Assessing multivariate analysis of GWAS for identification of genetic variants in Alzheimer's Disease

By

Priya Bhatt

A THESIS

Presented to the Department of Medical Informatics and Clinical Epidemiology

and the Oregon Health & Science University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Master of Science

November 2012

School of Medicine

Oregon Health and Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Priya Bhatt

has been approved

_____

Dr. Beth Wilmot
Mentor/Advisor

_____

Dr. Shannon McWeeney
Committee Member

_____

Dr. Eilis Boudreau
Committee Member

_____

Dr. Deniz Erten-Lyons
Committee Member

Acknowledgements

I would like to thank my thesis advisor, Dr. Beth Wilmot, for her guidance, mentorship and support during the duration of this study. In addition, I would like to extend my gratitude to my committee members, Dr. Deniz Erten-Lyons, Dr. Shannon McWeeney and Dr. Eilis Boudreau, for providing their expertise, encouragement and feedback. I would like to appreciate Oregon Health and Science University, particularly the Department of Medical Informatics and Clinical Epidemiology including the student body, faculty and staff for their support.

Finally, I would like to thank my parents for their unconditional love, patience and direction. Mom and Dad, I dedicate this thesis to you.

Table of Contents

List of Tables

List of Figures

Abstract

Assessing multivariate analysis of GWAS for identification of genetic variants in Alzheimer's Disease

Priya Bhatt

MS, Department of Medical Informatics and Clinical Epidemiology
at Oregon Health and Science University

November 2012

Thesis Advisor: Dr. Beth Wilmot

Alzheimer's Disease (AD) is the leading cause of dementia in the United States yet the genetics behind this complex disease remains unclear. With the exception of *Apolipoprotein E e4* (APOE-e4), more than 40 loci have been implicated as common genetic risk factors of AD but none of these have been confirmed. We completed a genome wide association study on 567 unrelated participants in the Alzheimer's Disease Neuroimaging Initiative (ADNI) data set. DNA samples were genotyped with the Illumina Human610-Quad BeadChip and 543,715 single nucleotide polymorphisms (SNPs) were included after undergoing quality control measures. Genome-wide association studies (GWASs) have successfully identified genetic associations to individual phenotypes in a univariate framework across many complex diseases including AD. However, the effort to detect pleiotropic associations, where multiple traits are associated with the same genetic loci, is far less common and has never been tried in an AD GWAS. Two multivariate methods, Principle Components Analysis (PCA) and Seemingly Unrelated Regression (SUR), were employed to determine the genetic

association of three quantitative correlated endophenotypes of AD. The PCA method incorporated hippocampal volume, ventricular volume and cognitive memory tests and the SUR method included hippocampal volume and cognitive memory tests. Our study identified 23 unique SNPs, with six SNPs found in common between both methods after adjusting for any biases. The PCA method found 21 SNPs (p-value < 10^{-5}) and the SUR method found eight SNPs (p-value < 10^{-5}). All of the identified genes have not been otherwise linked to AD, indicating a multivariate framework can provide new insight to genetic research of these phenotypes and AD.

Chapter 1: Introduction

1.1 Importance of determining novel genetic variants in AD

Alzheimer's Disease (AD) accounts for 50 to 80 percent of dementia in the United States.[1] In addition to AD's obvious mental and physical burden on patients, total payments for care average to three times higher for AD patients than for non-AD patients. A recent report by the Alzheimer's Association estimates the number of AD patients to nearly quadruple by 2050, thus creating a sense of urgency in AD research.[1] The general trend for current genome wide association studies (GWASs) of complex diseases is to collect multiple phenotypes of interest from a single study population and analyze the phenotypes individually using univariate analysis approach. This framework is limited because it ignores the possible genetic correlations between different traits that can help detect genes that have an effect on these multiple traits. These genes can be important components to understanding a given complex disease, and in our case, AD.[2] On the other hand, studying these traits jointly in a multivariate framework, where we predict the relationship of genetic variants to a number of traits together, can provide an insight to a complex disease that are not otherwise evident using a univariate approach. It has been long hypothesized that pleiotropy, where a genetic variation is associated with multiple traits, is an abundant phenomenon in complex diseases. By using the multivariate analysis methods to study these pleiotropic effects, we were able to determine novel genetic variants for AD while implementing and comparing two different multivariate analysis methods to suggest a promising

multivariate framework. Applying multivariate analysis methods on GWAS studies for complex diseases, allows future researchers to find additional genetic variants to enhance their knowledge of their disease of interest.

1.2 Multivariate methods are an innovative approach in AD research

Although past studies have been successful in finding novel genetic variants associated with AD, it is evident that there is much work to be done.[3,4] Univariate analysis methods have provided AD researchers with promising results, however we know this is suboptimal because published research in other complex diseases have found additional results by jointly analyzing the correlated phenotypes. By using a multivariate framework, we are able detect pleiotropic associations, where multiple traits are associated with the same genetic loci. Discovering and understanding these pleiotropic effects can provide essential clues to the nature and function of the genes associated with AD. Previous studies strongly suggest that multivariate analysis methods can identify novel genetic variants associated with AD and push the field of bioinformatics research in GWAS for complex diseases.

1.3 Specific Aims

There were two specific aims of this study.

1. Identify multiple outcomes as clinical important endophenotypes of AD within an AD GWAS data set.
2. Identify novel genetic variants by comparing the results from two different multivariate analysis techniques in the AD GWAS data set.

Chapter 2: Background

2.1 Understanding Alzheimer's Disease

Alzheimer's Disease (AD) is the leading cause of dementia in individuals over the age of 65 and is the sixth leading cause of death overall in the United States.[1,5] Affecting approximately 5.4 million Americans today .[1], AD is a progressive disease gradually worsening over time. Patients can live anywhere between four and 20 years from the onset of symptoms depending on age, additional health conditions and the severity of their AD.[1] There currently is no cure for AD and despite its prevalence in today's world, much about AD remains unknown.

2.1.1 Neuropathology of AD

The brain has over 100 billion neurons that connect and communicate with each other to create complex networks involved in specific functions and tasks. It is well known that AD is characterized by the loss to these neurons and synapses in the cerebral cortex and certain subcortical regions in the brain. Along with the loss of neurons, there is a build-up of amyloid plaques and neurofibrillary tangles found in the neuropathology of patients with AD. Amyloid plaques consist of dense, insoluble deposits of the beta-amyloid protein. Beta-amyloid is a part of a larger protein, *amyloid-precursor protein* (APP), which protrudes through the neuron membrane. An enzyme divides APP, creating the beta-amyloid fragments that migrate together to form the insoluble plaques. It is unclear whether the excessive amount of beta-amyloid

protein or a malfunctioning enzyme leads to the formation of amyloid plaques; however research shows that these plaques are present in AD patients.[6] Neurofibrillary tangles are insoluble twisted fibers made of *tau*, the protein that forms microtubules. Mutations of the *tau* protein, as seen in the pathology of AD patients, lead to the microtubule structures to collapse and result in tangles.[7]

2.1.2 Risk Factors of AD

A number of risk factors have been identified with AD, where age, family history and the presence of the *Apolipoprotein E e4* (APOE-e4) allele are arguably the most significant. Though AD is not a part of normal aging, the risk of developing the disease doubles every five years after the age of 65 and nearly half of those over the age of 85 have AD.[8] APOE-e4, a form *Apolipoprotein E* found on chromosome 19, has been found to be strongly associated with a higher risk for AD.[9] One allele of APOE-e4 increases the risk of AD by four and two copies of the APOE-e4 allele increases the risk of AD by ten. In addition, possession of two copies of the APOE-e4 allele increases the chance of AD symptoms occurring at a younger age.

2.1.3 Genetics of AD

As of late 2011, 15 GWASs have published results in the field of AD research. All GWAS results were detected utilizing a univariate framework. Through these univariate analyses of these GWASs, researchers have been able to identify approximately 40 AD susceptible loci other than APOE-e4. Recently, however, four additional genes show significant evidence to be susceptible genes for AD: CLU, CR1, PICALM and BIN1.[10,11]

APOE-e4 contributes to approximately 27% of the attributed risk where collectively the other susceptible genes make up for approximately 20% of the attributed risk of AD.[11] It is clear, however, APOE remains the single most important genetic risk factor for AD to date.

2.1.4 Clinical Criteria for the Diagnosis of AD

The National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disease Association (ADRDA) developed criteria for the diagnosis of AD in 1984. These criteria were recently updated in 2011 because of the significant research advancements that can now contribute to the diagnosis of AD.[13] Here, we will step through the outlined criteria by the NINCDS-ADRDA.[14]

In order to diagnose AD dementia, patients must first meet the criteria of all-cause dementia. Dementia is diagnosed when a patient expresses cognitive or neuropsychiatric symptoms that interfere with their ability to function doing daily activities or at work and, over time, show a declining level of functionality of these activities. In addition, patients' symptoms cannot be explained by delirium or other major psychiatric disorders. A thorough patient history provided by the patient and a knowledgeable family member along with bedside examination of the patient's mental health and neurophysiological tests must be completed to detect and confirm cognitive impairment. According to NINDCS-ADRDA, cognitive impairment is defined when patients show symptoms from at least two of the following domains: inability to remember new information, poor reasoning and inability to handle complicated tasks,

inability to understand visual representations and lack of spatial awareness, poorly functioning language, changes in personality and behavior. In addition to meeting the criteria for dementia, patients must also show a gradual onset and of these symptoms and a clear decline of cognition over time. Patients that also possess the causative genetic mutation for *amyloid precursor protein* (APP), *Presenilin-1* (PSEN1), or *Presenilin-2* (PSEN2) increases the certainty that the patient's symptoms is caused by AD pathology. Finally, despite all examinations that are completed to diagnose AD, the only definite method of diagnosis is brain autopsy after the patient's death.[14]

2.1.5 Neurophysiological Testing of Memory as an Endophenotype for AD

A patient's inability to remember family members or how to perform daily activities (ie: balancing their check book) are classic signs of AD. Normally, when a person learns new information, it is temporary held in short term memory until the hippocampus consolidates the information into long-term memory.[15] However, when the hippocampal region is damaged, the ability to convert from short-term memory to long-term memory is compromised,[16] and thus the patient lacks the ability to remember. An AD patient's short-term memory and long-term memory suffer and there are a number of tests to measure the function of both. In particular, to test a patient's general or long-term memory function clinicians use the Rey's Auditory Verbal Learning Test – Delayed Recall[17] (RAVLT-D) and the Logical Memory Test II – Delayed Recall[18] (LMT-D). Other tests can be used to measure a patient's ability to remain attentive, executive functioning, language, construction, and reasoning and judgment.

Since memory loss is a clear symptom of AD, declining scores of these memory tests, can serve as a strong endophenotype of AD.

RAVLT-D assesses a patient's verbal learning and memory. Though there are many variations of this test, patients listen to a list of fifteen words and repeat the words back to an administrator. They are then given an interference list of words, where thirty minutes later, patients are asked to recite back as many words they can from the initial list.[19] LMT-D asks patients to remember a story told to them by an administrator thirty minutes later. Administrators look for key words in the story and the final score is determined by the number of words that are mentioned.[20] These memory tests provide strong evidence of cognitive decline and thus become a critical endophenotype of AD.

2.1.6 Hippocampal Volume and Ventricular Volume as Endophenotypes

The hippocampal region in the brain plays a vital role in consolidating information from short-term memory to long-term memory and spatial navigation. The hippocampal region is one of the first regions to suffer damage in patients with AD, resulting in memory loss. Hippocampal atrophy could be detected in patients with AD and mild cognitive impairment (MCI) but not in cognitively normal individuals of the same age.[21] These rates of atrophy among AD patients range from 3% to 7% per year where cognitively normal individuals show a maximum atrophy rate of 0.9%.[22-24] It has also been indicated that delayed word recall tests are associated with hippocampal atrophy in AD patients as well as those with substantial neurofibrillary tangle

neuropathy.[25] Hippocampal volume, thus, is a crucial neurophysiological endophenotype of AD.

The ventricular system in the brain consists of numerous structures that hold cerebrospinal fluid, a colorless body fluid that protects the brain inside of the skull. The use of ventricular volume to determine the progress of AD has been supported by a number of recent studies and, in fact, brain atrophy rates are measured by ventricular expansion and thus an increase of cerebrospinal fluid space in the brain.[26] The rate of ventricular volume change is also highly correlated with the development of senile plaque and neurofibrillary tangles, two structures commonly found in the neuropathology of AD patients.[27] This evidence proves ventricular volume as a critical neurophysiological endophenotype of AD.

2.2 Genome-Wide Association Studies

A single nucleotide polymorphism (SNP) is a single base variation that occurs in a DNA sequence in over one percent of the human population. Since only three to five percent of the human genome codes for protein, the majority of SNPs are found in non-protein coding regions. Those SNPs that are within the protein coding regions are of particular interest because they may change the biological function of the protein. In humans, common diseases are not caused by a single variation within one gene and instead are a result of a number of genetic differences in multiple genes while also taking account for environmental factors and lifestyle choices. It is difficult to

determine environmental and lifestyle factors' impact on the disease process, but studying the genetic predisposition or likelihood of getting a disease is possible.

A genome wide association study (GWAS) aims to identify common genetic variants that are associated with a given trait or disease in a related or unrelated population. Often GWASs compare the genetic differences of patients with and without a given disease. Subjects provide a sample of DNA, typically extracted from their blood or saliva, from which millions of genetic variants are read using a SNP array. As opposed to a targeted approach, GWASs investigate the entire genome. Because the approach is non-candidate driven, GWASs determine the genes that are associated with a disease or trait but do not indicate which genes are causal.

2.2.1 GWASs of Brain Volume Atrophy and Cognitive Decline

Thus far there are two published GWASs that aim to find genetic association with hippocampal atrophy[28,29] and one published GWAS that aims to find association with ventricular volume[30]. The first study, a 2-stage GWAS for hippocampal atrophy, obtained data from two sources: Multi Institutional Research in Alzheimer's Genetic Epidemiology (MIRAGE) and the Alzheimer's Disease Neuroimaging Initiative (ADNI). Both datasets were analyzed individually and also in a meta-analysis approach that included African Americans and Caucasians ethnicities. According to the study's meta-analysis results, they identified four genes to be genome-wide significant (p-value $<1.0 \times 10^{-8}$).[28] The second GWAS for hippocampal atrophy identified 25 SNPs that mapped directly to 13 genes within the ADNI data set. Despite their findings, their analysis appeared to show some bias in the QQ plot that could have affected their outcome.[27]

Thus far, one published GWAS that aims to find association with ventricular volume and other regions in the brain including hippocampal volume. Interestingly, no significant genetic associations were found with ventricular volume and hippocampal volumes in a univariate framework.

Lastly, one published GWAS aimed to identify genetic variants associated with cognitive decline[31] by merging genetic data from two cohorts, the Religious Orders Study (ROS) and the Rush Memory and Aging Project (MAP). Their results identified APOE as genome-wide significant and associated with cognitive decline.

Chapter 3: Materials and Methods

3.1 Subjects

The Alzheimer's Disease Neuroimaging Initiative (ADNI) from 2004-2009 consisted of 822 unrelated patients recruited from 57 sites across the United States and Canada (adni.loni.ucla.edu). It is a multisite longitudinal study in an effort to support the research and discovery of the development of treatments to help hinder or halt the progression of AD. ADNI is funded by National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), pharmaceutical companies and non-profit organizations. Written informed consent was obtained from all participants or their families and a local institutional review board at each participating site approved the study. Of the 822 patients, there are 478 males and 344 females, 402 patients possess at least one of the APOE-e4 allele and they range from ages 55 to 91 at baseline.

ADNI participants were evaluated in six to twelve month intervals over a two to three year period, depending on the clinical diagnosis of the participant at baseline. Cognitive assessment was conducted at baseline for each participant. These neuropsychological tests measure immediate memory, verbal learning and memory, intelligence, attention and concentration, executive functions and language. In ADNI, these tests include the Alzheimer's Disease Assessment Scale – Cognitive[32], the American National Adult Reading Test[33], the Clock Drawing Test[34], the Logical Memory Test I - Immediate Recall[17], the Digit Span Forward[17], the Digit Span Backward[17], the

Category Fluency[35], the Trail Making Test[36-38], the Digit Symbol Substitution Test[17], the Boston Naming Test[39], the Logical Memory Test II - Delayed Recall (LMT-D)[17], and the Rey's Auditory Verbal Learning Test (RAVLT)[18].

In order for events to become permanent memories, these events must be consolidated by the hippocampus from short-term memory to long-term memory. However, in an AD brain, the hippocampus is unable to complete the task of converting these events into long-term memory. Since memory loss is a clear symptom of AD, declining scores of these memory tests can serve as a strong AD-related endophenotype. In particular, LMT-D and RVALT- Delayed Recall (RVALT-D) test the general memory function of a patient by asking a participant to recall a story or a list of words to the best of their ability after a period of 30 minutes. Though Logical Memory Test I – Immediate Recall also examines memory, this test focuses the immediate memory function of a patient. For the purposes of this study, we wished to study the general memory function and thus chose to incorporate the combined scores of RAVLT-D and LMT-D as endophenotypes.

Imaging and volumetric data is also available from ADNI. These data were collected by 1.5 Tesla (T) magnetic resonance imaging (MRI), 3.0 T MRI and positron emission tomography (PET) imaging methods at qualified data collection sites. MRI is a structural imaging method and is often considered the preferred neuroimaging method for AD.[40] MRI allows for precise measurements of three-dimensional volumes of hippocampal, ventricular and other related regions in the brain. Evidence of hippocampal atrophy, for example, can indicate the progression of AD. 3.0T MRI

possesses twice the magnetic field strength of 1.5T MRI and is said to provide a better noise-to-signal ratio to better differentiate gray matter from other tissue in the brain, though 1.5T field strength of MRI is still predominantly used today.[41] PET is a functional imaging method that enables clinicians and researchers to examine processes in the brain in a noninvasive manner.[42] For example, PET scans can help show a reduction of glucose levels in brain regions important in memory. Volumetric data was collected from all ADNI participants using 1.5T MRI and PET however only 25% of these participants were also screened with 3T MRI. Participants were screened approximately every six to twelve months, depending on the clinical diagnosis of the individual. We chose to limit our data collected using 1.5T MRI because of our interest in the rate of volume change of the hippocampal and ventricular regions of the brain in contrast to other biomarkers that can be studied using PET data. We chose to use the 1.5T MRI over the 3.0T MRI because more participants were screened with the lower magnetic field strength and therefore maximizing our sample size.

