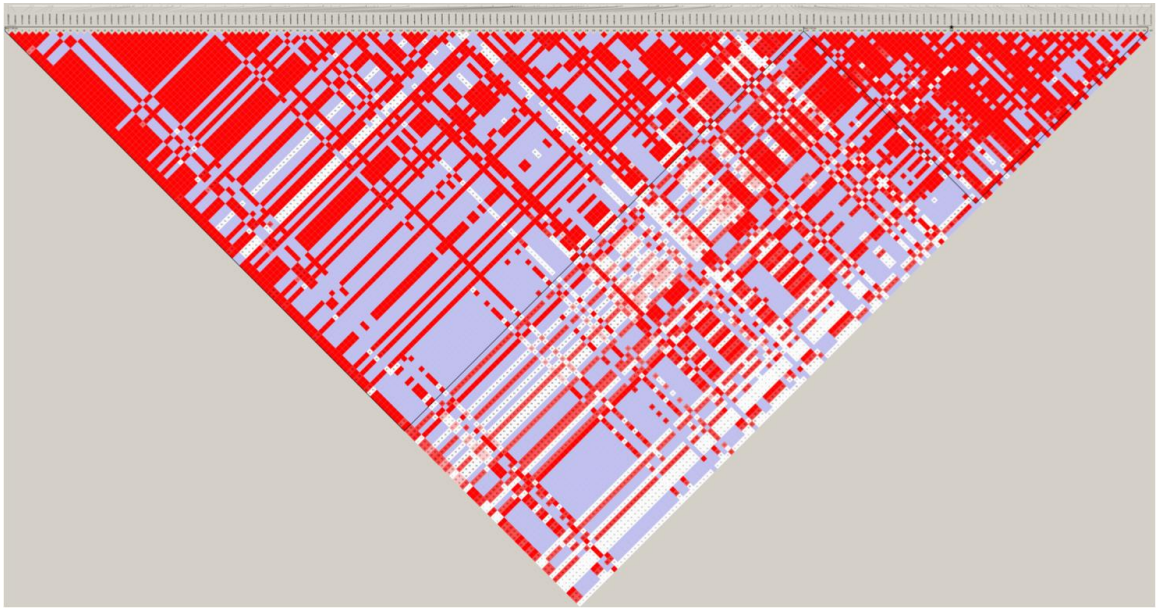
m) *ZCWPW1*



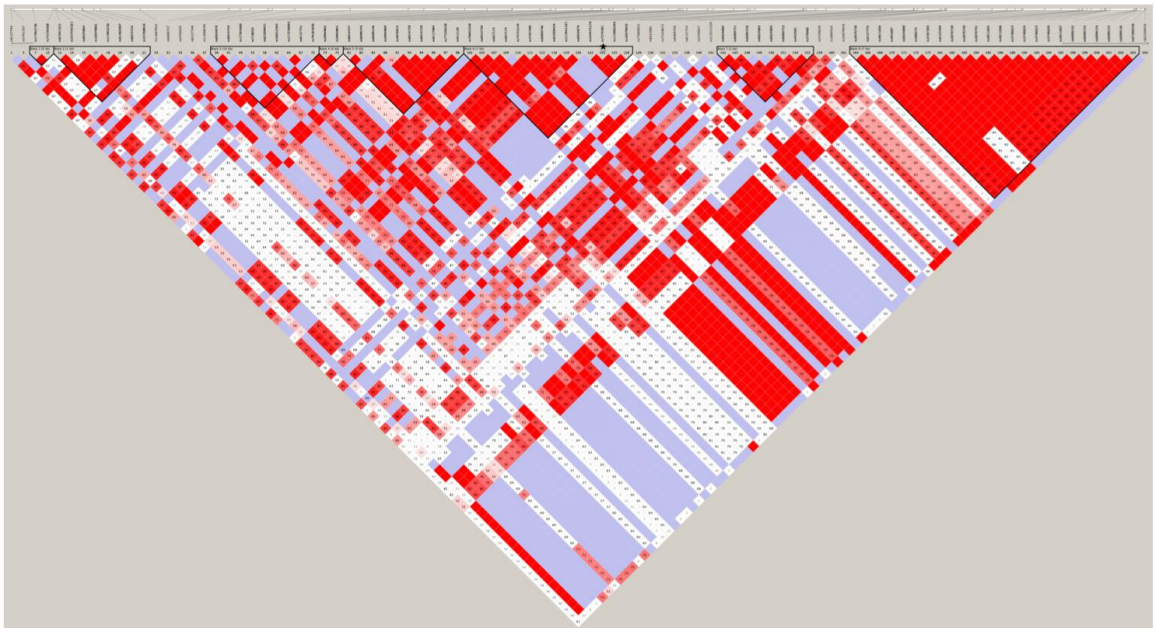
n) *CELF1*



o) *FERMT2*

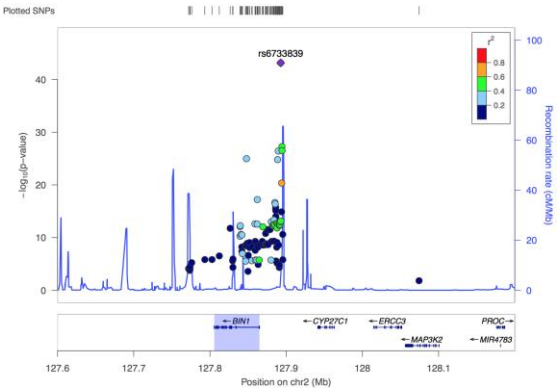


p) *CASS4*

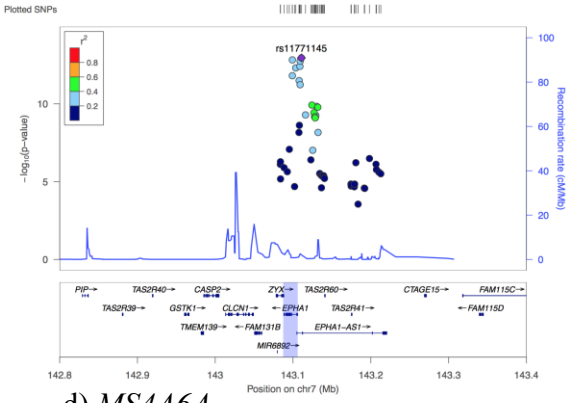


Appendix B

