# EXPLORING THE GENETIC ARCHITECTURE OF LATE-ONSET ALZHEIMER DISEASE IN AN AMISH POPULATION

By

Anna Christine Cummings

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements for

the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

December, 2012

Nashville, Tennessee

Approved:

Professor Dana C. Crawford

Professor Jonathan L. Haines

Professor William K. Scott

Professor Michael G. Tramontana

Professor Bingshan Li

To my husband, Christopher, and son, Titus

And

To my parents, Bernard and Christine Davis

## ACKNOWLEDGEMENTS

The work presented in this dissertation was supported by NIH grants AG019085 to Jonathan L. Haines and AG019726 to William K. Scott, a Discovery Grant from Vanderbilt University, and Michael J Fox Foundation grants. I would like to thank all individuals and communities for so graciously participating in these studies. None of this work would have been possible without them.

The work presented here was guided and greatly improved by the input from my thesis committee: Dana Crawford (my committee chair), William Scott, Bingshan Li, and Michael Tramontana. Special acknowledgements are due to my mentor, Jonathan Haines. I am especially grateful for his expertise, guidance, time, and patience.

I am thankful for all the members of the Haines lab (Nathalie Schnetz-Boutaud, Ping Mayo, Brent Anderson, Melissa Allen, Jacob McCauley, William Bush, Kylee Spencer, Sharon Liang, Rebecca Zuvich, Olivia Veatch, Mary Davis, Joshua Hoffman, and Laura D'Aoust) for being helpful in so many ways and for creating an enjoyable, fun, and collaborative work environment.

I would also like to thank all members of the CHGR. So many people have played a role in this work for which I am very grateful. A special thanks is due to Lan Jiang who patiently provided much training for genetic analyses in the Amish. I would also like to thank all of my fellow students who have been so kind and supportive along the way.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF APPENDICES

# CHAPTER I

## INTRODUCTION

Alzheimer Disease (AD) is the most common cause of dementia, affects over 5 million individuals over the age of 65 in the United States (1), is the fifth-leading cause of death in the United States for individuals over the age of 65 (2), and is an increasingly serious public health issue. AD is a progressive neurodegenerative disorder of the brain characterized by loss of memory and cognitive abilities, development of neuropsychiatric symptoms and behavioral changes, and loss of daily independent function. With inadequate treatments and no cure, the nature of this disease puts a heavy burden on individuals, their families, caregivers, and society as a whole. With our aging population, this burden will only increase as the number of affected individuals is expected to triple by 2050 (1).

AD can be divided into two categories: early-onset and late-onset. Individuals younger than 65 have the early-onset form but account for less than 5% of all AD cases (3). Dominant mutations in three genes cause susceptibility to the majority of early-onset familial AD: amyloid precursor protein [APP] (4) and presenilin 1 and 2 [PS1, PS2] genes (5-7). However, these three genes combined only contribute to less than 2% of all cases of AD. The much more common form, late-onset Alzheimer disease (LOAD), describes AD when it occurs in individuals older than or equal to 65 (8). Unlike early-onset where most of the genetic risk is identified and follows a simple Mendelian pattern, the majority of the genetic risk of LOAD is unexplained and has a much more

complex architecture. My thesis work, therefore, is aimed to better understand the genetic architecture of LOAD and to identify at least one novel LOAD risk locus. In this chapter, I provide an overview of LOAD, including the known pathophysiology, diagnoses, risk factors, and the search for genetic risk factors. I will also present the background information about the Amish population and rationale for using the Amish as our study population.