3.2 Clinical Data Cleaning to Determine Clinical Outcomes

When determining inclusion criteria for a GWAS analysis, sample size is an important factor to consider. In addition to sample size, however, creating a homogenous sample is equally as crucial. The majority of the patients were of Caucasian descent (n=763), thus any patients who were not Caucasian, primarily African American and Latino, were excluded (n=59). Any patients that experienced a stroke prior to or during the study and any patients who were diagnosed with other forms of

dementia or Parkinson's Disease were also excluded from the study (n=126). (Appendix 1, Appendix 3).

In order for the multivariate analysis to be successful, each of the three correlated quantitative clinical outcomes needed to have sufficient data for each patient: hippocampal volume, ventricular volume, and memory tests. The brain volumes were derived from MRI 1.5T imaging data only. From these data, we required that at least two data points for each patient be present for the averaged volume of the left and right hippocampal regions and for the averaged volume of the left and right ventricular regions. Participants also needed scores from the RAVLT-D and LMT-D cognitive memory tests at baseline in order to be included in this study. Any patients that lacked sufficient data in either memory test or volumetric data were excluded (n=75). Lastly, any patients with a genotypic call rate lower than 85% were removed from the study (n=2). After filtering based on these criteria, 567 participants remained in the study (Table 1).

	Control (*n*=158)	**MCI (*n*=277)**	**AD (*n*=132)**
Gender (M/F)	79/79	178/99	72/60
APOE (E4+/E4-)*	44/114	156/121	88/44
Age (mean)	62-90 (75.84)	55-89 (74.51)	55-91 (74.93)
* E4+ = patient has at least one APOE e4 allele; E4- = patient does not have an APOE e4 allele			

Table 1: Descriptive statistics of Final data set (n=567)

To normalize the phenotypic data we determined the rate of change for the average hippocampal volume and the average ventricular volume. These rates of change required data from two time points and followed Formula 1.

$$\frac{[(Volume\ at\ last\ time\ point - Volume\ at\ first\ time\ point)/Volume\ at\ first\ time\ point]}{(Last\ time\ point - First\ time\ point)}$$

(Formula 1)

The rates of change for both volumes were log-base 2 transformed to obtain a normal distribution. The two cognitive memory tests, RAVLT-D and LMT-D, were combined by calculating an average Z-score to improve reliability. As expected, these three quantitative phenotypic outcome measures were correlated (Table 2, Appendix 3). Hippocampal volume and ventricular volume were inversely correlated where the memory tests had nearly the same magnitude of correlation with both brain volumes but in opposite directions. Individual phenotype-covariate association analyses assessed which covariates should be incorporated in the model, however because of their clinical relevance, sex, age and APOE genotype were always included though the covariates may not have been significant for each phenotype (Appendix 2).

	Age	Sex	APOE Genotype	Cognitive Memory Tests	Hippocampal Volume	Ventricular Volume
Age	1	-0.050	-0.133	-0.043	-0.048	-0.305
Sex	-0.050	1	-0.025	0.005	-0.008	0.115
APOE Genotype	-0.133	-0.025	1	-0.108	-0.301	0.299
Cognitive Memory Tests	-0.043	0.005	-0.108	1	0.198	-0.183
Hippocampal Volume	-0.048	-0.008	-0.301	0.198	1	-0.453
Ventricular Volume	-0.305	0.115	0.299	-0.183	-0.453	1

Table 2: Pearson's Correlation of Phenotypes and Covariates
Correlation coefficients of the phenotypes (memory tests, hippocampal volume and ventricular volume) and covariates (age, sex and APOE Genotype)

3.3 Genotypic Data Cleaning to Determine Final Sample Size

Determining the quality of the genetic data is equally as important as studying the clinical data in a GWAS. ADNI used Illumina 610 Quad array with 620,901 SNP markers. SNP call rates less than 95%, a minor allele frequency less than 1%, and any mitochondrial SNPs or CNV markers reduced the number of SNP markers to a final genotypic data set of 543,715 SNPs. These genotypes were coded under an additive model as a function of the number of minor alleles (ie: 0, 1, or 2).

3.4 Statistical Analysis Methods

Traditionally GWAS is studied in a univariate framework where SNP markers predict the outcome of a given trait associated with a given disease. At the time of data collection for complex diseases, several correlated phenotypes are recorded but usually studied individually. By joining these correlated phenotypes, we are able to identify genetic loci that are associated with all of the AD-related phenotypes in the model.

Determining genetic association of these phenotypes together can provide opportunities to find pleiotropic genes that may have a more central role in functional pathways. In addition, more exact modeling may bring forth to more accurate predictions of one or more phenotypes within the model (Appendix 4).

The Principle Components Analysis (PCA) method and the Seemingly Unrelated Regression (SUR) method were chosen for this study and have been successfully implemented in other complex disease GWAS studies in the past.[43,44] The SUR method was used as a bivariate approach by implementing the cognitive memory tests and the rate of change of hippocampal volume as phenotypic outcomes. The PCA also included the rate of change of ventricular volume as the third outcome. Our significance threshold was p–value < 10^{-5} for SNPs of interest and however SNPs with strong association (p–value < 10^{-8}) were further scrutinized. Both methods required the use of R, version 2.15.0 (http://www.r-project.org). Table 3 highlights the important packages and functions used in this analysis.

Function	Package	Version	Use
systemfit()	Systemfit[45]	1.1-12	SUR
prcomp()	Basic	2.15.0	PCA
lm()	Basic	2.15.0	Linear models
qtscore()	GenABEL[46]	1.7-2	Genomic Control
GetClosestGeneInfo()	NCBI2R	1.4.4	Annotation

Table 3: Important R packages and functions

In addition, SNPs were annotated to the closest RefSeq gene (hg19) on the chromosome within 100kb. A brief summary of each method and an outline of our analysis can be found in the following sections.

3.4.1 Bivariate Model: Seemingly Unrelated Regression (SUR)

The SUR model is a system of linear regression models that allows for the possibility of different predictor variables. By incorporating different explanatory variables in the model to predict phenotypic traits, we are able to recognize that certain variables may only predict the outcome of one or some of the phenotypes. In addition, these linear regression models are "related" by their correlated error terms. A classic bivariate SUR system can be written as:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

(*Formula 2*)

In matrix notation the system can be written as $Y = X\beta + \varepsilon$, where Y is a vector of phenotypic variables, X is a diagonal matrix of explanatory variables, β is the vector of the coefficients and ε is a vector of the residual error terms.

We applied the SUR model to test the association of two correlated quantitative phenotypes: cognitive memory tests and the rate of change of hippocampal volume. Though SUR has the capability to incorporate more than two phenotypes in the model, biologically the rate of change in hippocampal volume and the scores cognitive memory tests are closely related. It is important to create a meaningful study design by choosing traits that are not only related to the disease but also are clinically and biologically

relevant to each other. The phenotypes are predicted by each SNP marker and the following covariates: age, sex and APOE genotype. In addition, the coefficients for all variables in the model remained unrestricted. There is no association to either one or both of the phenotypes under the null hypothesis. We were able to obtain the overall F-statistic by comparing the observed model to a null model where the SNP coefficients were zero.

3.4.2 Multivariate Model: Principle Components Analysis (PCA)

The PCA model transforms a set of correlated traits or phenotypes into an equal number of uncorrelated, or orthogonal, linear combinations called principle components (PCs). The number of PCs is equivalent to the number of phenotypes incorporated in the analysis. Each PC is determined by calculating the maximum variability possible in the data under the constraint that they be orthogonal with the other PCs (Figure 1).

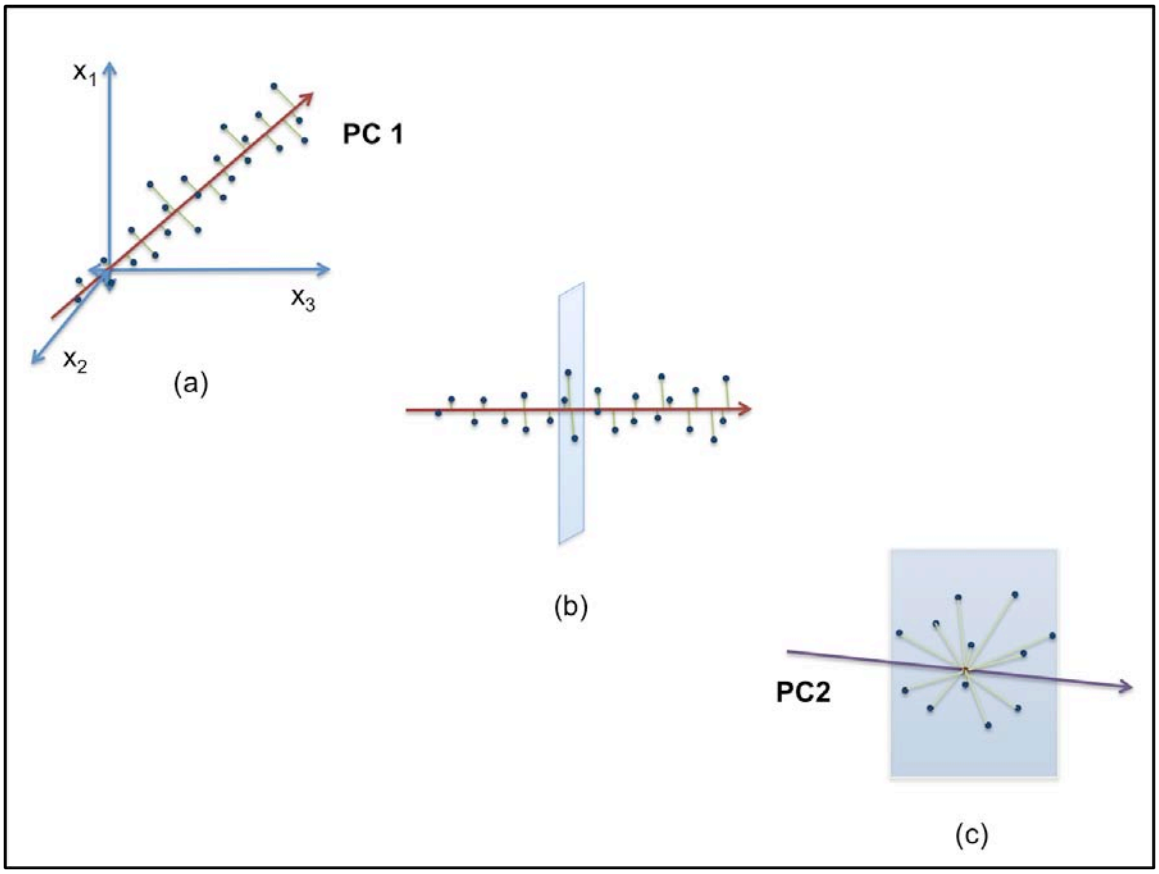


Figure 1: Graphical representation of how PCs are developed
(a) Determine the best fit line for all your data points and *(b)* rotate the axis to an orthogonal plane *(c)* and determine the best fit line based on this axis. PCA is an iterative process until 100% of the variation is explained. The number of PCs is equivalent to the number of phenotypes there are in the analysis. (Figure adapted from http://www.xlstat.com/en/learning-center/tutorials/principal-coordinate-analysis-pcoa-with-xlstat.html)

These PCs are then used in a traditional linear regression model where they are predicted by a SNP marker along with age, sex, and APOE genotype as covariates. Traditionally the first PCs that explain approximately 80% of the total variance are only taken into consideration with the belief that the remaining PCs, which explain little of the variance, are noise and irrelevant. In contrast, in 1998 Hadi & Ling showed that the

PCs which contribute little to the total variance may nonetheless significantly account for the variance in the response variable.[47]

By applying the PCA model to test the association of three correlated quantitative phenotypes, cognitive memory tests, the rate of change of hippocampal volume and the rate of change of ventricular volume, we obtained three PCs that all acted as individual response variables in a linear regression model. We obtained SNP *p*-values for each regression model.

Chapter 4: Results

4.1 Results from the SUR Method

Two phenotypes were incorporated in this analysis: memory tests and the rate of change of hippocampal volume. The bivariate SUR model identified eight significant SNPs with a *p*-value < 10^{-5}, with two SNPs having a *p*-value < 10^{-6} (Table 4, Figure 2.). A QQ Plot of the association results is presented in Figure 3 indicates no significant biases were present in the data.

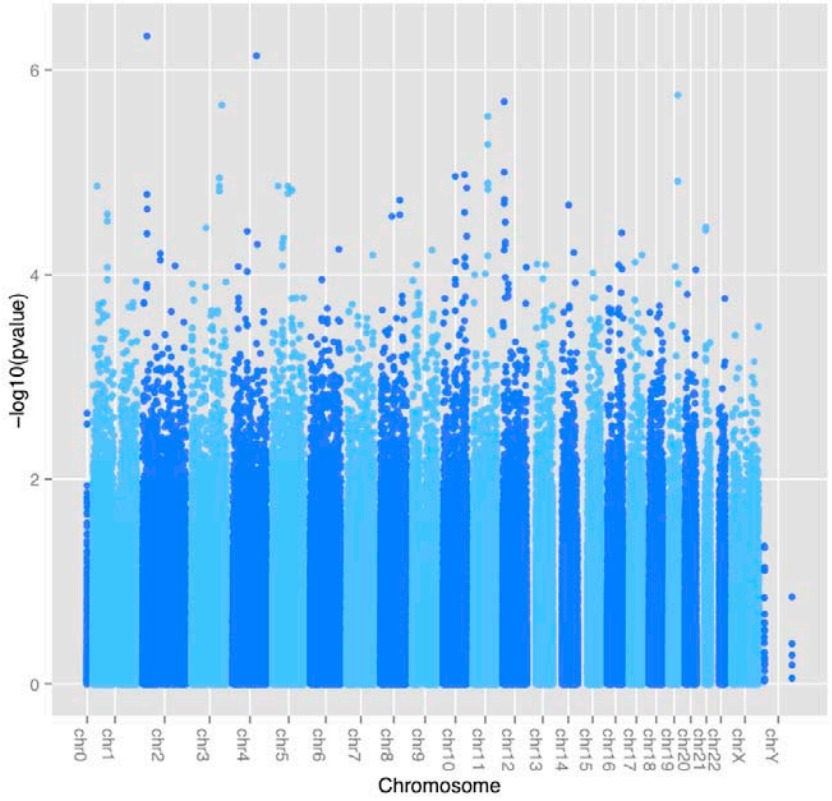


Figure 2: Manhattan Plot of the results from the SUR method
Manhattan plot of the $-\log_{10}$ of observed *p*-values. Plots of univariate results can be found in the Appendix 6.

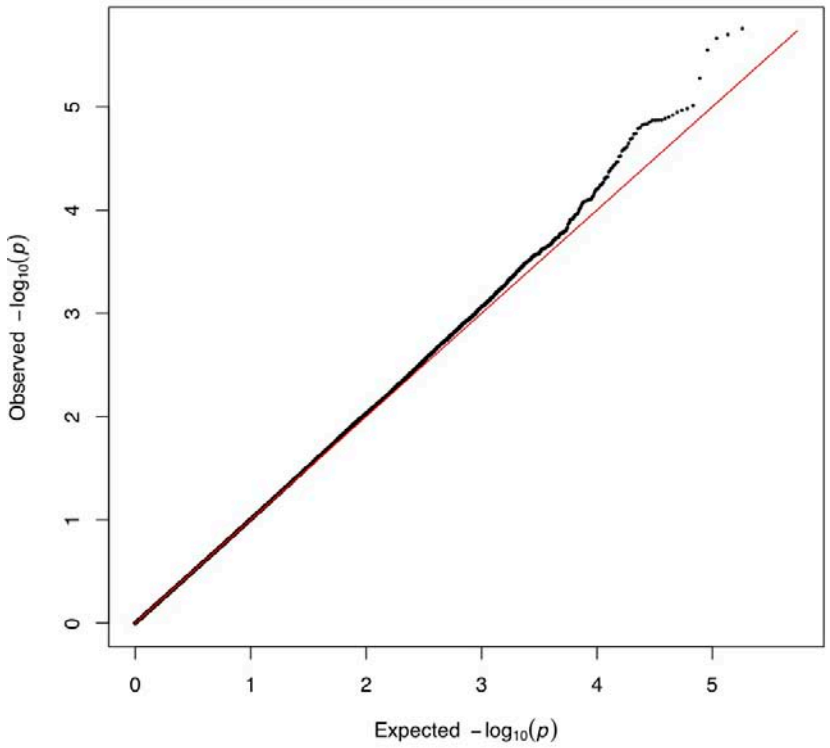


Figure 3: QQ Plot of the association results for the SUR method
QQ Plots of univariate results can be found in Appendix 6.

A *post-hoc* univariate analysis was completed for comparison with the bivariate SUR model. In the univariate approach, nine SNPs were found to be significantly associated with the rate of change of hippocampal volume and 11 SNPs were found to be significantly associated with memory tests (Appendix 6). The top four SNPs in the univariate analysis for rate of change of hippocampal volume and the top SNP the univariate analysis for the cognitive memory tests were also found to be highly significant in the bivariate SUR approach (Table 5-6). No SNPs were found to be genome-wide significant with a p-value < 10^{-8} by the SUR model or the univariate analyses.