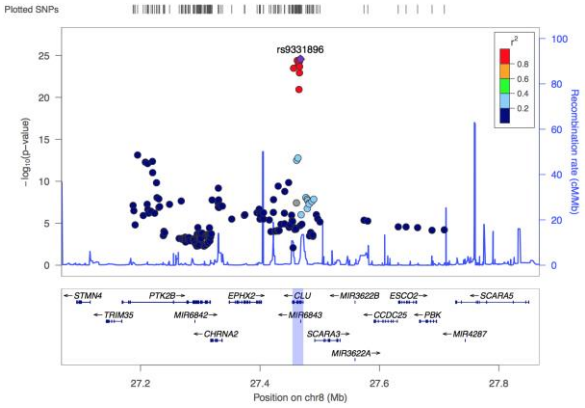
a) *BIN1*



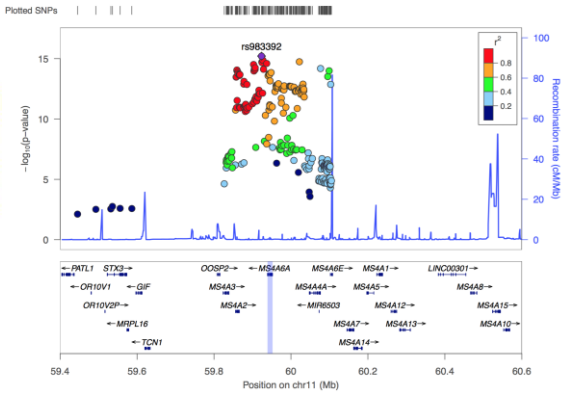
b) *EPHA1*



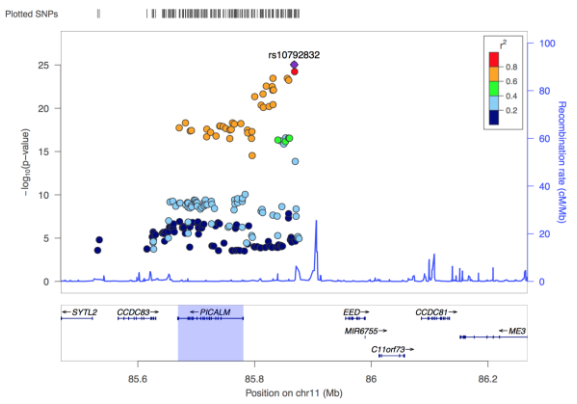
c) *CLU*



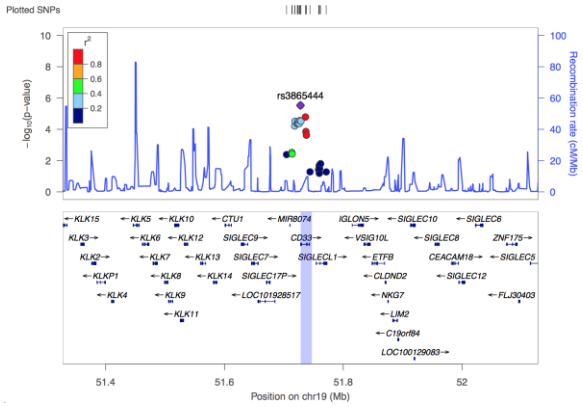
d) *MS4A6A*



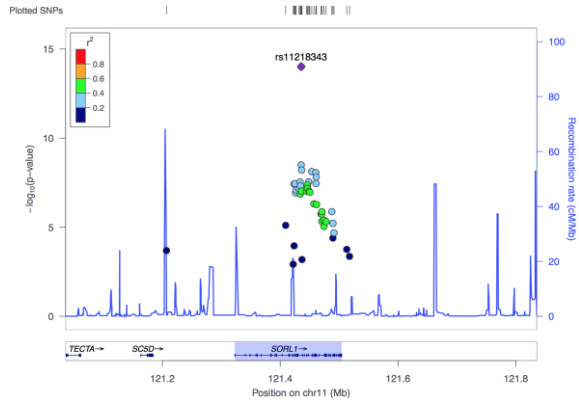