## Pathophysiology and diagnosis of Alzheimer disease

Although the pathophysiology of AD is yet to be completely explained, the initiation of AD is thought to be triggered by the generation of peptide oligomers (amyloid beta, also known as beta amyloid) from amyloid precursor protein (APP) in the brain.(9-11). The accumulation of amyloid beta (Aβ), called plaques, occurs outside the neurons and is almost always accompanied by twisted strands of hyperphosphorylated tau protein, called tangles, inside the neurons. However, there are rare cases of plaque-only and tangle-only Alzheimer disease (12;13). Recent advances in studying blood and cerebrospinal fluid biomarkers detecting the level of beta amyloid in the brain have potential to detect the pathophysiological process (14;15). The plaques, as they interfere with synaptic communication, and the tangles, as they inhibit essential transportation inside the neurons, are hypothesized to cause the nerve cell damage and death in the brain, which is characteristic of Alzheimer patients. The cell death is so extensive that the shrinking of the brain is visible via neuroimaging measures, another category of biomarkers.

The timing of the initiation of these deleterious events remains a mystery, but advances in biomarker technologies detecting the Aβ pathophysiology and the subsequent neurodegeneration are providing clues. For LOAD, most symptoms and diagnoses of AD begin after the age of 65; however, changes in the brain due to AD might begin as early as 20 years before the onset of symptoms (16). To reflect that knowledge, the National Institute on Aging (NIA) and the Alzheimer Association proposed new diagnostic criteria in 2011, but these have not yet been implemented in clinical practice. These new criteria define three stages of Alzheimer disease: 1) preclinical Alzheimer disease, 2) MCI (mild cognitive impairment) due to Alzheimer disease, and 3) dementia due to Alzheimer disease. Preclinical Alzheimer disease can be diagnosed only on the basis of biomarker detection since no symptoms have occurred at this stage. When noticeable changes in cognition appear in addition to the biomarker evidence of AD pathophysiology, an individual would be diagnosed with 'MCI due to Alzheimer disease.' Not everyone with MCI goes on to develop dementia due to Alzheimer disease. Therefore, biomarker testing could help to distinguish those who will go on to develop AD from those who will not progress. The final stage, stage three, of AD occurs when a person exhibits the clinical symptoms to make a diagnosis of probable or possible AD, as is the current practice for diagnosing probable or possible AD. At this stage biomarkers are merely for confirmation of the underlying pathophysiology. Definite AD can only be defined with a post-mortem pathology report. While there are three stages defined, clinicians and researchers also recognize that AD evolution is a continuum and that the boundaries between the stages can be blurred (16). These new criteria will allow much earlier diagnoses, and therefore, much earlier introductions of interventions that could be much more effective. More effort is

needed in the area of biomarker research to develop standards to implement these new diagnostic criteria in clinical practice.

The previously established and currently used diagnostic criteria were developed by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer Disease and Related Disorders Association (NINCDS ADRDA) in 1984 (17). These classical criteria rely on clinical symptoms as reported by patients and informants and as indicated by neuropsychological testing to assess eight areas of cognition that can be impaired in individuals with AD: orientation, memory, language skills, praxis, attention, visual perception, problem-solving skills, and social function/daily living (17). These diagnoses are at least 80% accurate (>90% for individuals with dementia) (18); however, a definite diagnosis of AD can only be established with an autopsy.

We have employed the NINCDS ADRDA criteria in the studies presented in Chapters III and IV (See Chapter III for more details). Because we have no autopsy information we are not able to make any definite diagnoses of AD, only probable and possible AD. Therefore one limitation of our study is that some of the seemingly cognitively normal patients and some of the MCI patients could go on to develop AD. To minimize those potential issues, individuals diagnosed with MCI are classified as 'unknown' in our analyses, and we only seek individuals over the age of 65 for our study to lessen the chances of enrolling a control that might go on to develop AD. We also follow up all individuals (typically after three years) with 'unclear' diagnoses, including MCI, to check for regression to AD status.

## Epidemiology of and risk factors for Alzheimer disease

All the data to date suggests that a complex combination of genetic and environmental components determine if someone will get LOAD (19). Many risk factors are under debate, but the most commonly accepted are: age, lifestyle, cardiovascular disease (CVD) risk factors, race, education, head trauma, family history, and genetics (3).