P-value	(a) Bivariate SUR	(b) Univariate Models	
		Hippocampal Volume	Cognitive Memory Test
$10^{-7} < p < 10^{-6}$	2	1	2
$10^{-6} < p < 10^{-5}$	6	8	9
$10^{-5} < p < 10^{-4}$	70	39	55
TOTAL of SNPs $p < 10^{-5}$	**8**	**9**	**11**
TOTAL of SNPs $p < 10^{-4}$	**78**	**48**	**66**

Table 4: Number of significant SNPs determined by SUR
(a) shows the number of significant SNPs associated with hippocampal volume and memory tests as determined by the bivariate SUR method. (b) presents the number of significant SNPs associated with hippocampal volume and memory tests in a univariate framework

SNP Name	F-statistic	P-value	Closest Gene	SNP Location (bp)	Chromosome Location
rs1653725	29.174	4.62E-07	KLHL29	23806335	2p24.1
rs6848146	28.313	7.11E-07	TCONS_l2_0002 1296	132757887	4q28.3
rs788338	26.517	1.75E-06	MYH14	50778543	19q13.33
rs7294478	26.240	2.01E-06	C1RL	7266805	12p13.31
rs1511592	26.080	2.17E-06	EGFEM1P	168594416	3q26.2
rs10898028	25.552	2.83E-06	FAM181B	82456494	11q14.1
rs492923	24.302	5.28E-06	FAM181B	82478824	11q14.1
rs730165	23.065	9.80E-06	C1RL-AS1	7270804	12p13.31

Table 5: Top associated SNPs determined by the SUR model
List of the significant SNPs associated with hippocampal volume and memory tests as determined by the bivariate SUR method. Each SNP was mapped to its closest RefSeq gene (hg19) found within 100kb

(a) Cognitive Memory Test					
SNP Name	**Beta Coefficient**	**P-value**	**Closest Gene**	**SNP Location (bp)**	**Chromosome Location**
rs1653725	-0.291	2.35E-07	KLHL29	23806335	2p24.1
rs10898028	-0.267	6.94E-07	FAM181B	82456494	11q14.1
rs492923	-0.287	1.09E-06	FAM181B	82478824	11q14.1
rs7294478	-0.204	1.38E-06	C1RL-AS1	7266805	12p13.31
rs11233276	-0.233	2.84E-06	FAM181B	82411650	11q14.1
rs1021595	-0.234	2.84E-06	FAM181B	82413892	11q14.1
rs11583823	-0.403	2.85E-06	RUNX3	25299473	1p36
rs717178	-0.212	3.74E-06	LOC100499405	9395920	12p13.31
rs2036135	-0.225	3.75E-06	FAM181B	82448738	11q14.1
rs1432268	-0.244	4.12E-06	KLHL29	23770434	2p24.1
rs4752092	0.202	4.19E-06	TCONS_00018348	119448066	10q26.11
(b) Hippocampal Volume					
SNP Name	**Beta Coefficient**	**P-value**	**Closest Gene**	**SNP Location (bp)**	**Chromosome Location**
rs6848146	0.00091	1.57E-07	TCONS_l2_00021296	132757887	4q28.3
rs4533608	-0.003	2.41E-06	--	154398140	3q25.2
rs696854	-0.0014	3.14E-06	GPR149	154102701	3q25.2
rs6762590	-0.00145	3.20E-06	GPR149	154124532	3q25.2
rs171711	-0.0012	3.30E-06	MIR4280	86471345	5
rs30394	-0.0013	3.93E-06	RASA1	86490293	5q13.3
rs6882746	-0.0007	3.95E-06	--	31659869	5p13.3
rs11597160	-0.001	4.37E-06	TCONS_00018071	132273788	10q26.3
rs1159082	0.0008	8.66E-06	PPIAP22	20199970	21q21.1

Table 6: Top associated SNPs determined by univariate approach
List of the significant SNPs associated with (a) memory tests and (b) hippocampal volume as determined by the bivariate SUR method. Each SNP was mapped to its closest RefSeq gene (hg19) found within 100kb.

Multiple significant SNPs mapped to KLHL29, or *kelch-like 29 (Drosophila)*, and was significantly associated with the both phenotypes by the bivariate SUR approach (p-value < 4.62×10^{-7}) and in the univariate analysis for the memory tests (p-value < 2.35×10^{-7}). KLHL29, located on chromosome 2 (Figure 4, Appendix 8), has not yet been

reported to be associated with hippocampal volume, cognitive memory tests, ventricular volume or AD but is expressed in brain tissue.

A long non-coding RNA (lncRNA) on chromosome 4, TCONS_l2_00021296, was highly significantly associated with both phenotypes by the bivariate SUR (*p*-value < 7.11×10^{-7}) and was also most significant result in the univariate approach for hippocampal volume (*p*-value < 1.57×10^{-7}). lncRNAs are strongly implicated to be involved the regulation of protein-coding genes at both the transcriptional and post-transcriptional levels, and large number have been identified to affect a various cellular and developmental pathways. Thus, it would make sense that irregularities in lncRNAs could contribute to many complex diseases, including AD.[48]

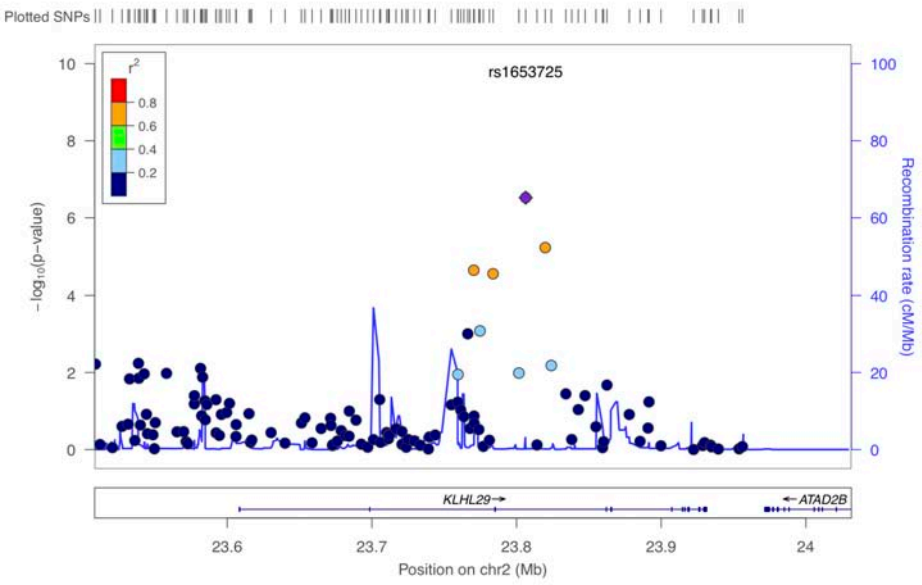


Figure 4: Focus Plot of KLHL29
Additional SNPs, but of lesser significance, are found within the KLHL29 region and are in linkage disequilibrium thus supporting our result. Region plots of all significantly determined SNPs are in Appendix 8.

MYH14, or *myosin, heavy chain 14, non-muscle*, is a gene that encodes for the myosin superfamily found on chromosome 19. Myosin-related genes are involved in actin-dependent motor proteins and regulate cytokinesis, cell polarity and cell molality. Mutations in MYH14 have been linked to a form of autosomal dominant hearing impairment and to the neurodegenerative disease, Charcot Marie Tooth disease, affecting the peripheral nervous system (Appendix 8).[49] This gene was determined highly significant by the bivariate SUR method (p-value < 1.75×10^{-6}) and was not found to be highly significant by either *post-hoc* univariate analyses.

C1RL, or *complement component 1, r subcomponent-like*, is found on chromosome 12 and encodes for a protein belonging to the serine protease family, but lacks any protease function. It still, however, plays an important role in the activation/catalytic process of these proteases. C1RL is expressed in a wide range of human tissue, including but not limited to liver, kidney, pancreas, placenta, lung, and spleen. Though the biological functions of the gene remain unknown, C1RL could play roles in the regulation of protease activity in inflammation or immune responses.[50] Furthermore, C1RL-AS1, *C1RL antisense RNA 1*, was also determined highly significant by the bivariate SUR method (C1RL: p-value < 2.01×10^{-6}, C1RL-AS1: p-value < 9.80×10^{-6}).

EGFEM1P, *EGF-like and EMI domain containing 1, pseudogene*, and FAM181B, family *with sequence similarity 181, member B*, were also determined highly significant by the SUR method (EGFEM1P: p-value < 2.17×10^{-6}, FAM181B: p-value < 2.83×10^{-6}, p-value < 5.28×10^{-6}). Little is known about these two genes however EGFEM1P is expressed in the hippocampus.

4.2 Results from PCA Method

All three phenotypes were incorporated in this analysis: cognitive memory tests, the rate of change of hippocampal volume and the rate of change of ventricular volume, and thus three PCs were studied. Though the first two PCs made up over 81.83% of the total variance, previous studies have shown that the PCs that account for little of the total variance are often those that explain the phenotypic variables. Thus all three PCs were incorporated in our analysis (Table 7). In addition, the PCs explain the joint variation of multiple phenotypes so the portion of the phenotypic variation that makes up each PC is unknown. To confirm the variation in the covariates in the PCs were not biased in one PC to the other, pairwise comparison plots of the three PCs highlighting the covariates in the model as well as the 57 data collection sites were created (Appendix 4). The plots did not show any evidence of a bias to any covariate or sites in the data.

PC	Percentage of Variance
PC1	52.72%
PC2	29.1%
PC3	18.17%
All PCs	100%

Table 7: Percentage of Variance explained by each PC
Shows the percentage of variance explained by each PC in the PCA analysis for hippocampal volume, ventricular volume and memory tests.

QQ plots of the association results for PC3 showed a deviation from normality and Genomic Control analysis was then performed to account for this bias (Appendix 7). After the Genomic Control, the PCA multivariate analysis method identified 21 significant SNPs with a *p*-value < 10^{-5} (Table 8, Appendix 7). Of these 21 SNPs, the

second PC (PC2) was associated with 13 of these SNPs, however none of these SNPs were found significant at the genome-wide significance threshold. Five SNPs were found significant in the first PC (PC1), which made up more than half of the total variation in this analysis. PC3, the PC that accounted for the least variation at 18.17%, found three significant SNPs, to which all mapped to unique genes. Table 9 shows remaining the associated SNPs from the PCA method by each PC. Not surprisingly, each PC uniquely identified a set of significant SNPs that mapped to different genes. We expect this result because each PC accounts for a distinctive part of the variation in within the data.

P-value	PC1	PC2	PC3
$10^{-7} < p < 10^{-6}$	0	4	0
$10^{-6} < p < 10^{-5}$	5	9	3
$10^{-5} < p < 10^{-4}$	45	49	35
TOTAL of SNPs $p < 10^{-5}$	**5**	**13**	**3**
TOTAL of SNPs $p < 10^{-4}$	**50**	**62**	**38**

Table 8: Number of significant SNPs determined by PCA
Shows the number of significant SNPs associated with PC1, PC2 and PC2 determined by the PCA method.

(a) PC1				
SNP Name	**P-value**	**Closest Gene**	**SNP Location (bp)**	**Chromosome Location**
rs6882746	2.27E-06	--	31659869	5p13.3
rs1253107	2.93E-06	JKAMP	59952947	14q23.1
rs1475394	4.77E-06	LPHN2	82412099	1p31.1
rs30394	5.00E-06	RASA1	86490293	5q13.3
rs10518661	5.16E-06	LPHN2	82381380	1p31.1
(b) PC2				
SNP Name	**P-value**	**Closest Gene**	**SNP Location (bp)**	**Chromosome Location**
rs1653725	3.01E-07	KLHL29	23806335	2p24.1
rs7294478	5.75E-07	C1RL-AS1	7266805	12p13.31
rs730165	6.96E-07	C1RL-AS1	7270804	12p13.31
rs10898028	9.32E-07	FAM181B	82456494	11q14.1
rs3782924	1.80E-06	C1RL-AS1	7262154	12p13.31
rs2263090	2.73E-06	SLFN11	33718077	17q12
rs9989026	3.02E-06	C1RL	7258451	12p13.31
rs492923	4.30E-06	FAM181B	82478824	11q14.1
rs788338	4.39E-06	MYH14	50778543	19q13.33
rs7498145	4.93E-06	ACSM2B	20547076	16p12.3
rs11681555	5.95E-06	KLHL29	23819830	2p24.1
rs788332	6.27E-06	MYH14	50782462	19q13.33
rs1291361	9.23E-06	HEBP1	13155166	12p13.1
(c) PC3				
SNP Name	**P-value**	**Closest Gene**	**SNP Location (bp)**	**Chromosome Location**
rs9684216	6.09E-06	TRIML2	189031028	4q35.2
rs696854	6.10E-06	GPR149	154102701	3q25.2
rs6835799	9.64E-06	SORCS2	7586924	4p16.1

Table 9: Top associated SNPs determined by the PCA model
List of the significant SNPs associated with (a) PC1 (b) PC2 and (c) PC3 as determined by the PCA method. Each SNP was mapped to its closest RefSeq gene (hg19) found within 100kb.

4.2.1 Mapped genes significantly associated to PC1

Four of the five SNPs determined by PC1 mapped to genes within 100kb of the SNP. JKAMP, *JNK1/MAPK8-associated membrane protein*, found on chromosome 14 (*p*-value < 2.93×10^{-6}) encodes a protein located within the endoplasmic reticulum and aids in the degradation of misfolded proteins by recruiting proteasomes and the components involved in endoplasmic-reticulum-associated protein degradation (ERAD).[51]

Abnormally functioning JKAMP could prevent ERAD from eradicating misfolded proteins and potentially lead to oxidative stress and cell death which has been suggested to be related to AD, Parkinson's Disease, and diabetes.[52]

LPHN2, *Latrophilin-2*, is a member of the latrophilin subfamily of G-protein coupled receptors and was associated to PC1 in our analysis (p-value < 5.16×10^{-6}). Latrophilins are suggested to function in cell adhesion and signal transduction. LPHN2 a candidate gene for its involvement in the development of breast cancer however it's specific role remains unidentified.[53]

RASA1, or *RAS p21 protein activator 1*, was the third gene found significantly associated with PC1 (p-value < 5.00×10^{-6}). The gene encodes for a protein that is responsible for inactivating the Ras protein, which is involved in cellular signal transduction. Mutations of the RASA1 gene have been linked to capillary malformation-arteriovenous malformation syndrome (CM-AVM) and Parkes Weber syndrome. CM-AVM is characterized by enlarged capillaries that increase blood flow close to the skin's surface and vascular abnormalities that affect blood circulation that can lead to abnormal bleeding, seizures, heart failure and even death. Parkes Weber syndrome presents similar vascular abnormalities to CM-AVM and also usually involves the over growth of one limb. Mutations in RASA1 have not yet been linked to AD.[54,55]

4.2.2 Mapped genes significantly associated to PC2

Significant genetic variants associated with PC2 mapped to eight unique genes. Of these eight genes, five genes (KLHL29, C1RL-AS1, FAM181B, C1RL and MYH14) were

also found to be associated with the joint association of hippocampal volume and the cognitive memory tests in the bivariate SUR model.

Found on chromosome 17, SLFN11, or *Schlafen-11*, (*p*-value < 2.73×10^{-6}) has been recently identified to selectively inhibit HIV protein expression. Though the specific mechanism is unknown, the natural inhibition of HIV protein synthesis could point to why some HIV-positive patients never experience the symptoms of AIDs.[56]

ACSM2B, *acyl-CoA synthetase medium-chain family member 2B*, (*p*-value < 4.93×10^{-6}) is one of five genes that encode for enzymes that catalyze the activation of medium chain length fatty acids. The gene is highly expressed in the liver and kidney and studies have linked this gene to traits of insulin resistance syndrome and type 2 diabetes.[57]

HEBP1, or *heme binding protein 1*, (*p*-value < 9.23×10^{-6}) may be involved in heme regulation or biosynthesis. It is also involved in the pathway to generate F2L, a crucial protein required for tissue repaired and regulation of the inflammatory process. HEBP1 is widely expressed in human tissue, including brain.[58]

4.2.3 Mapped Genes significantly associated to PC3

Three genetic variants associated with PC3 mapped to three unique genes: TRIML2 (*p*-value < 6.09×10^{-6}), GPR149 (*p*-value < 6.10×10^{-6}) and SORC2 (*p*-value < 9.64×10^{-6}). TRIML2, *tripartite motif family-like 2*, has yet to be associated with AD or the phenotypes discussed in this study. Little is also known about GPR149, *G protein-coupled receptor 149*, however the deletion of this gene increases fertility in mice.

SORCS2, or *sortilin-related VPS10 domain containing receptor 2*, is found on chromosome 4 and is highly expressed in human brain tissue (Appendix 8). As part of the related VPS10-domain receptor family, SORCS2 is one of five type 1 transmembrane proteins that interact with neurotrophins and neuropeptides. Lastly, family members of SORCS2, SORCS1 and SorLA, have been implicated to be associated with AD.

4.3 Common Findings of SUR and PCA

There were six highly significant SNPs found by both the SUR and the PCA methods. These SNPs mapped to four genes: KLHL29, FAM181B, C1RL-AS1, MYH14. Interestingly, all six SNPs were significantly associated with PC2 of the PCA method (Table 10).

Chromosome	Gene	P-values (PC2)	P-values (SUR)
2	**KLHL29**		
	rs1653725	3.01×10^{-7}	4.62×10^{-7}
11	**FAM181B**		
	rs10898028	9.31×10^{-7}	2.83×10^{-6}
	rs492923	4.30×10^{-6}	5.28×10^{-6}
12	**C1RL-AS1**		
	rs7294478	5.75×10^{-7}	2.01×10^{-6}
	rs730165	6.96×10^{-7}	9.80×10^{-6}
19	**MYH14**		
	rs788338	4.39×10^{-6}	1.75×10^{-6}

Table 10: Common genes found by both methods
Shows the four common genes, that mapped to six SNPs associated to PC2 in the PCA method and hippocampal volume and memory tests in the bivariate SUR method.

None of these genes have been previously implicated to be associated with AD, hippocampal atrophy, ventricular volume, or cognitive memory tests. These genes warrant follow-up studies to understand their involvement in relation to AD, or at least one or more of the phenotypes in the study.

Chapter 5: Conclusions

We employed two multivariate methods that allow for correlated quantitative phenotypes that are clinically relevant to AD. Both methods handle correlated phenotypes differently. With the SUR method, it is possible to have different linear regression models for each phenotype while retaining the correlation between the error terms.. The PCA method transforms the correlated phenotypes to uncorrelated, or orthogonal, linear combinations called PCs that serve as the response variable in a linear regression model. To our knowledge this is the first multivariate analysis study of an AD GWAS with unrelated subjects.