e) *PICALM*



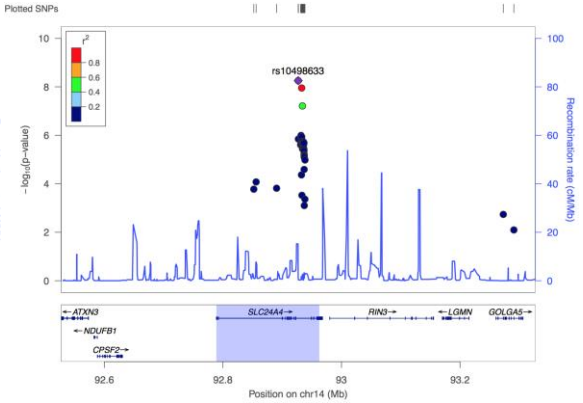
f) *CD33*



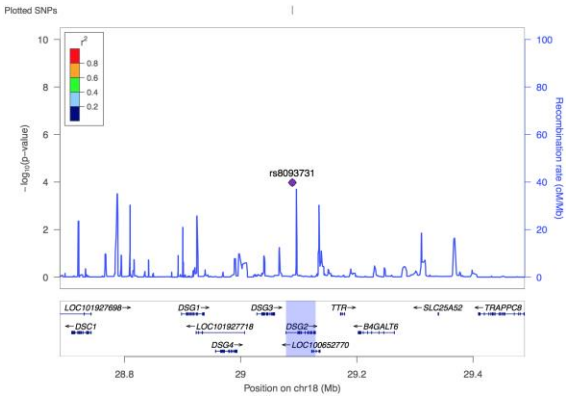
g) *SORL1*



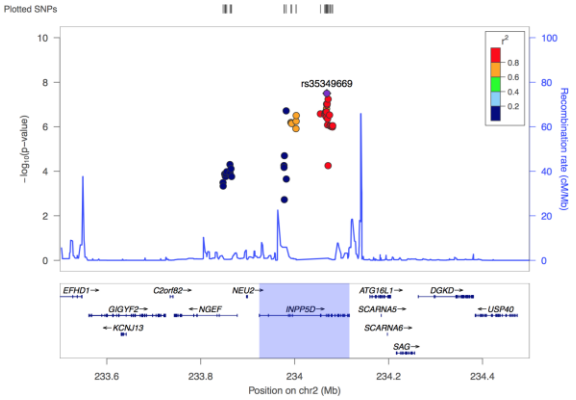
h) *SLC24A4-RIN3*



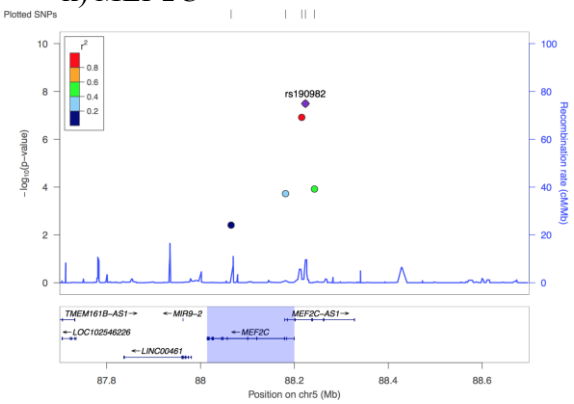
i) *DSG2*