The single greatest risk factor for Alzheimer disease is age. While some difficulty with memory is typical of the normal aging process, the symptoms of Alzheimer disease are much more severe—to the point where they dramatically interfere with everyday tasks. The population prevalence of LOAD dramatically increases with the age of the individuals (1). Each 5-year incremental increase in age doubles the percent of people with AD (20). With no cure nor reliable treatment for Alzheimer disease, the increased life expectancy of our population will continue to increase the prevalence of Alzheimer disease.

A person's lifestyle includes a variety of factors—physical and mental activity, diet, sleeping habits, and smoking—that can affect a person's likelihood of developing Alzheimer disease. Some of these lifestyle factors overlap factors modulating cardiovascular disease risk. In fact, substantial research supports that there is a strong connection between a person's heart health and brain health (21). Diet-related factors such as high cholesterol, obesity, and type 2 diabetes have all been linked to AD risk in addition to CVD (21-24). Conversely, adherence to a 'Mediterranean diet', which is rich in vegetables, fruits, legumes, grains, and unsaturated fatty acids, seems to protect against both AD and CVD (25). This heart health–brain health link could at least partially explain the higher risk for LOAD for African Americans and Hispanics

compared to whites (26;27) since African Americans and Hispanics also have higher risks for CVD. Sedentary lifestyle and smoking also place people at higher risk for developing AD and CVD (28-30).

Mental activity, including years of education also impacts LOAD risk. Fewer years of education increase a person's likelihood of developing AD, even after adjusting for socioeconomic status (31). The reason for this correlation is unknown, and it is unclear which is the cause and which is the effect. Proposed explanations include the idea that educated individuals have more of a 'cognitive reserve', helping those individuals to withstand the attacks on their cognition (32;33).

Moderate to severe head injuries also increase risk to Alzheimer disease by 2 to 4 and a half fold (34;35). The head injuries must result in unconsciousness or post-traumatic amnesia lasting for at least 30 minutes for moderate head injury and more than 24 hours for severe head injury to confer these risks. There has been some evidence for an interaction between moderate to severe head injury and *APOE* ε4 carrier status (34;36;37). Mild head injuries do not seem to confer the same risks, but more research is needed in this area.

The risk factors described above, if shared among family members, would at least partially explain why individuals with a family history of LOAD have a greater chance of developing AD. However, family history increasing one's risk for developing LOAD also suggests that genetic factors are involved. A person who has a sibling with AD is 4-5 times more likely to also develop AD compared to the general population (38-40). The estimates of the heritability of AD using twin studies ranges from 58% to 80% (41;42). The remainder of this chapter discusses the previous and more recent methods,

including the approaches taken in this thesis work, used to discover Alzheimer disease risk genes.

**The search for genetic risk factors for late-onset Alzheimer disease**

The first LOAD risk gene, *APOE*, was identified in 1993 via genetic linkage mapping in combination with gene function information (43-48). The common *APOE* ε4 allele increases susceptibility to both early- and late-onset AD, and the ε2 allele decreases risk (49-51). Considering LOAD is a genetically complex disease, *APOE* has a strong effect with an odds ratio generally greater than 3 (alzgene.org), but at most explains 22% to 50% of the 80% genetic effect of LOAD (42;50;52).

The remaining genetic component for LOAD risk remained unexplained for the next 16 years despite an abundance of effort, including linkage studies, candidate gene association studies, and genome-wide-association studies. From the late 1990's into the early 2000's, many genome-wide linkage studies using microsatellite markers were attempted (53-63). Regions on chromosomes 9, 10, and 12 seemed particularly promising, but no gene could be confirmed for those regions. Additionally, associations between LOAD and candidate genes on every chromosome in the human genome were reported, but none could be consistently replicated (64).