Our GWAS results show that carefully studying the relationship of quantitative endophenotypes with AD strongly promotes the identification of new genetic variants. The bivariate SUR model discovered eight SNPs that have not previously been found. These significant SNPs show a pleiotropic effect on the rate of change in hippocampal volume and cognitive memory tests, a phenomenon known to be ubiquitous in complex diseases like AD. In addition, the bivariate SUR model showed a strong association to SNPs that were the top hit in the *post-hoc* univariate analyses of both traits. The multivariate PCA model found 21 associated SNPs: five SNPs were associated with PC1, 13 SNPs were associated with PC2, and three SNPs were associated with PC3. As previously noted, traditionally the first PCs that account for 80% or more of the total variance are analyzed further with the notion that the remaining PCs are noise within the data, however others indicate that PCs that explain the little amount of the total

variance are often those that explain SNP association with the response variables. By including PC3 in our analysis, we determined three significant SNPs, each of which uniquely mapped to a gene. The SUR method determined eight unique genes and the PCA model identified 16 unique genes none of which have been previously identified to be associated with the hippocampal volume, ventricular volume, cognitive memory tests or AD.

As previously mentioned, two published GWASs studied genetic association with hippocampal volume, one published GWAS studied the genetic association with ventricular volume and one published GWAS studied the genetic associated with cognitive decline or memory. All of these studies were conducted using a univariate framework. Of these studies, the GWASs for ventricular volume[30] and cognitive decline[31] did not identify any novel genetic variants. The two GWASs determining association to hippocampal volume identified 17 genes together, none of which were identified in our study.[28,29] The first GWAS study for hippocampal volume identified novel four gene regions as a result of testing the association of imputed SNPs.[28] The second GWAS study identified novel 13 genes to be associated with hippocampal volume, however there was a significant bias in their association results, as depicted by their QQ Plots.[29] The study did not correct for the bias and warrants further scrutiny of their results.

In addition to these studies there have been a number of published GWASs, yet the results of our study remain unique. Determining genetic association in AD is complex because it accounts for multiple factors. Our study focuses on testing for the

association of two or three AD-related traits together, choosing to remove many other factors that impact the disease.

Examining phenotypic traits provide a more focused approach to studying complex disease like AD by avoiding any variability or noise in the traits for which we hope to find genetic association. A possible explanation as to why these results have not been otherwise identified could attest to our careful outcome selection and exclusion criteria to obtain a clean phenotypic data. Our results show that using multivariate analysis methods in an AD GWAS can bring out associations for correlated quantitative phenotypes and further characterize the nature of these complex diseases. Just as multiple traits help define a disease, these SNPs that are associated to multiple traits can provide insight to how these traits define a disease and warrant follow-up research. Our findings in this study need to be replicated in a similar cohort that possesses imaging, clinical and genotypic data.

5.1 Further work

The PCA method allows us to detect SNPs that are associated with the phenotypes we selected; however because of the unique joint nature of this analysis the relationship of the SNP to a given phenotype is unknown. Further research into the specific variation comprising each PC may improve our understanding of the relationships of the SNPs and the phenotypes.

AD impacts millions of Americans today. Our approach identified a number of novel genetic variants. By targeting phenotypic traits that are strongly associated with AD and analyzing them in two different multivariate methods, we were able to identify a

number of novel genetic variants. Uncovering genetic variants can further our understanding and provides new targets for treatment to the millions of people affected by AD in addition to advancing the fields of bioinformatics and AD research.

References

1. Alzheimer's Association. (2012). Alzheimer's Disease Facts and Figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association.*

2. Liu YZ, Pei YF, Liu JF, Yang F, Guo Y, Zhang L, Liu XG, Yan H, Wang L, Zhang YP, Levy S, Recker RR, Deng HW. (2009). Powerful bivariate genome-wide association analyses suggest the *SOX6* gene influencing both obesity and osteoporosis phenotypes in males. *PLoS ONE, 4*(8), e6827. doi:10.1371/journal.pone.0006827

3. Tanzi, RE. (2012). The Genetics of Alzheimer Disease. *Cold Spring Harb Perspect Med*. 2:a006296.

4. Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, Pahwa JS, Moskvina V, Dowzell K, Williams A, et al. (2009). Genome-wide association study identifies variants at *CLU* and *PICALM* associated with alzheimer's disease. *Nature Genetics, 41*, 1088-1093. doi:10.1038/ng.440

5. Biller, José. *Practical Neurology*. Philadelphia: Lippincott Williams and Wilkins. 2009. Print.

6. Ghiso J & Frangione, B. (2002). Amyloidosis and Alzheimer's Disease. *Advanced Drug Delivery Reviews*. 54(12): 1539-1551. doi: 10.1016/S0169-409X(02)00149-7

7. Perl D. (2010). Neuropathology of Alzheimer's Disease. *Mt Sinai J Med*. 77(1): 32–42. doi:10.1002/msj.20157

8. "Alzheimer's Disease: Risk Factors." *Mayo Clinic*. 2011.

9. Kim J, Basak J, Holzman D. (2009). The Role of Apolopoprotein E in Alzheimer's Disease. Neuron. 63(3): 287-303 doi: 10.1016/j.neuron.2009.06.026

10. Guereiro R, Hardy, J. (2011). Alzheimer's disease genetics: Lessons to improve disease modeling. *Biochemical Society Transactions, 39*, 910-916. doi:10.1042/BST0390910

11. Bertram, L. (2011). Alzheimer's genetics in the GWAS era: A continuing story of 'Replications and refutations'. *Current Neurology and Neuroscience Reports, 11*, 246-253. doi:10.1007/s11910-011-0193-z

12. Jack CR Jr, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, Theis B, Phelps CH. (2011). Introduction to the recommendations from the National

Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*. 7:257-262.

13. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*. 7:263-269.

14. Harris M, Ivnik R, Smith, G. (2010). Mayo's Older Americans Normative Studies: Expanded AVLT Recognition Trial Norms for Ages 57 to 98, Journal of Clinical and Experimental Neuropsychology, 24:2, 214-220

15. Hartley T, Bird CM, Chan D, Cipolotti L, Husain M, Vargha-Khadem F, Burgess N. (2007). The hippocampus is required for short-term topographical memory in humans. *Hippocampus*. 17(1): 34–48.

16. Graham KS, Hodges JR. (1997). Differentiating the Roles of the Hippocampal Complex and the Neocortex in Long-Term Memory Storage: Evidence From the Study of Semantic Dementia and Alzheimer's Disease. Neuropsychology. 11(1): 77-89.

17. Wechsler D. (1987). Wechsler Memory Scale--revised manual. Psychological Corp., San Antonio.

18. Rey A. (1964). L'examen clinique en psychologie. Presses Universitaires de France, Paris.

19. Spreen, Otfried, & Esther Strauss. A Compendium of Neurophsychological Tests: Administration, Norms and Commentary. 1st ed. New York: Oxford University Press, 1991. Print.

20. Schuff N, Woerner N, Boreta L, Kornfield T, Shaw LM, Trojanowski JQ, Thompson PM, Jack CR Jr, Weiner MW. (2009). MRI of hippocampal volume loss in early Alzheimer's disease in relation to Apoe genotype and biomarkers. *Brain* 132, 1067–1077

21. Jack CR Jr, Petersen RC, Xu Y, O'Brien PC, Smith GE, Ivnik RJ, Tangalos EG, Kokmen E. (1998). Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. *Neurology*. 51:993–9.

22. Laakso MP, Lehtovirta M, Partanen K, Riekkinen PJ, Soininen H. (2000). Hippocampus in Alzheimer's disease: a 3-year follow-up MRI study. *Biol Psychiatry*. 47:557–61.

23. Raz N, Rodrigue KM, Head D, Kennedy KM, Acker JD. (2004). Differential aging of the medial temporal lobe: a study of a five-year change. *Neurology*. 62:433–8.

24. Mortimer JA, Gosche KM, Riley KP, Markesbery WR, Snowdon DA. Delayed recall, hippocampal volume and Alzheimer neuropathology: Findings from the Nun Study. (2004). *Neurology.* 62:428. doi: 10.1212/01.WNL.0000106463.66966.65

25. Nestor SM, Rupsingh R, Borrie M, Smith M, Accomazzi V, Wells JL, Fogarty J, Bartha R. (2008). Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain*. 131: 2443-2454. doi:10.1093/brain/awn146

26. Silbert LC, Quinn JF, Moore MM, Corbridge E, Ball MJ, Murdoch G, Sextpm G. Kaye JA. (2003). Changes in premorbid brain volume predict Alzheimer's disease pathology. *Neurology*. 61: 487–92.

27. Erten-Lyons D, Dodge H, Woltjer R, Silbert, L, Howieson D, Kramer P, and Kaye J. (2012). Neuropathological Basis of Age-Associated Brain Atrophy. *Archives of Neurology*. In press.

28. Melville SA, Buros J, Parrado AR, Vardarajan B, Logue MW, Shen L, Risacher SL, Alzheimer's Disease Neuroimaging Initiative, Kim S, Jun G, DeCarli C, Lunetta KL, Baldwin CT, Saykin AJ, Farrer LA. (2012). Multiple Loci Influencing Hippocampal Degeneration Identified by Genome Scan. *Ann Neurol*. 72(1):65-75.

29. Potkin SG, Guffanti G, Lakatos A, Turner JA, Kruggel F, Fallon JH, Saykin AJ, Orro A, Lupoli S, Salvi E, Weiner M, Macciardi F; Alzheimer's Disease Neuroimaging Initiative. (2009). Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One*. 4(8):e6501.

30. Furney SJ, Simmons A, Breen G, Pedroso I, Lunnon K, Proitsi P, Hodges A, Powell J, Wahlund LO, Kloszewska I, Mecocci P, Soininen H, Tsolaki M, Vellas B, Spenger C, Lathrop M, Shen L, Kim S, Saykin AJ, Weiner MW, Lovestone S; Alzheimer's Disease Neuroimaging Initiative; AddNeuroMed Consortium. (2011). Genome-wide association with MRI atrophy measures as a quantitative trait locus for Alzheimer's disease. *Mol Psychiatry*. 16(11):1130-8.

31. De Jager PL, Shulman JM, Chibnik LB, Keenan BT, Raj T, Wilson RS, Yu L, Leurgans SE, Tran D, Aubin C, Anderson CD, Biffi A, Corneveaux JJ, Huentelman MJ; Alzheimer's Disease Neuroimaging Initiative, Rosand J, Daly MJ, Myers AJ, Reiman EM, Bennett DA, Evans DA. (2012). A genome-wide scan for common variants affecting the rate of age-related cognitive decline. *Neurobiol Aging*. 33(5):1017.e1-15.

32. Rosen, WG, Mohs RC, Davis KL. (1984). A new rating scale for Alzheimer's disease. *Am J Psychiatry*. 141(11), 1356-64

33. Grober E, Sliwinski M, (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *J Clin Exp Neuropsychol.* 13(6), 933-49.

34. Goodglass H, Kaplan E. (1983). The assessment of aphasia and related disorders. Lea and Febiger, Philadelphia.

35. Morris JC, Heyman A, Mohs RC, Hughes JP, van Belle G, Fillenbaum G, Mellits ED, Clark C. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* 39(9), 1159-65.

36. Partington JE, Leiter RG. (1949). Partington's Pathway Test. *The Psychological Service Center Bulletin*. 1, 9-20.

37. Reitan R, Wolfson D. (1985). The Halstead-Reitan Neuropsychological Test Battery. Neuropsychology Press, Tucson.

38. Reitan RM, (1958). Validity of the Trail-Making Test. *Perceptual Motor Skills* 8, 271-6.

39. Kaplan E, Goodglass H, Weintraub S. (1983). The Boston Naming Test. Lea and Febiger, Philadelphia.

40. Doody RS, Stevens JC, Beck C, Dubinsky RM, Kaye JA, Gwyther L, Mohs RC, Thal LJ, Whitehouse PJ, DeKosky ST, Cummings JL. (2001). Practice parameter: management of dementia (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*. 56(9):1154-66.

41. Ho AJ, Hua X, Lee S, Leow AD, Yanovsky I, Gutman B, Dinov ID, Leporé N, Stein JL, Toga AW, Jack CR Jr, Bernstein MA, Reiman EM, Harvey DJ, Kornak J, Schuff N, Alexander GE, Weiner MW, Thompson PM; Alzheimer's Disease Neuroimaging Initiative. (2010). Comparing 3 T and 1.5 T MRI for tracking Alzheimer's disease progression with tensor-based morphometry. *Hum Brain Mapp*. 31(4):499-514.

42. Silverman DH, Small GW, Chang CY, Lu CS, Kung De Aburto MA, Chen W, Czernin J, Rapoport SI, Pietrini P, Alexander GE, Schapiro MB, Jagust WJ, Hoffman JM, Welsh-Bohmer KA, Alavi A, Clark CM, Salmon E, de Leon MJ, Mielke R, Cummings JL, Kowell AP, Gambhir SS, Hoh CK, Phelps ME. (2001). Positron emission tomography in evaluation of dementia: Regional brain metabolism and long-term outcome. *J Am Med Assoc*. 286(17):2120-7.

43. Saint-Pierre A, Kaufman JM, Ostertag A, Cohen-Solal M, Boland A, Toye K, Zelenika D, Lathrop M, de Vernejoul MC, Martinez M. (2011). Bivariate association analysis in selected samples: application to a GWAS of two bone mineral density phenotypes in males with high or low BMD. *Eur J Hum Genet*. 19(6):710-716

44. Bolormaa S, Pryce JE, Hayes BJ, Goddard ME. (2010). Multivariate analysis of genome-wide association study in dairy cattle. *J Dairy Sci*. 93(8): 3818-3833.

45. Henningsen A, Hamann JD. (2007). systemfit: A Package for Estimating Systems of Simultaneous Equations in R. *Journal of Statistical Software.* 23(4), 1-40. URL http://www.jstatsoft.org/v23/i04/

46. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics*. 23(10): 1294-6.

47. Hadi AS, Ling RF. (1998). Some cautionary notes on the use of principle components regression. *The American Statistician*. 52(1): 15-19.

48. Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS. (2010). Non-coding RNAs: regulators of disease. *J Pathol*. 220: 126-129.

49. Choi BO, Kang SH, Hyun YS, Kanwal S, Park SW, Koo H, Kim SB, Choi YC, Yoo JH, Kim JW, Park KD, Choi KG, Kim SJ, Züchner S, Chung KW. (2011). A complex phenotype of peripheral neuropathy, myopathy, hoarseness, and hearing loss is linked to an autosomal dominant mutation in MYH14. *Hum Mutat*. 32(6):669-77.

50. Lin N, Liu S, Li N, Wu P, An H, Yu Y, Wan T, Cau X. (2004). A novel human dendritic cell-derived C1r-like serine protease analog inhibits complement-mediated cytotoxcity. *Biochem Bioph Res Co*. 321: 329-336.

51. Tcherpakov M, Broday L, Delaunay A, Kadoya T, Khurana A, Erdjument-Bromage H, Tempst P, Qui X, DeMartino GN, Ronai Z. (2008). JAMP Optimizes ERAD to Protect Cells from Unfolded Proteins. *Mol Bio Cell*. 19: 5019-5028.

52. Haynes CM, Titus EA, Cooper AA. (2004). Degradation of Misfolded Proteins Prevents ER-Derived Oxidative Stress and Cell Death. *Mol Cell*. 15: 767-776.

53. White GR, Varley JM, Heighway J. (2000). Genomic structure and expression profile of LPHH1, a 7TM gene variably expressed in breast cancer cell lines. *Biochim Biophys Acta*. 1491: 75-92.

54. Boon LM, Mulliken JB, Vikkula M. (2005). RASA1: variable phenotype with capillary and arteriovenous malformations. *Curr Opin Genet Dev*. 15: 265-269

55. Eerola I, Boon LM, Mulliken JB, Burrows PE, Dompmartin A, Watanabe S, Vanwijck R, Vikkula M. (2003). Capillary malformation-arteriovenous, a new clinical genetic disorder caused by RASA1 mutations. *Am. J. Hum. Genet*. 73:1240–1249.

56. Li M, Kao E, Gau X, Sandig H, Limmer K, Pavon-Eternod M, Jones TE, Landry S,Pan T, Weitzman MD, David M. (2012). Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. *Nature*. 491(7422):125-8.

57. Boomgaarden I, Vock C, Klapper M, Döring F. (2009). Comparative analyses of disease risk genes belonging to the acyl-CoA synthetase medium-chain (ACSM) family in human liver and cell lines. *Biochem Genet*. 47(9-10):739.

58. Devosse T, Dutoit R, Migeotte I, De Nadai P, Imbault V, Communi D, Salmon I, Parmentier M. (2011). Processing of HEBP1 by cathepsin D gives rise to F2L, the agonist of formyl peptide receptor 3. *J Immunol*. 187(3):1475-85.

Appendix

Appendix 1: Inventory of ADNI Clinical Files

All data files described here can be retrieved with special permission from ADNI administrators (adni.loni.ucla.edu).