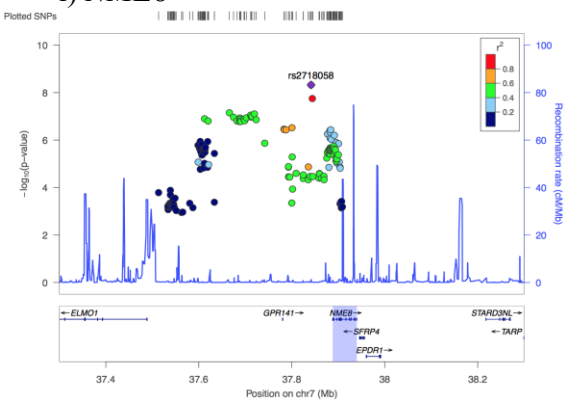
j) *INPP5D*



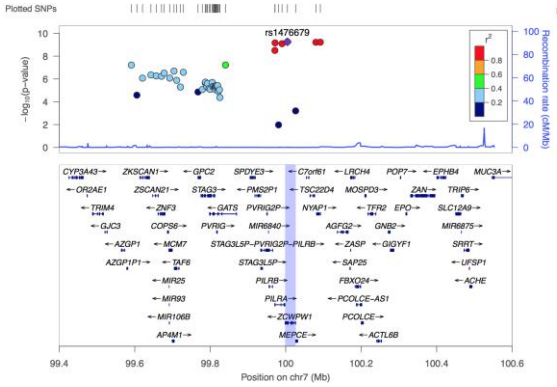
k) *MEF2C*



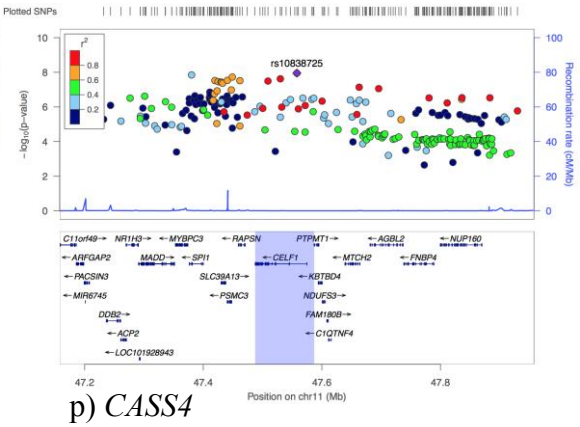
l) *NME8*



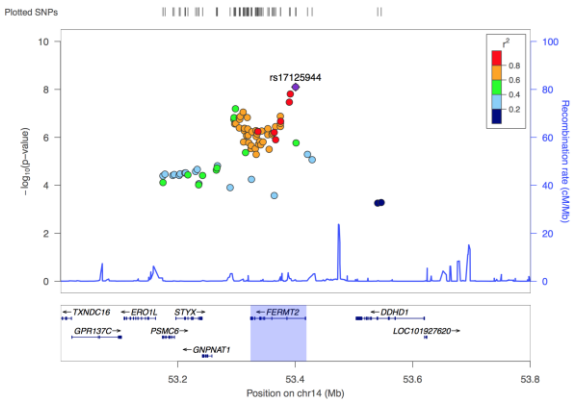
m) *ZCWPW1*



n) *CELF1*



o) *FERMT2*



p) *CASS4*