With the initiation of the International HapMap project in 2002, common variations, in the form of single-nucleotide polymorphisms (SNPs), were characterized and catalogued in different populations for the entire genome. The genotypes for the SNPs provide estimates of linkage disequilibrium (LD) across the genome (65-67). This effort led to the emergence of the genome-wide association study (GWAS) design, in

which 250,000 to more than 1,000,000 SNPs are genotyped, to scan the genome for associations to disease. The first GWAS for LOAD were published in 2007 (68;69), but the only significant locus identified was *APOE*. Eight subsequent GWAS, two of which only focused on candidate genes (Grupe et al and Feulner et al), also failed to generate any novel LOAD genes that could be discovered and replicated at a genome-wide significance level (70-77). All of the first ten GWAS discovery and replication datasets were under 2,000 cases and 2,000 controls. Due to the heterogeneous nature of LOAD and the small effect sizes likely to be detected, in retrospect these studies were all underpowered to find anything but *APOE* which has a fairly large effect size. From these studies we learned that much of the remaining genetic effects would require more powerful approaches to be discovered.

Then in 2009 both Harold et al and Lambert et al published the first consortia-derived GWAS for LOAD. With almost 6,000 total cases and more than 10,000 total controls, Harold et al identified and replicated SNPs near *CLU* and *PICALM* at genome-wide significance (78). Concurrently, Lambert et al replicated the *CLU* finding and achieved genome-wide significance for *CR1* when they combined their discovery and replication datasets (79) with 6,000 total cases and more than 8,600 total controls. Lambert et al also found nominal significance for *PICALM*. *CLU* was also replicated at genome-wide significance by Seshadri et al in 2010 (80). Also at genome-wide significance, Seshadri et al replicated *PICALM* and identified *BIN1* as a potential LOAD gene, which would later be confirmed. In 2011, Naj et al published the replication of *BIN1*, *CR1*, *CLU*, and *PICALM*, and also identified and confirmed *MS4A*, *CD2AP*, *EPHA1*, and *CD33* as novel LOAD loci (81). In the same issue of Nature Genetics,

Hollingworth et al provided additional support for all of the same genes and also added *ABCA7* as a tenth LOAD susceptibility locus (82).

Recently, *SORL1* has also been gaining support as a LOAD gene. After mixed results in previous candidate gene studies, recent studies using larger sample sizes have confirmed the association of SNPs in *SORL1* with LOAD (83-86).

One of the most recent advances in genome technology has been next-generation sequencing, including whole genome sequencing. Despite the many successes of GWAS, the heritability of most common diseases remains unexplained. GWAS only query common variation by genotyping a subset of the known SNPs and relying on LD to capture other SNPs not directly genotyped. Therefore, portions of the genome are inevitably underrepresented by GWAS because of they simply could not be included in the design of the SNP genotyping assay and because of differences in LD between different datasets. Rare variants are ignored when designing GWAS, and although rare variants could be tagged by a GWAS, it would take a follow-up sequencing study to identify the rare variant as the causal variant. The realization of the limitations of GWAS and the availability of this new technology has generated increased interest in rare variants (more about this in Chapter IV). Taking this approach, Jonsson et al discovered a protective mutation in *APP* for LOAD (Jonsson et al 2012). Before Jonsson et al's report, mutations in *APP* were only known to be causative for early-onset AD.

To date, there are twelve known LOAD genetic loci: *APOE*, *CR1*, *CLU*, *PICALM*, *BIN1*, *MS4A*, *CD2AP*, *EPHA1*, *CD33*, *ABCA7*, *SORL1*, and *APP* (Table 1.1). However, these genes combined do not explain the entire genetic component of Alzheimer disease. Although earlier reports suggested that *APOE* might explain as much as 50% of the 80% heritability, So et al estimated that 18% of the total variance in AD risk, or 23% of the

~80% heritability of AD, can be explained by SNPs in *APOE*, *CR1*, *CLU*, and *PICALM*. *CR1*, *CLU*, and *PICALM* combined only contribute ~1% of the variance (52). Naj et al estimated the population attributable fractions for *CR1*, *BIN1*, *CD2AP*, *EPHA1*, *CLU*, *MS4A4*, *PICALM*, *ABCA7*, and *CD33* to individually range from 3% to 6% (81). The population attributable fraction is the percentage of AD cases that could be prevented if the risk factor was removed. This calculation is not the same as the percentage of total variance in AD, also known as locus-specific heritability, calculated by So et al (52). Therefore, as much as 75% of the heritability of LOAD could remain to be explained.