Appendix 1 is designed to provide an overview of the clinical data present in the ADNI data set. I have organized the information by file name where I have provided the information from the initial ADNI protocol to obtain the data (when applicable and/or available), the background and detailed information of the data itself.

adni_adas_2009-09-01.csv

Alzheimer's Disease Assessment Scale (ADAS) - Cognitive

- Brief cognitive test to assess learning and memory, language production, language comprehension, constructional praxis, ideational praxis and orientation
- This test is not timed.

adni_adasscores_2009_09_01.csv

- Scores from adni_adas_2009-09-01.csv
  - Q1: Word Recall → Score = 10 – (average of 3 trials)
  - Q2: Commands → Score = 5 – (# of commands correctly completed)
  - Q3: Construction → Score = 4 – (# of constructions correctly drawn)
    - Score of 5 means none were correct
  - Q4: Delayed Word Recall →Score = 10 – (# remembered)
  - Q5: Naming → Score = 17 – (# correctly named)
    - Score of 18 = none
  - Q6: Ideational Praxis → Score = 5 – (#correctly completed)
    - Score of 6 = none
  - Q7: Orientation → Score = 8 (#correctly)
    - Score of 9 = none
  - Q8: Word Recognition → Score = 12 – (#correct)
    - Score of 12 = none
  - Q9: Recall Instructions → Score = 0 if no help, 1 if need help
  - Q10: Spoken Language → Score = 0 if no help, 1 if need help
  - Q11: Word Finding → Score = 0 if none, 1 impaired
  - Q12: Comprehension → Score = 0 if none, 1 impaired
  - Q13: No data
  - Q14: Number Cancellation
- TOTAL11: all Qs except for Q4 and Q14
- TOTALMOD: all Qs

adni_addcomm_2009-09-01.csv

- Additional comments

adni_adsxlist_2009-09-01.csv

- List of symptoms/diagnosis
- Nausea, vomiting, diarrhea, constipation, abdominal discomfort, sweating, dizziness, low energy, drowsiness, blurred vision, headache, dry mouth, shortness of breath, coughing, palpitations, chest pains, urinary discomfort, urinary frequency, ankle swelling, musculoskeletal pain, rash, insomnia, depressed mood, crying, elevated mood, wandering, fall.
    - 1 = present, 0 = absent

adni_apoeres_2009-09-01.csv

- Results of APOE genotyping
- Genotypes of both alleles
- Info of how blood was handled

adni_arm_2009-09-01.csv

- Gives diagnosis (NL, MCI, or AD) and scan assignment
- (1.5 Tesla MRI only PET & 1.5T MRI, 1.5T MRI & 3T MRI)
    - Diagnosis
        - NL : 1 & 4 & 7
        - MCI: 2 & 5 & 8
        - AD: 3 & 6 & 9
    - Scan assignment
        - 1.5T MRI only: 1 & 2 & 3
        - PET & 1.5MRI: 4 & 5 & 6
        - 1.5T MRI & 3T: 7 & 8 & 9

adni_biomark_2009-090-01.csv

- Information about biomarker samples
- 4 biomarker samples: blood (serum), plasma, urine, cerebrospinal fluid (CSF)
    - Each biomarker has info for time of collection, amount collected, centrifuged time, transfer time, volume transferred, time frozen

adni-blchange_2009-09-01.csv

- Overall summary to explain if there was a “clinically relevant” change (in relation to baseline) overtime (ie: changes in MMSE. ADAS, etc.)

adni_blscheck_2009-09-01.csv

- checklist of symptoms at baseline
- nausea, vomiting, diarrhea, constipation, abdominal discomfort, sweating, dizziness, low energy, drowsiness, blurred vision, headache, dry mouth, shortness of breath, coughing, palpitations, chest pains, urinary discomfort, urinary frequency, ankle swelling, musculoskeletal pain, rash, insomnia, depressed mood, crying, elevated mood, wandering, fall

adni_cdr_2009-09-01.csv

- Clinical Dementia Rating (CDR): numeric scale to quantify severity of symptoms of dementia
- 6 Categories: Memory, Orientation, Judgment, Problem Solving, Community Affairs, Home & Hobbies, Personal Care
    - Memory is primary, remaining categories are secondary
- CDR Score = Memory (M) if 3 secondary categories are given same scores as memory.

CDR Composite Rating	Symptoms
0	None
0.5	Very mild
1	Mild
2	Moderate
3	Severe

- conducted at every clinic visit after month 6 visit

adni_faq_2009-09-01.csv

- measures activities of daily living(can patient do it w/o assistance
    - ie: writing checks, shopping, playing bridge/chess etc
- administered at baseline & every subsequent in clinic visit

adni_fhq_2009-09-01.csv

- provide info of whether mother/father suffered from dementia and/or AD

adni_gdscale_2009-09-01.csv

- ADNI uses shorter version of GDS. Originally GDS is a 30 item questionnaire used to identify depression in the elderly.

- Completed at screen, 12 months, 24 months, 36months
- A score of > 5 indicates further examination for depression is required (for shorter version only)(SEE TOTALSCORE COLUMN)

adni_hcres_2009-09-01csv

- Evidence from observational epidemiological studies suggest that higher levels of plasma total homocysteine may be associated with inc. risk of AD.(and also stroke & Parkinson's Disease)
- Ranges: Female 4.9 – 11.6 µmol/L, elevated > 10.4 µmol/L
  Male : 5.9 – 15.3 µmol/L, elevated > 11.4 µmol/L
- This file gives amount of total plasma homocysteine in µmol/L.

adni_inclusio_2009-0901.csv

- A list of questions, where the answers determines patient's inclusion in the study
- Completed at screening phase.

adni_indfemog_2009-09-01.csv

- Provides demographic info of patient's study partner (gender, occupation, relationship to patient)

adni_labtests_2009-09-01.csv

- Tests completed at c=screening
- Blood & urine

adni_loclab_2009-09-01.csv

- From CSF, WBC count, RBC count, protein results(mg/dL), glucose results(mg/dL)

adni_medhist_2009-09-01.csv

- A yes/no response if the patient has a clinically significant history of problems in the area
- Completed at screening.

adni_mmse_2009-09-01.csv

- Test used to screen for cognitive impairment
- Scores range from 0 – 30
  - ≥ 25 is considered normal

    - < 10 is considered severe impairment
    - 10 -19 is considered moderate impairment
    - 19-24 is considered mild impairment /AD
- MMSCORE = Total score in the data file

adni_modhach_2009-09-01.csv

- Modified Hachinski: test used to screen/differentiate vascular dementia from degenerative forms.
- ≥7 = vascular dementia
- HMSCORE = total score in the data file.

adni_mri3meta_2009-09-01.csv

- 3T MRI Scan Information
- Follows the ADNI MRI Tech Procedure Manual (pages 21 – 27)
- For each step in manual a yes/no record was kept to indicate what was completed/not completed for the given patient

adni_mrib1calib_2009-09-01.csv

- MRI B1 Calibration
- Head coil or body

adni_mrimeta_2009-09-01.csv

- 1.5T MRI Scan Information
- Follows the ADNI MRI Tech Procedure Manual (pages 21 – 27)
- For each step in manual a yes/no record was kept to indicate what was completed/not completed for the given patient

adni_mrimpro_2009-09-01.csv

- MPRAGE Process (T1 weighted, v common)
- Provides information regarding the scanning process
- More necessary info is in adni_mrimprank

adni_mrimprank_2009-09-01.csv

- **MPRAGE RANKING**
    - Ranks the quality of the scan
- Also have the monthly visits

adni_mrinclusio_2009-09-01.csv

- MRI Subject inclusion→ can patient be in study?
- Information regarding scan type (1.5T/3T) and any medical image issues
    - Surgery, infarction, hemorrhage, trauma, devanomaly, metallic, lesion, nph, atrophy, edema
- Indicates if patient "passes" and if they can continue in study (based on MRI scan alone)

adni_mriphantom_2009-09-01.csv

- MRI phantom (QC)
    - Tests performance of MRI system
    - PRESENT data field indicates if system was present & accurate or not

adni_mriprot_2009-09-01.csv

- MRI Protocol
    - Appears to test quality of images
    - Appears to provide info for what technology was used & if scan passed.

adni_mriquality_2009-09-01.csv

- MRI Quality

adni_mriread_2009-09-01.csv

- MRI Clinical Read:
- Indicates if the patient was screened with 1.5T/3T/ or both
- Indicated if patient is compatible with inclusion/exclusion criteria
- Indicated if patient is clinically suitable to remain in study.

adni_neurobat_2009-09-01.csv

- Neuropsychological Battery
    - Assessment of possible physical aspects of neurological damage
- Clock Drawing Test (pg. 88 – 89)
    - Recorded yes /no results based on drawings.
- Logical Memory Test 1 – immediate recall (pg.91)
    - Patient must recall immediately (verbatim) a story that was just read to them.
- Digit Span Forward (pg.92)

- Used to test working memory by reading # sequence & asking patient to repeat
- Total score is # patient got correct
- Forward length is the length of highest digit sequence patient was able to repeat back.

- Digit Span Backward (pg.93-94)
  - Used to test working memory by reading # sequences to the patients & asking to repeat the backwards.
  - Total score(DSPANBAC) is total # patient got correct
  - Backward (DSPANLTH) is the length of highest digit sequence patient was able to say in reverse.
- Category Fluency Test (pg. 95-96)
  - Used to measure semantic memory (verbal fluency, language)
  - Patient asked to give examples of a given topic in an allotted amount of time (60 seconds)
  - The adequate number of responses is dependent on age

Age	#Response(adequate)
<65-69	15
70-74	15
75-79	14
80-84	13
85>	11

  - 1st Topic: Animals (CARANIMSC)
  - 2nd Topic: Vegetables (CATVEGESC)

- Trail Making Test (pg. 97 – 98)
  - Test of processing speed & executive function
  - Part A
    - Ask patient to connect dot #1 – 25 in order in 150second
    - Scoring based on the time it takes. (TRAASCOR)
  - Part B
    - Ask patient to connect dots #1 – 13 & A - L alternating in order in 300 seconds
    - Scoring based on the time it takes. (TRABSCOR)
  - For A & B errors
    - Errors of commission: connects dots in the incorrect sequence each occurrence is marked
    - Error of omission: fails to draw connecting line between 2 dots in the correct sequence. Occurs when patient runs out of time
  - Scoring Thresholds

	Avg. Score	Deficient	Rule of Thumb
Part A	29sec	>78sec	Most in 90 sec
Part B	75sec	>273sec	Most in 3 min

- Digit Symbol Substitution Test (pg.99)
  - Test sensitive to brain damage, dementia, age, depression
  - 90 second timed test
  - Total # points max is 93
- Logical Memory Test II- Delayed recall (pg.100)
  - Scoring same as logical MT I.
- Boston Naming Test (pg. 100 – bottom)
  - Tests the ability to name line drawings of objects
  - Only odd # problems were used in ADNI testing
  - Max score = 30 (in ADNI)
  - If semantic or phonemic, clues were given, it is recorded.
- American National Adult Reading Test – ANART (pg. 87)
  - Tests ability to correctly read & pronounce words
  - Score # correct, # incorrect

adni_neuroexm_2009-09-01.csv

- Neurological Exam
- Completed at screening
- Abs/pres: visual impairment, auditory impairment, tremor
- Norm/Abnorm: Level of consciousness, cranial nerves, motor strength, cerebellar- fingers to nose, cerebellar – heel to shin, sensory, deep tendon reflexes, plantar reflexes, gait.

adni_npiq_2009-09-01.csv

- Neuropsychiatric Inventory Q
  - Yes/no or none/mild/moderate/severe assessment of patient based on their responses to questions.ie: Does patient believe others are stealing from them? Does patient act as if they hear voices?

adni_pdxconv_2009-09-01.csv

- Diagnostic Summary
- Completed Bl, M6, M12, M24, M36
- 1. indicates current diagnosis. (NL, MCI, AD)
- 2. Indicates a conversion (NL→MCI, NL→AD, MCI→AD) or a reversion (MCI→NL, AD→NL, AD→MCI)
- 3. Describes the MCI, AD, or other dementia or Parkinson's

adni_petmeta_2009-09-01.csv

- PET Scan Information
- Prior to PET scan plasma (10 vials) were given to each patient.
- File provides info of plasma, as well as details of the scanning process.

adni_petqc_2009-09-01.csv

- PET QC Tracking
- Indicates if PET scans are acceptable & which frames, if any are unacceptable.
- Whether scans pass QC.

adni_physical_2009-09-01.csv (pg 108)

- Physical Exam
- Completed at screening
- normal/abnormal responses assessed by clinician

adni_pibmeta_2009-09-01.csv

- PiB Scan information
- MCI n = 48, AD n = 24, NL n = 24
- PET scans using Pittsburg Compound B (PiB) (used to image B-amyloid plaques in neuronal tissue)
- Indicates scanner info, PiB dose, motion issues

adni_pibqc_2009-09-01.csv

- PiB QC Tracking
- Indicates if PET scans W/ PiB are acceptable or not & if scans pass QC

adni_ptdemop_2009-09-01.csv

- Participant Demographic Information
- Completed at screening
- Gender, DOB, handedness, marital status, education, mental retardation, primary occupation, most recent occupation, retired date, type of residence, language for testing, primary language, year of onset for AD symptoms, ethnicity, race

adni_recadv_2009-09-01.csv

- Adverse Events/Hospitalizations Log
- Record of all new symptoms & all symptoms that worsen in frequency or severity.
- Indicates info, such as if patient went to a hospital because of event, if it was life threatening, medication changes or new prescribed

adni_recbllog_2009-09-01.csv

- Documentation of Baseline Symptoms Log
- Records all symptoms (recorded by test, no numerical assessment) at time of baseline visit
- The severity, date ceased

adni_reccmeds_2009-09-01.csv

- Concurrent Medication Log
- All medications from screen visit (up to 3 months prior)
- Listed and any while in ADNI

adni_recmhist_2009-09-01.csv

- Medical History
- Provides descriptive information of patient medical history if it is current

adni_registry_2009-09-01.csv

- Registry
- Records of patient's involvement/termination

adni_roster_2009-09-01.csv

- RID matching w/ participant ID

adni_treatdis_2009-09-01.csv

- Early discontinuation and withdrawal
- If patient decides to withdraw from ADNI & why

adni_visits_2009-09-01.csv

- Codes for visits

VISCODE	VISNAME	VISORDER
Sc	Screening	1
Bl	Baseline	2
M06	Month 6	3
M12	Month 12	4
M18	Month 18	5
M24	Month 24	6
M30	Month 30	7
M36	Month 36	8
Uns1	unscheduled	
F	Screen fail	
Nv	Not yet determine	11
M42	Month 42	9
M48	Month 48	10

adni_vitals_2009-09-01.csv

- Vital signs
- Patient's weight, height, SBP, DBP, pulse rate, respiration per min, temperature

adni_avgjacob_2009-09-01.csv

- Average Jacobian Temporal (Paul Thompson's Lab)
- Jacobian maps reflect the percentage of tissue change over time.
- This file is specific to the Temporal Lobe

adni_bsi_2009-09-01.csv

- boundary shift integral summaries
- a measure of cerebral volume changes derived from registered repeat 3-D MRI scans
- BSI determines the total volume through which the boundaries of given cerebral structure has moved

adni_conversion_2009-09-01.csv

- Conversions of patients from MCI → AD, MCI→NL, NL→MCI, AD→MCI, NL→AD.

- Gives patient ID, the month of the visit, the conversion was assessed & what the specific conversions are
- There are 19 conversions in total

adni_strokesum_2009-09-01.csv

- Stroke Summary
- Provides information if patient had stroke, the severity of the stroke, amount of white matter in whole brain, location, stroke type

adni_uaspmvbm_2009-09-01.csv

- Voxel based Morphometry
  - Neuroimaging analysis technique that allows investigation of focal differences in brain anatomy

adni_ucbpet_2009-09-01.csv

- PET ROI Analysis (UCB)
- Brain ROI for regions
- Gives voxels as well as PET values

adni_ucsdvol_2009-09-01.csv

- Derived Volumes
- Volumes of whole brain, ventricles, left hippocampus, right hippocampus ,left mid temporal, right mid temporal, left inferior temporal, right inferior temporal, left fusiform, right fusiform, left entorhinal, right entorhinal

adni_ucsfatrphy_2009-09-01.csv

- Regional Atrophy Rates
- Summary of regional atrophy rates between 1st and last scan. Summarizes changes in both temporal lobes

adni_ucsffresfr_2009-09-01.csv

- Longitudinal Free Surfer
- Freesurfer measures volumes of sub/cortical structures, computes thickness
- Longitudinal looks at changes over time
- For each region – volume, surface area, Cortical Thickness Avg. , Cortical Thickness SD

adni_ucsffsl_2009-09-01.csv

- Longitudinal FreeSurfer
- (same as previous file)

adni_ucsfsx_2009-09-01.csv

- Cross – Sectional FreeSurfer
- (same as previous file)

adni_ucsfsnlvox_2009-09-01.csv

- SNT Hippocampal Volumes
- Uses MRI data
- Volumes of hippocampus (Right & Left)

adni_upennspare_2009-09-01.csv

- Spatial Patterns of Abnormalities for Recognition of early AD
- Score indicates a presence of an AD like spatial pattern of brain atrophy
    - If +, = presence
    - If -, = absence

adni_uucacir_2009-09-01.csv

- UUPET Analysis

adni_uwovent_2009-09-01.scv

- Ventricular Volumes

Appendix 2: Individual Phenotype-Covariate association analysis

The following six tables provide the results of individual phenotype (or PC) to covariate association analysis.

Table A2.1 Association between covariates and Composite Z-score of Cognitive Memory Tests (AVLT and LMT):

Covariate	β_0 **Coefficient**	**T-test**	***p*-value**
Sex	0.01055	0.170	0.865
APOE Genotype	-0.16064	-2.642	0.00847
Age	-0.004173	-0.935	0.350
Diagnosis Status	-0.15967	-3.773	<0.001

Table A2.2. Association between covariates and Rate of Change in Log of Ventricular Volume:

Covariate	β_0 **Coefficient**	**T-test**	***p*-value**
Sex	0.0011683	2.781	0.00559
APOE Genotype	0.0029676	7.449	<0.001
Age	-2.212e-04	-7.624	<0.001
Diagnosis Status	0.0025752	9.480	<0.001

Table A2.3. Association between covariates and Rate of Change in Log of Hippocampal Volume:

Covariate	β_0 **Coefficient**	**T-test**	***p*-value**
Sex	-6.496e-05	-0.317	0.752
APOE Genotype	-0.0014332	-7.411	<0.001
Age	-1.931e-05	-1.308	0.191
Diagnosis Status	-0.0015589	-12.389	<0.001

Table A2.4. Association between covariates and PC1:

Covariate	β_0 **Coefficient**	**T-test**	***p*-value**
Sex	0.1609	1.505	0.133
APOE Genotype	0.85725	8.624	<0.001
Age	-0.020737	-2.704	0.00705
Diagnosis Status	0.84005	12.90	<0.001

Table A2.5. Association between covariates and PC2:

Covariate	β_0 **Coefficient**	**T-test**	***p*-value**
Sex	-0.09778	-1.230	0.219
APOE Genotype	-0.15781	-2.016	0.0443
Age	0.017667	3.108	0.00198
Diagnosis Status	-0.14715	-2.691	0.00733

Table A2.6. Association between covariates and PC3:

Covariate	β_0 **Coefficient**	**T-test**	***p*-value**
Sex	0.14439	2.307	0.0214
APOE Genotype	0.0006762	0.011	0.991
Age	-0.036601	-8.589	<0.001
Diagnosis Status	-0.09182	-2.120	0.0345

Appendix 3: Distribution of plots

The following plots show the distributions of the data for the final number of subjects (n=567).

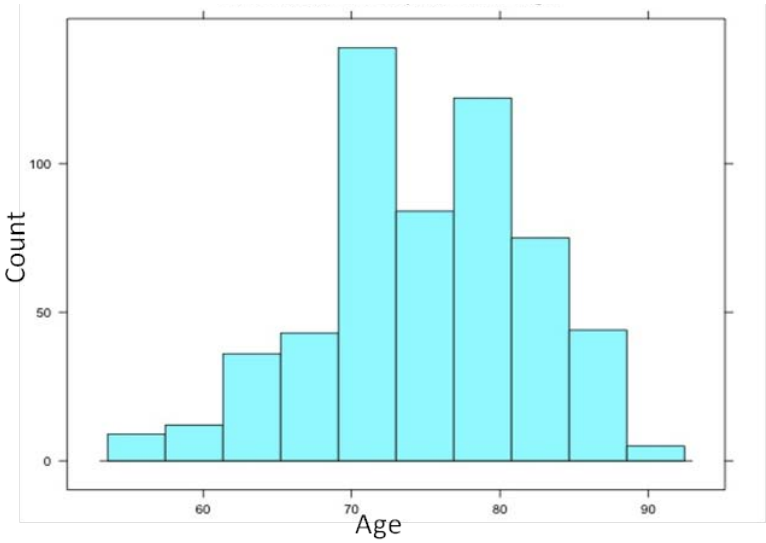


Figure A3.1 Distribution of Patient's age
Overall normal distribution of patient's age at baseline (n=567).

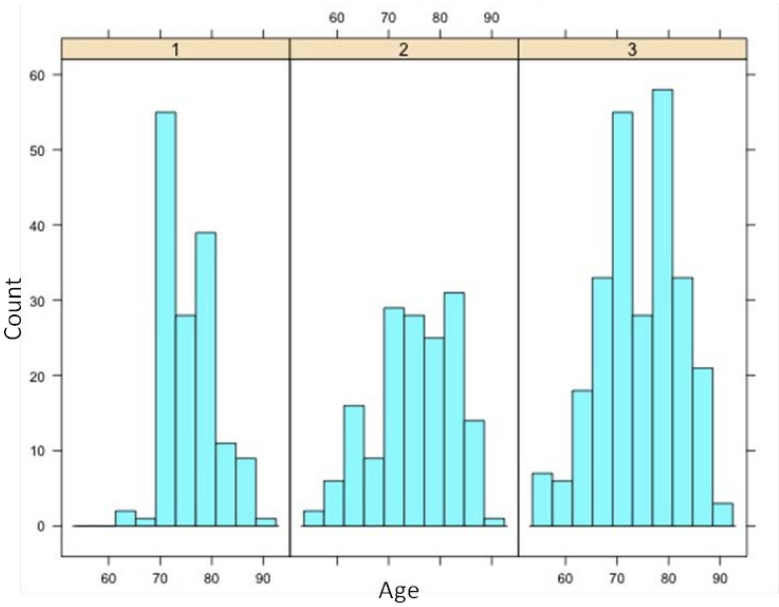


Figure A3.2 Distribution of Patient's age by diagnosis status
Overall normal distribution of patient's age at baseline (n=567) where 1 = Healthy control, 2 = Mild Cognitive Impairment, 3 = Alzheimer's Disease.

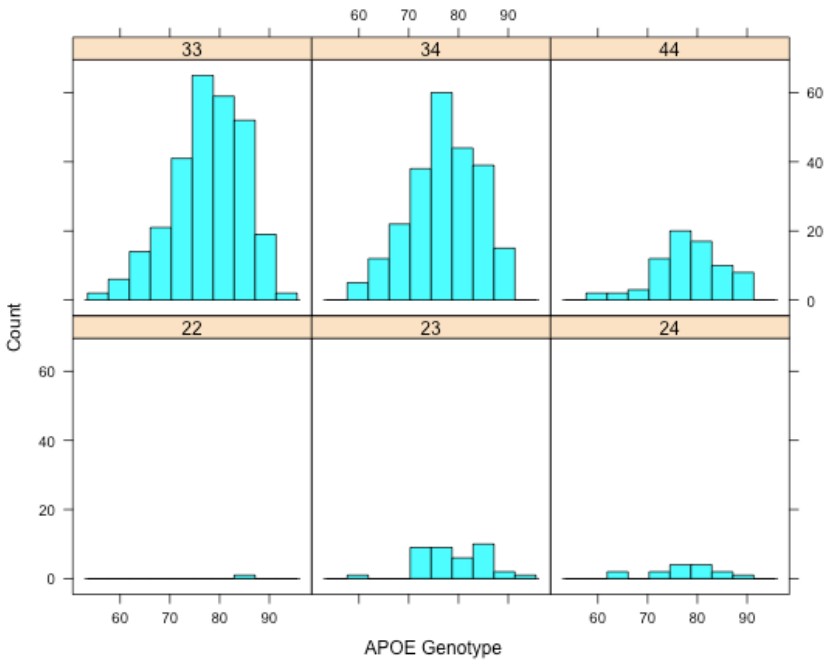


Figure A3.3 Distribution of APOE alleles e2, e3, e4.
Overall normal distribution of patient's age at baseline (n=567) where 22 = e2e2, 23 = e2e3, 24 = e2e4, 33 = e3e3, 34 = e3e4, 44 = e4e4. We also see an insufficient amount of data for patients with APOE genotypes of e2e2, e2e3 and e2e4. As a result, APOE Genotype was transformed into a binary variable indicating the presence or absence of at least 1 APOE e4 allele.

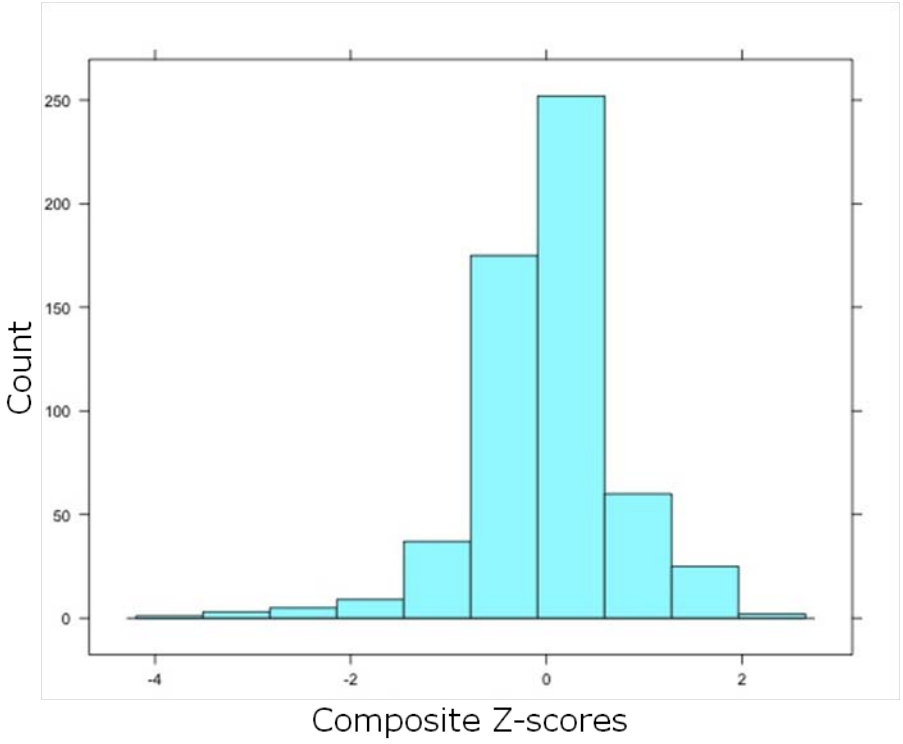


Figure A3.4 Distribution of the Composite Z-scores of the Cognitive Memory Tests
Overall normal distribution of Composite Z-scores of the Cognitive Memory Tests for all patients (n=567).

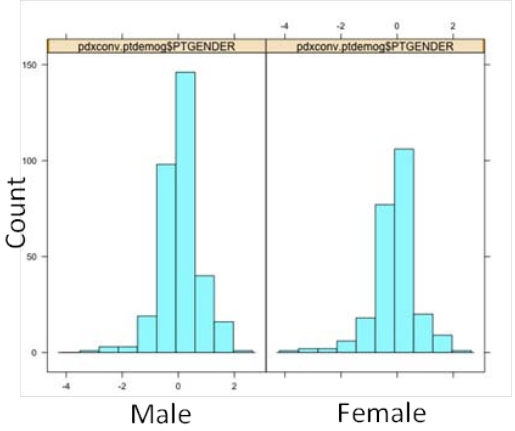


Figure A.3.5. Distribution Composite Z-score of the Cognitive Memory Tests by gender.
Overall normal distribution of memory test z-scores for all patients by gender (n=567)

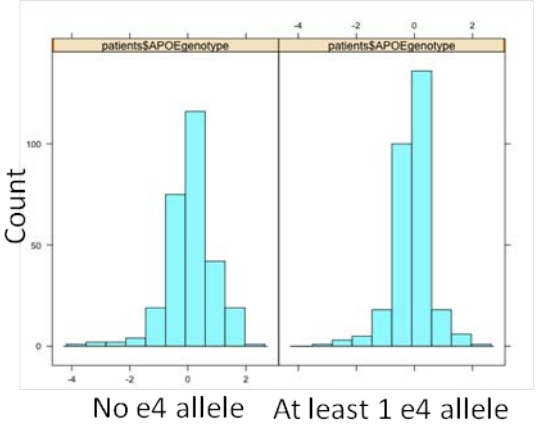


Figure A.3.6 Distribution Composite Z-score of the Cognitive Memory Tests by APOE genotype.
Overall normal distribution of memory test z-scores for all patients by binary variable APOE Genotype (n=567).

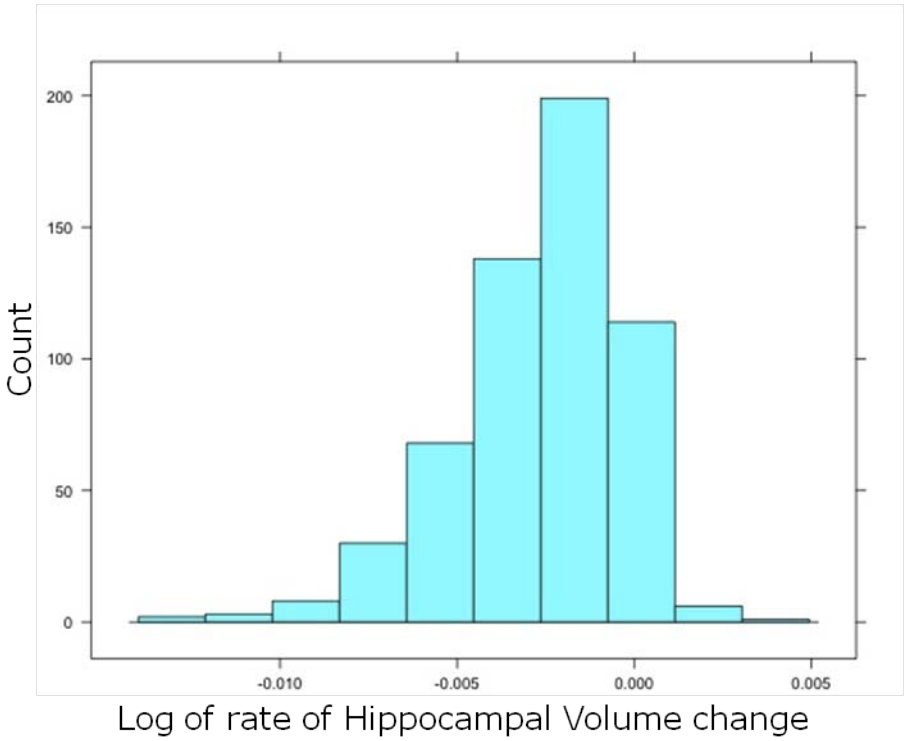


Figure A3.7 Distribution of the log base-2 of rate of hippocampal volume change
Overall normal distribution of rate of hippocampal volume change for all patients (n=567).

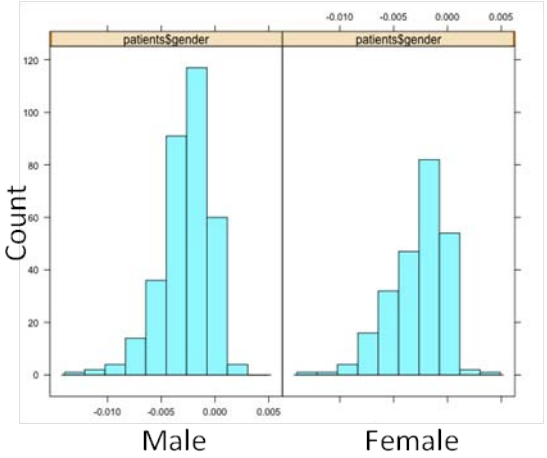


Figure A3.8. Distribution of the log base-2 of rate of hippocampal volume change by gender
Overall normal distribution of rate of hippocampal volume change for all patients by gender (n=567)

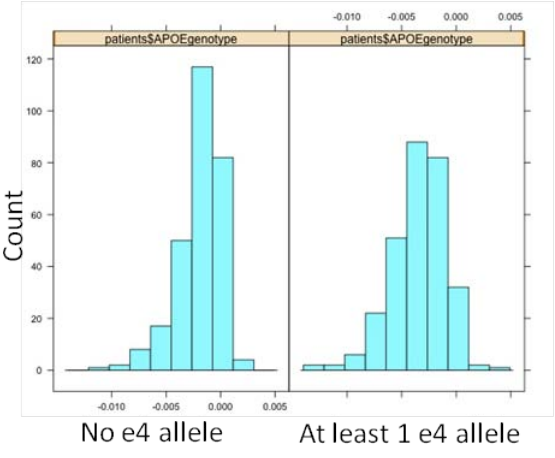


Figure A3.9. Distribution of the log base-2 of rate of hippocampal volume change by APOE genotype
Overall normal distribution of rate of hippocampal volume change for all patients by binary variable APOE Genotype (n=567).

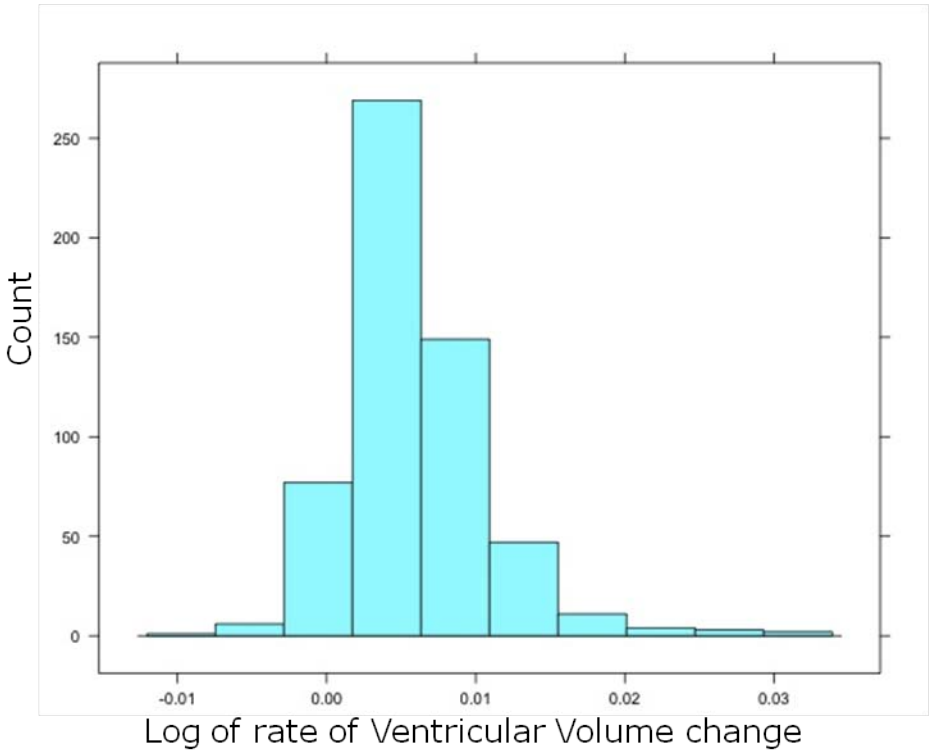


Figure A3.10 Distribution of the log base-2 of rate of ventricular volume change.
Overall normal distribution of rate of ventricular volume change for all patients (n=567).