**Table 1.1 Late-onset Alzheimer disease genes.** 'Study' indicates the first study to publish a significant association for the gene. 'SNP/allele' and 'OR' are the SNPs or alleles and odds ratios that the indicated study initially published.

| Gene | Study | SNP/allele | OR |
|---|---|---|---|
| **APOE** | **Corder et al** | **E4/E2** | **3.78** |
| **CLU** | **Harold et al** | **rs11136000** | **0.86** |
| **PICALM** | **Harold et al** | **rs3851179** | **0.86** |
| **CR1** | **Lambert et al** | **rs6656401** | **1.21** |
| **BIN1** | **Seshadri et al** | **rs744373** | **1.13** |
| **MS4A** | **Naj et al** | **rs4938933** | **0.88** |
| **CD2AP** | **Naj et al** | **rs9349407** | **1.14** |
| **EPHA1** | **Naj et al** | **rs11767557** | **0.85** |
| **CD33** | **Naj et al** | **rs3865444** | **0.88** |
| **ABCA7** | **Hollingworth et al** | **rs3764650** | **1.23** |
| **SORL1** | **Rogaeva et al** | **multiple SNPs** | **1.70-1.84** |
| **APP** | **Jonsson et al** | **rs63750847** | **0.189** |

While large-scale studies in the general population have produced most of the recent discoveries of LOAD genes, these approaches will not identify all of the genetic variations underlying LOAD. Genetic heterogeneity, i.e. different genes in different

groups and individuals contributing to LOAD susceptibility, complicates additional gene discoveries and replication of discoveries, particularly in studies of the general population, which has been the standard practice. The small attributable risk of each polymorphism to the overall genetic variance in the general population makes many studies underpowered (64).

## The utility of genetically isolated populations

Using a more genetically homogeneous study population is one approach to overcome this problem. Isolated populations are a valuable resource for genetic studies (87-89). The isolated expansion of the population from a small number of founders restricts the introduction of new genetic variation(90), so it can be expected that these unique groups' genomes would contain a more homogeneous set of disease risk genes. Many isolated populations have large families and often keep extensive genealogy records, making extended pedigree construction feasible. Linkage analysis of large pedigrees has proven to be a valuable tool for genetic studies, particularly for AD since all four of the first verified AD genes (*APP*, *PSEN1*, and *PSEN2* for early-onset and *APOE* for late-onset) were initially localized via linkage analysis. The recent discovery of a mutation in APP conferring protection for LOAD was performed in an isolated Icelandic population (Jonsson et al 2012). Studies of population isolates for AD have also been performed in an isolated Finnish population (63), in a Netherlands population (91), and in the relatively isolated Caribbean Hispanics (92).

## The Amish

The Amish communities of middle Ohio (Holmes County) and northern Indiana (Elkhart, LaGrange, and Adams counties) are a genetically isolated founder population, originating from two waves of immigration of Swiss Anabaptists, seeking freedom from religious persecution, into the U.S. In the early 1700's the first wave of immigration brought the Anabaptists to Pennsylvania. In the early 1800's some of these immigrants moved to Holmes County, OH, while a second wave of immigration from Europe established more Amish communities in other areas of Ohio and Indiana (including Adams County). Later, Elkhart and LaGrange County Amish communities were started by some of the Amish from Pennsylvania and Ohio (including Holmes County) moving to these new locations (93-95).

The Amish marry almost exclusively within the community and have large families, providing pedigrees with multiple affected individuals for analyses. The Anabaptist Genealogy database (AGDB) (96;97) and the Swiss Anabaptist Genealogical Association (SAGA) keep thorough family history records, providing necessary and critical pedigree information. Because of their faith, the Amish lead a strict and traditional lifestyle and, therefore, have more homogeneous environmental exposures than the general population.