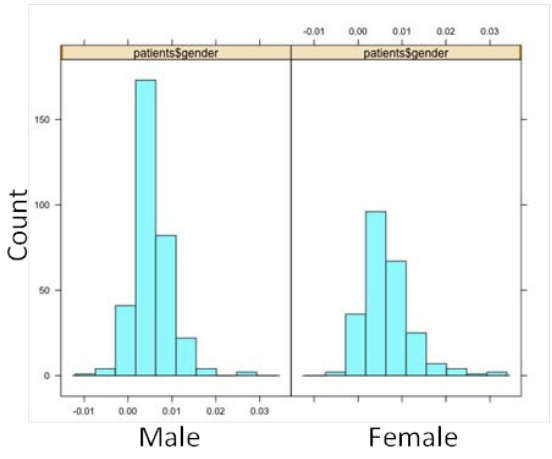


Figure A3.11 Distribution of the log base-2 of rate of ventricular volume change by gender.
Overall normal distribution of rate of ventricular volume change for all patients by gender (n=567)

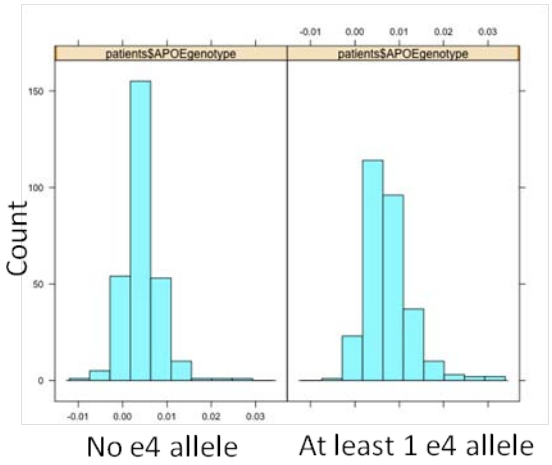


Figure A3.12 Distribution of the log base-2 of rate of ventricular volume change by APOE genotype
Overall normal distribution of rate of ventricular volume change for all patients by binary variable APOE Genotype (n=567).

Appendix 4: Genotypic Quality Control Distributions

These distributions show the SNP call rates and the Minor Allele Frequencies of the final genotypes incorporated in the study (n=543715 SNPs)

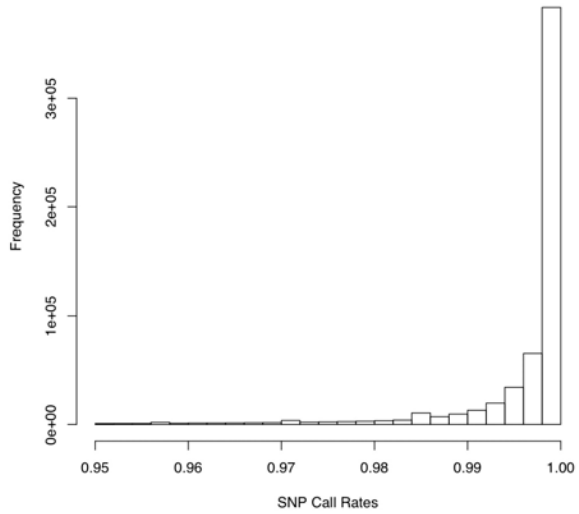


Figure A4.1 Histogram of SNP Call Rates (n = 543715 SNPs)
Majority of SNPs possess a call rate of 100%. Any SNPs with a call rate < 85% were removed from the study

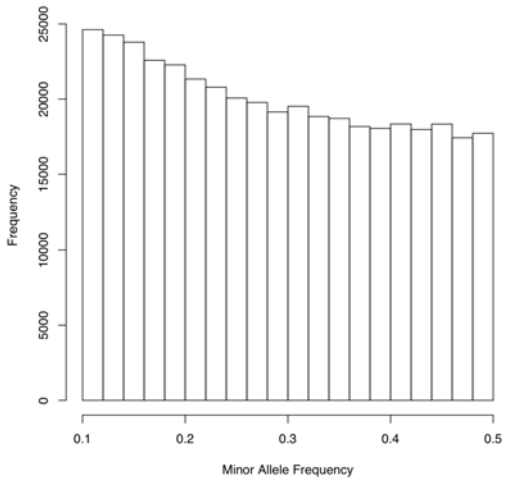


Figure A4.2 Histogram of Minor Allele Frequencies (n = 543715 SNPs)
Any SNPs with a minor allele frequency < 10% were removed from the study

Appendix 5: Pairwise comparisons of PCs, highlighting covariates in study

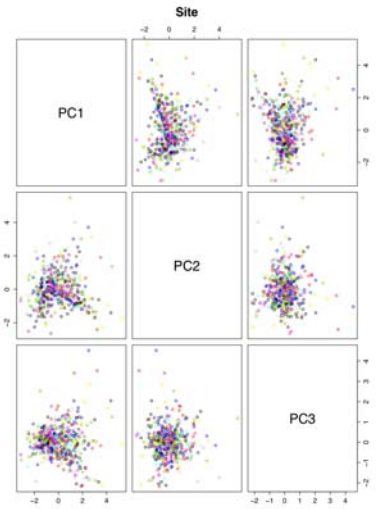


Figure A5.1. Pairwise comparison of PCs (color coding indicates each site (n=57) for data collection)

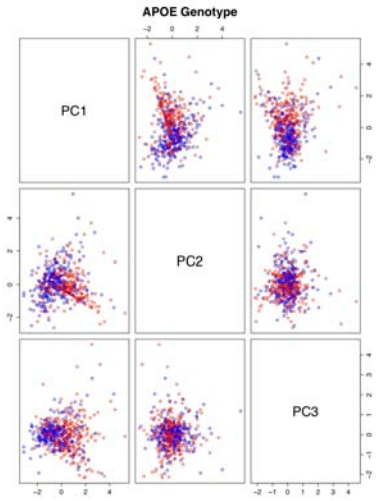


Figure A5.2. Pairwise comparison of PCs (color coding indicates APOE genotype (E4+ or E4-))

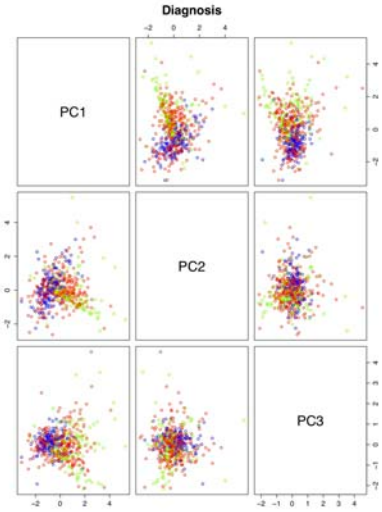


Figure A5.3. Pairwise comparison of PCs (color coding indicates diagnosis)

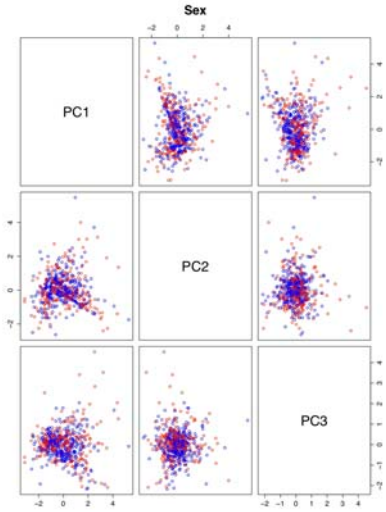
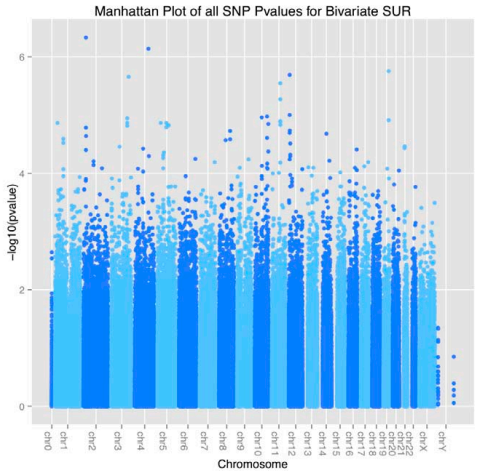


Figure A5.4. Pairwise comparison of PCs (color coding indicates gender)

Appendix 6: Manhattan Plots and QQ Plot for SUR and Univariate Methods

a.



b.

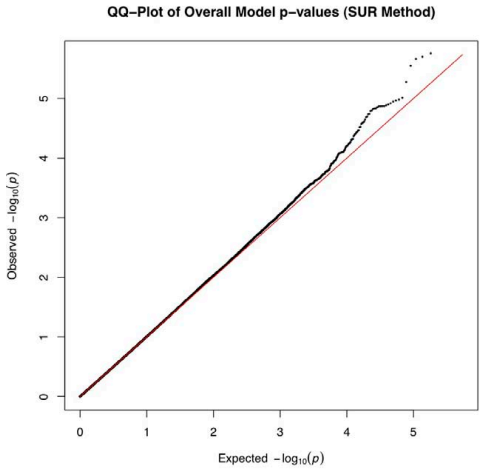
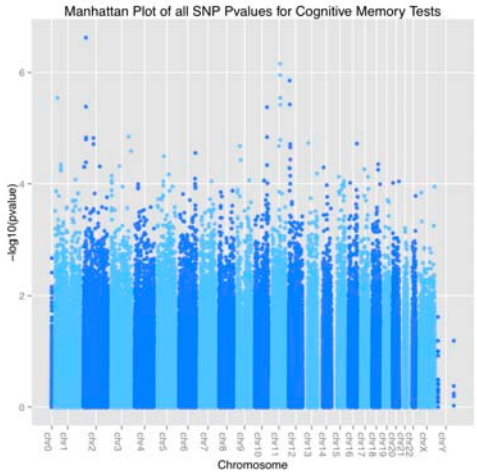


Figure A6.1. (a) Manhattan Plot and (b) QQ Plot the Results from the Bivariate SUR Method

a.



b.

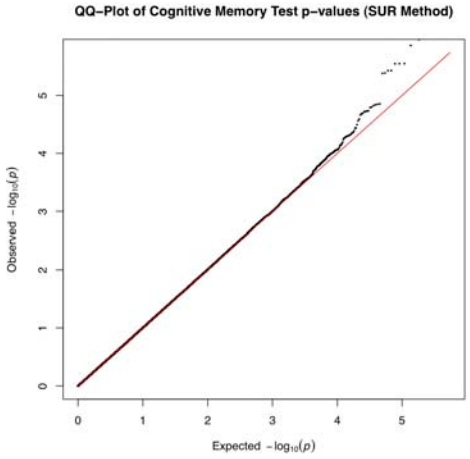
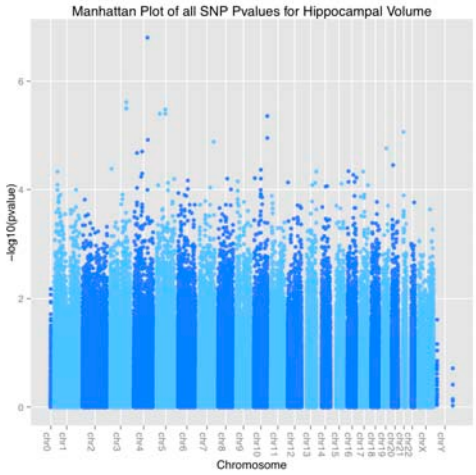


Figure A6.2. (a) Manhattan Plot and (b) QQ Plot the Results from Univariate association results for Cognitive Memory Tests

a.

b.



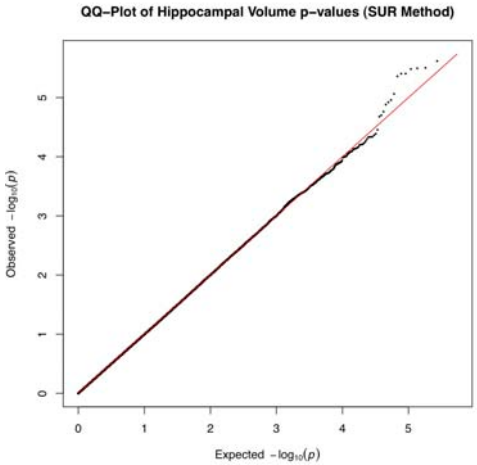


Figure A6.3. (a) Manhattan Plot and (b) QQ Plot the Results from Univariate association results for Hippocampal Volume

Appendix 7: Manhattan Plots and QQ Plot for PCA Method

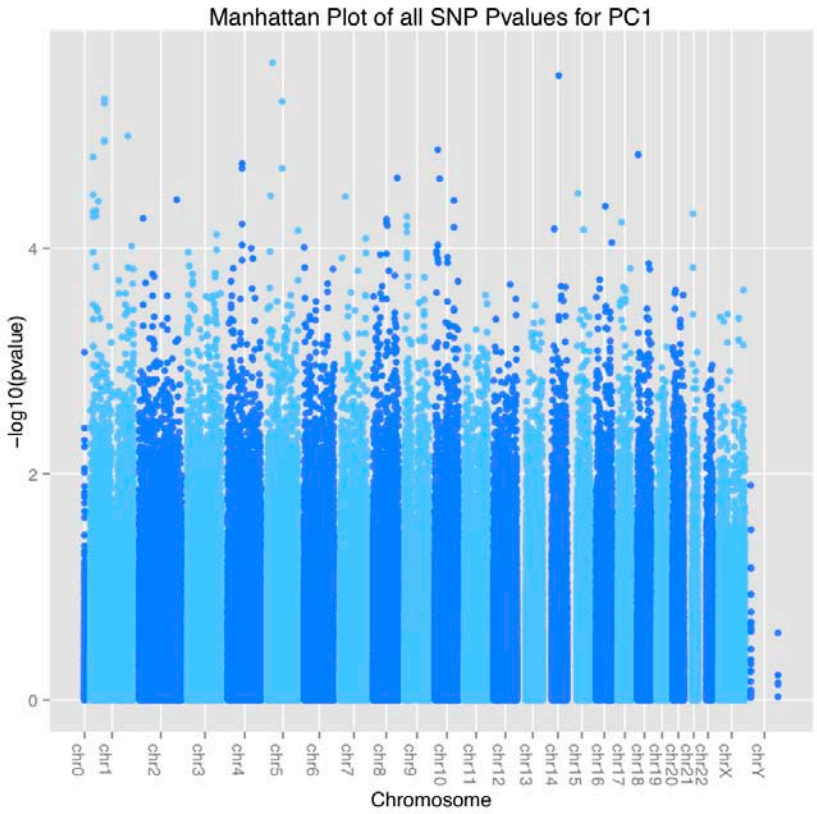


Figure A7.1 Manhattan plot of PC1 association results

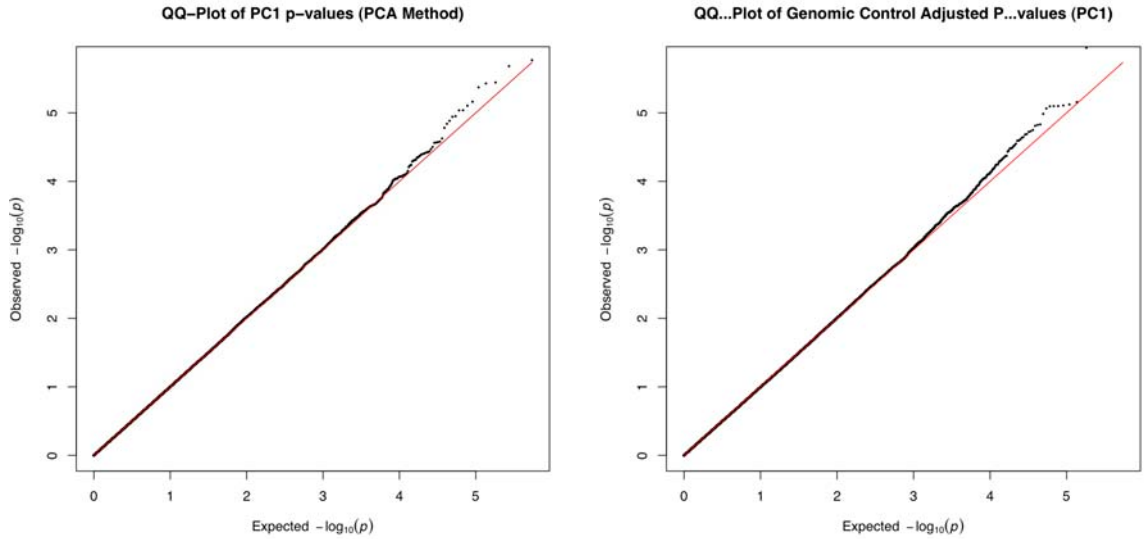


Figure A7.2 QQ plot of PC1 association results before (left) and after (right) Genomic Control

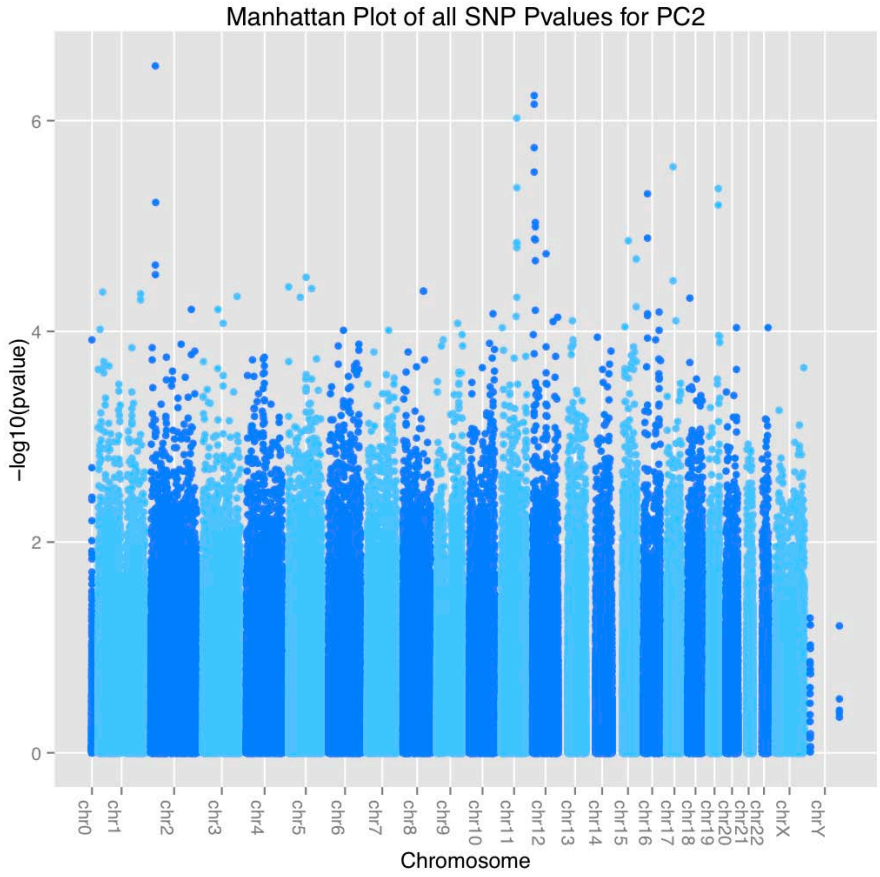


Figure A7.3 Manhattan plot of PC2 association results

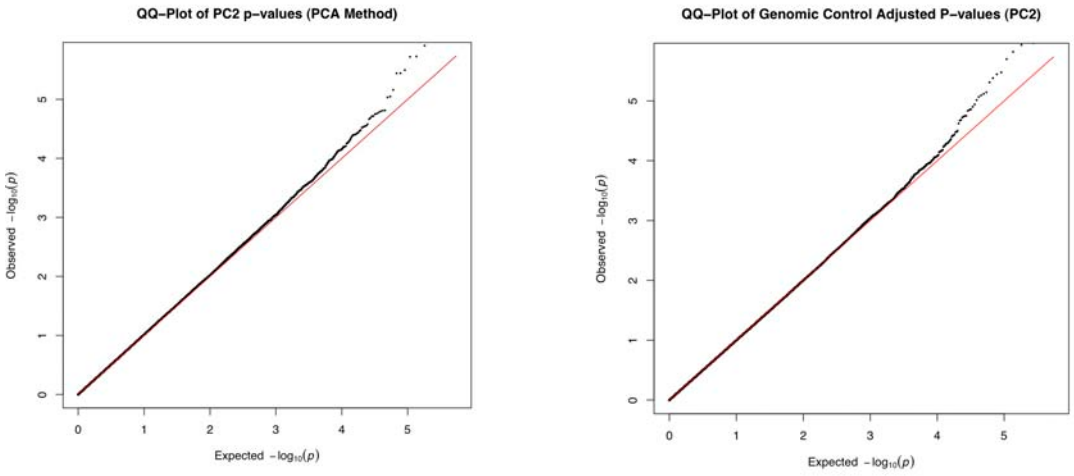


Figure A7.4 QQ plot of PC2 association results before (left) and after (right) Genomic Control