Compared to the general population in which many genes are contributing to LOAD, the relatively homogeneous Amish population is likely to contain a smaller set of risk alleles, each with a theoretically increased population attributable risk, thereby increasing detection power. The relatively recent expansion of the population from a small number of original founders plus isolation results in this reduced amount of

genetic variation (90). The risk alleles found in the Amish population, however, should be a subset of the risk alleles in the general population. This approach has already proven valuable as the Amish communities of Pennsylvania have aided the discovery of genetic risk factors for complex diseases including the discovery of an *APOC3* mutation, R19X, that is associated with a healthier lipid profile and less coronary artery calcification (98).

A previous study in the Adams County Amish reported lower level of cognitive impairment in individuals ≥65 years old, even accounting for lower levels of former education(99-101). The Adams County Amish have a lower frequency of the ε4 *APOE* risk allele compared to the general population, while the *APOE-4* frequency in the Elkhart, LaGrange, and Holmes Counties is similar to the general population(102). Another study in the Pennsylvania Amish found the *APOE-4* frequency to be similar to the *APOE-4* frequency in the general population, although the Pennsylvania Amish also have a lower prevalence of dementia(103). Therefore, not only is it expected that the Amish genome harbors fewer AD risk genes, the risk is likely to be found in other genes besides *APOE*.

This unique lower *APOE-4* frequency in the Adams County Amish, along with other linkage data, mitochondria and Y-chromosome data, has suggested some degree of genetic heterogeneity between the Adams County and the Holmes, Elkhart, and LaGrange County Amish. However, all four counties combined still represent a very closely related population.

## Previous work

We have a solid relationship established with the Amish communities of Indiana with the help of Dr. Gene Jackson, who has had a long-standing relationship of over 40 years with them. His pioneering efforts opened the door for our efforts examining dementia in the Indiana Amish. He also provided the stimulus to expand our studies to the Amish communities in Ohio. At the onset of this thesis work, about 26% of the Ohio and Indiana Amish population age 65 or older had already been contacted, and nearly 90% of those contacted have agreed to participate in the study. Over 900 DNA samples, more than 125 of which are from individuals diagnosed with AD, had been collected.

We first examined kinship coefficients in this dataset. A kinship coefficient is the probability that two alleles at a randomly chosen locus, one from individual i and the other from individual j, are identical by decent (i.e. came from the same common ancestor). The more related the individuals are, the more alleles they should share, and the higher the kinship coefficient will be between the two individuals. Some examples of kinship coefficients expected for various relationships are in Table 1.2.

**Table 1.2. Expected kinship coefficients for some familial relationships**

| Relationship | kinship coefficient |
|---|---|
| Parent-Offspring | 0.25 |
| Full Siblings | 0.25 |
| Half Siblings | 0.125 |
| First Cousins | 0.0625 |
| Second Cousins | 0.015625 |

We compared the average kinship coefficient between those diagnosed with Alzheimer disease in our study and those that were cognitively normal. We did not include the individuals with other non-Alzheimer cognitive impairments in the calculation. The kinship coefficients were based on known pedigree structure to estimate the expected genetic sharing, not actual genetic data. The average kinship coefficient for all pairs of LOAD individuals was 0.0129 and 0.0116 for all pairs of cognitively normal individuals. To test if this difference was statistically significant, we performed the nonparametric two sample Wilcoxon rank-sum (Mann Whitney) test because the kinship coefficients were not normally distributed. We saw that the difference between cases and controls was significant ($p<1x10^{-5}$); however, we did not take into account any correlations among the pairs. This significant difference suggests that there is a genetic component of Alzheimer disease in this Amish cohort. This difference could also be due to a more common environmental exposure in the Alzheimer patients; however, because the Amish have more homogeneous environmental exposures than the general population, this explanation is less probable. The cases, on average, share more of their genomes, and that shared portion likely harbors risk to Alzheimer disease.

Previous linkage scans in small subsets of the current dataset yielded multiple candidate chromosomal regions for LOAD, but none of the results were striking enough to warrant extensive follow-up in the regions (53;56) (Table 1.3). These previous studies mostly involved microsatellite marker linkage screens with smaller sample sizes and average of marker densities of 11 cM (53) and 7 cM (56). We also have had some changes in affection status since those previous publications. Therefore it is possible for future studies to find other significant loci and to not replicate these results. With the

advancement of genotyping technologies, doing a genome-wide SNP (single nucleotide polymorphism) screen followed by next-generation sequencing in these isolated populations is now a reasonable task. This type of screen should be more comprehensive than the previous microsatellite screens, allowing for more precise locus identification which can then be followed up with sequence analysis.

**Table 1.3 Regions with LOD score >3 in previously published linkage scans in subsets of the current Amish dataset.**

| Publication | Total sample size | Number with AD | Chromosomal region | LOD |
|---|---|---|---|---|
| Ashley-Koch et al 2004 | 24 | 10 | 11p | 3.1 (multipoint) |
| Hahs et al 2006 | 115 | 40 | 4q31 | 3.01 (two-point) |

## Summary

In summary, LOAD is an incredibly complex neurodegenerative disorder affecting many elderly individuals and is only expected to increase in prevalence. Some LOAD risk genes have been identified, but a large portion of the heritability remains to be explained. The Amish communities of Ohio and Indiana provide a means to overcome some of the complexity and heterogeneity that hinders many genetic studies. The many advantages of this isolated population in combination with the advancements in genome technology have been employed for this thesis work and will hopefully help to shed light on the genetic architecture of LOAD.

# CHAPTER II

# QUALITY CONTROL PROCEDURES FOR A GENOME-WIDE STUDY IN AN AMISH POPULATION

## Introduction

A genome-wide association study (GWAS) is one approach for studying a disease with complex etiology like Alzheimer disease. Complex diseases do not show a Mendelian pattern of inheritance. To attempt to predict the inheritance of complex diseases, the hypothesis that common variants explain common diseases emerged (104). GWAS is based on the common disease/common variants hypothesis (104) and involves genotyping 250,000 to more than one million single nucleotide polymorphisms (SNPs) across the genome. The set of SNPs can also be used to conduct a genome-wide linkage scan if family data is available. The broad coverage of the genome eliminates the need to select candidate genes prior to genotyping, and also provides a much denser coverage of the genome than previous genome-wide linkage scans. We have taken this high-throughput approach to study Alzheimer disease in an Amish population (as discussed in Chapters I and III).

With high-throughput methods also come more potential for errors in the dataset because, by necessity, less attention is given to the individual variants. It is impossible to evaluate the quality of each individual genotype, requiring evaluation by descriptive statistics of the SNPs and samples to determine outliers. Combining this large volume of SNP genotypes with the large and complicated family structure of the Amish

produces an extra layer of complexity and the need for careful quality control procedures on both the SNP level and the sample level to produce an accurate dataset to analyze.

Because the focus of a GWAS is on common variation, SNPs with low minor allele frequencies (typically <5%) are removed from analysis. Power to detect an association decreases with lower minor allele frequencies, and spurious associations can arise with low allele frequencies. Unexpected allele and genotype frequencies can also indicate failed genotyping. Therefore, checking SNPs for Hardy-Weinberg equilibrium is a common quality control procedure. In doing so, observed genotype frequencies are compared to expected genotype frequencies. Assumptions of Hardy-Weinberg equilibrium include random mating, a large population size, and no inbreeding. Because our dataset violates those assumptions, we did not check SNPs for Hardy-Weinberg equilibrium in this study. Per SNP genotyping performance can also be assessed by calculating genotyping efficiency, i.e. the percentage of samples for which a genotype could reliably be determined at each SNP. The genotyping efficiency at which a SNP should be deemed 'poor' is debatable as there is always a trade-off between quality and quantity of data available to analyze.

The same trade-off comes into play when determining which samples to eliminate from analysis. Those with low genotyping efficiency (the percentage of called genotypes for each sample) need to be eliminated since the genotypes that are available for those individuals might not be reliable. Besides genotyping errors, DNA handling and/or plating errors could occur prior to genotyping resulting in possible sample mix-ups. One way to detect a sample mix-up is to compare reported gender with the gender

determined by the heterozygosity rate for SNPs on the X chromosome since the heterozygosity rate should be very minimal in males and much higher in females.