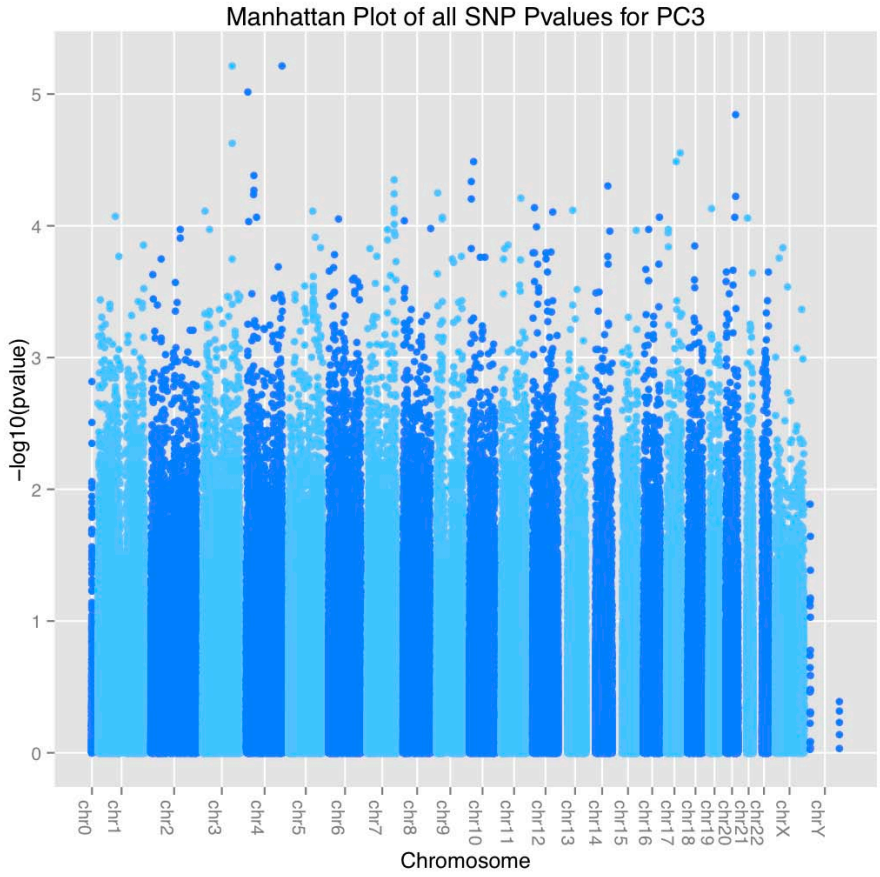


Figure A7.3 Manhattan plot of PC3 association results

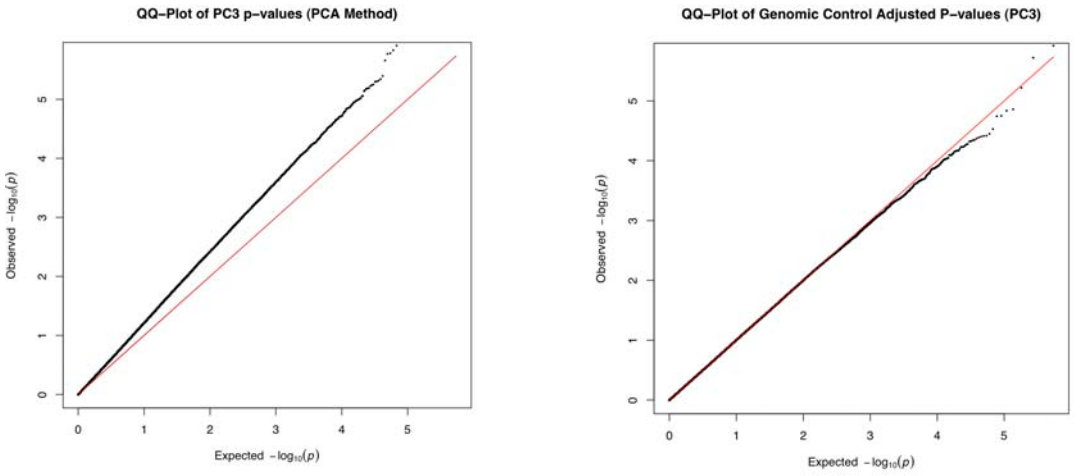


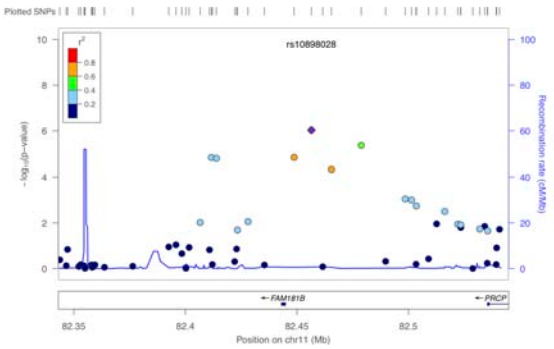
Figure A7.6 QQ plot of PC3 association results before (left) and after (right) Genomic Control

Appendix 8: Focus Plots

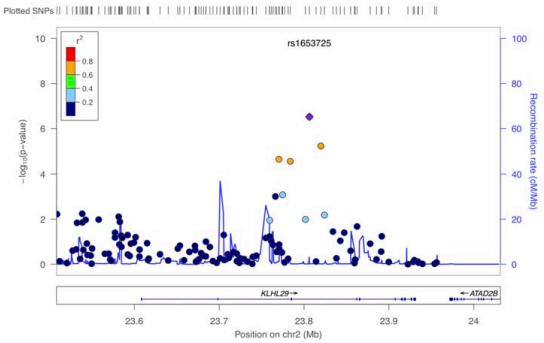
Focus plots for genes that annotated to SNPs with p-values < 10^{-5}.

Focus plots for genes annotated to significant SNP determined by Bivariate SUR method

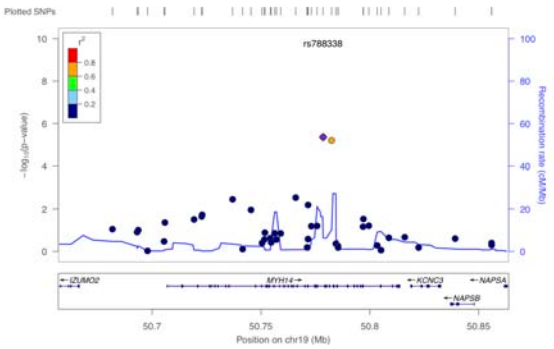
FAM181B



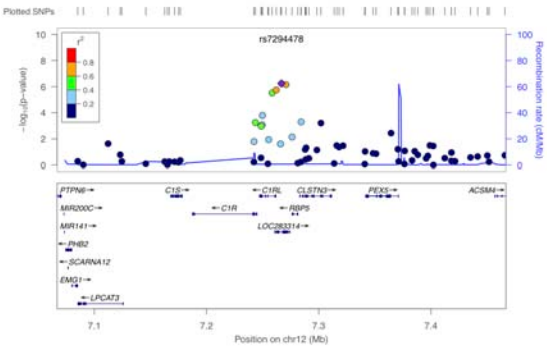
KLHL29



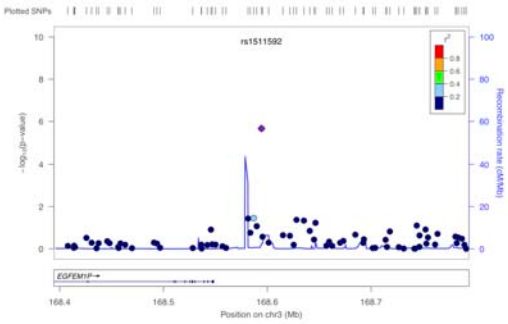
MYH14



C1RL-AS1

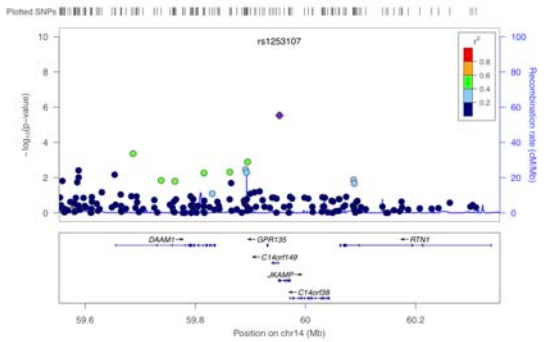


EGFEM1P

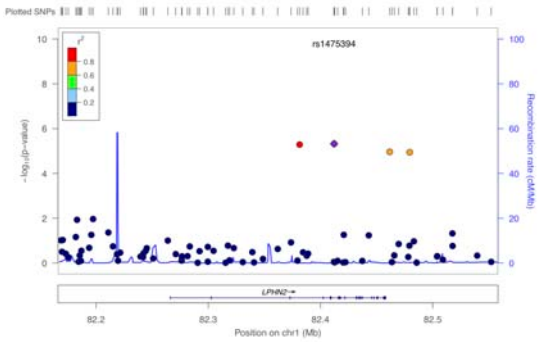


Focus plots for genes annotated to significant SNP associated with PC1 (PCA Method)

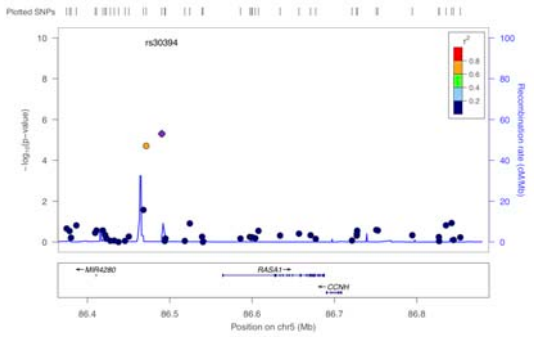
JKAMP



LPHN2

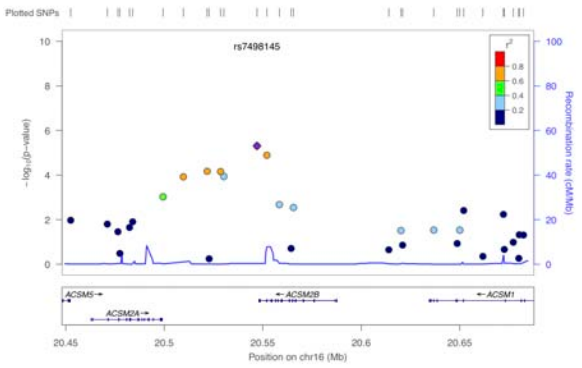


RASA1

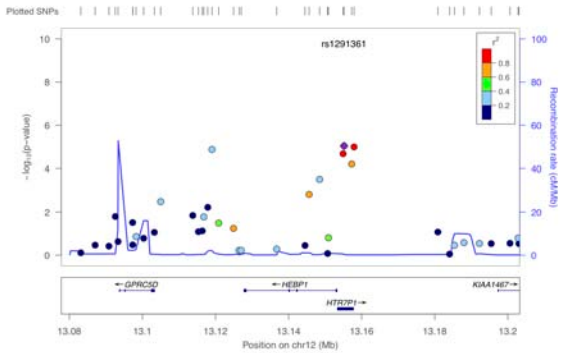


Focus plots for genes annotated to significant SNP associated with PC2 (PCA Method)

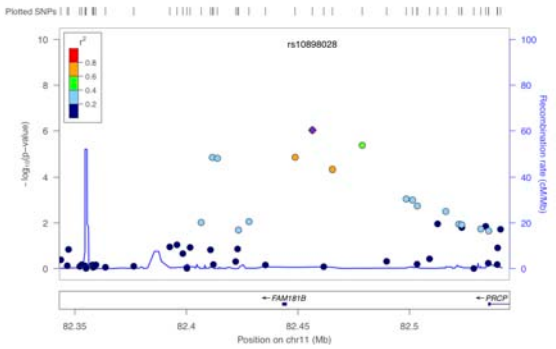
ACSM2B



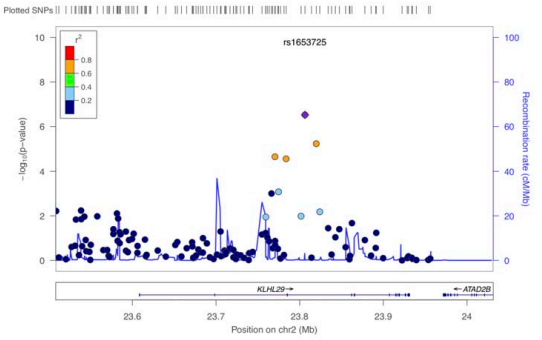
HEBP1



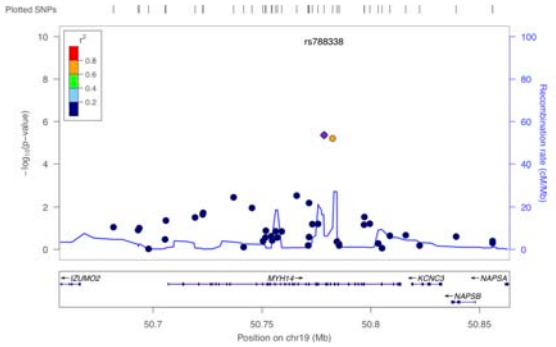
FAM181B



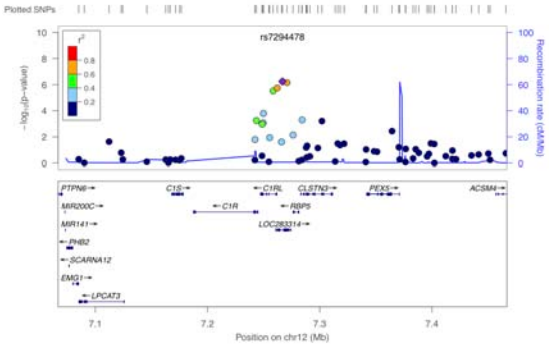
KLHL29



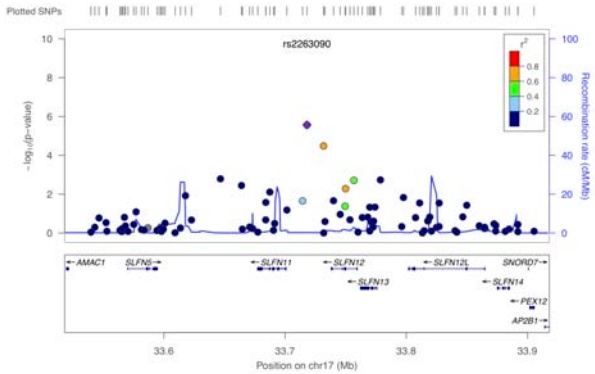
MYH14



C1RL-AS1

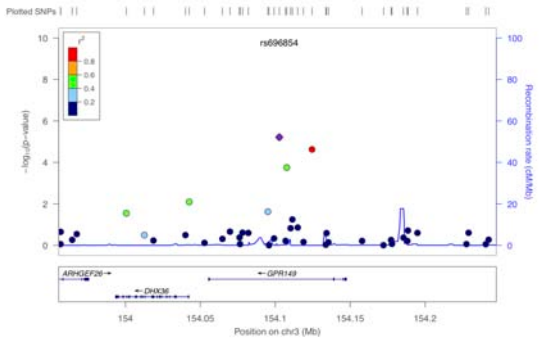


SLFN11

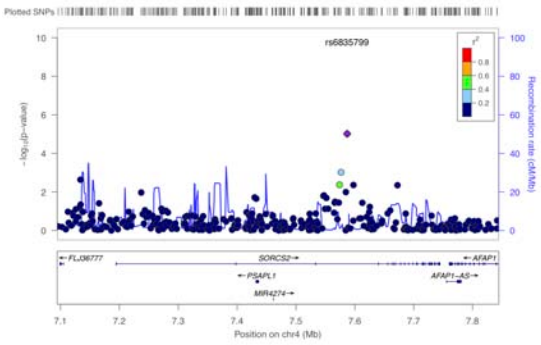


Focus plots for genes annotated to significant SNP associated with PC3 (PCA Method)

GPR149



SORCS2



TRIML2