Another way to detect a sample mix-up is to examine the genetic ethnicity of each sample. For the Amish, the race/ethnicity should be similar between individuals, so within project sample mix-ups would not be detected but other project samples could be detected if they were accidentally plated. The program Structure provides a plot to visualize clustering of individuals compared to individuals with known racial/ethnic decsent, i.e. samples from the HapMap project (105). Because the Amish population is a founder population of European descent, it is also of particular interest to see if and how they cluster with the HapMap CEU (Eurpean-descent) dataset.

Another important aspect of quality control when studying the Amish is to verify the accuracy of the pedigree relationships. It is the genetic relationships that allow us to perform our studies and therefore the accuracy of which need to be maintained. Aberrant connections in the pedigree would greatly impact linkage analysis, which directly relies on pedigree relationships as it tests for co-segregation of genetic loci and the trait of interest in the pedigrees. Pedigree errors could also distort association results when family relationships are used to correct for the nonindependence of the genotypes. As discussed in Chapter I, the pedigree information is provided by the Anabaptist Geneaology Database (AGDB) (96;97). The accuracy of the pedigree and demographic information from AGDB is outstandingly reliable. However, in rare cases, reported family relationships might not coincide with actual genetic relationships, for instance in the case of an unreported adoption. The large number of available SNPs allows us to compare the reported pedigree relationships with the genetic relationships from calculations of average identical-by-state (IBS) allele

sharing as a proxy for identical-by-descent (IBD) allele sharing. Not only does checking expected IBS to actual IBS serve to verify pedigree relationships, but it can also detect sample swaps or duplicates that might have occurred while plating the DNA.

## Methods

### *Genotyping*

Genotyping was performed on the Affymetrix Genome-Wide SNP Array 6.0, as described in Chapter III. Initial quality control performed by the genotyping laboratory resulted in a dataset with 830 samples and 906,598 SNPs. Two individuals are not in AGDB and another individual could not be connected into the same pedigree with the rest of the individuals. These three individuals were removed before running any additional quality control measures since they would not be useful for analysis. Therefore, 827 individuals and 906,598 SNPs were evaluated using the quality control procedures described below.

### *Sample and SNP genotyping efficiency*

PLATO (PLatform for the Analysis, Translation, and Organization of large-scale data) was used to calculate per-sample and per-SNP genotyping efficiencies. We used a sample genotyping efficiency threshold of 95% and a SNP genotyping efficiency threshold of 98% to remove poor performing samples and SNPs from the dataset. We chose a more liberal sample cut-off since we removed low efficiency samples before removing low efficiency SNPs.

***Minor allele frequency***

Minor allele frequencies were calculated, not taking pedigree relationships into account, using PLATO. Minor allele frequencies were also calculated using the MQLS (Modified Quasi-Likelihood Score) test, which incorporates kinship coefficiencts when calculating allele frequencies to correct for the high degree of relatedness in the dataset. The correlation between the two sets of minor allele frequencies was 0.99. We removed SNPs that had a minor allele frequency <5% calculated by either method.

***Gender***

PLATO was used to calculate the per-sample percentage of heterozygosity for SNPs on the X chromosome. Distributions of the X chromosome heterozygosity percentages were compared between the reported males and the reported females to determine outliers.

***Mendelian errors***

PLATO was used to remove any genotypes that are impossible based on the genotypes at the same SNPs for a parent and a child.

***Race/Ethnicity***

Homogeneity of the ethnicity of all individuals in the study was determined using Structure. Structure is a clustering method that uses allele frequencies to probabilistically assign individuals to populations. Genotypes from individuals of known ethnicity, such as those from HapMap, can be used to guide the clustering. PLATO was used to create a dataset of 1000 randomly chosen markers from our GWAS

dataset. We only used the 124 most 'unrelated' Amish for this analysis. Genotype data from the HapMap CEU (Eurpean-descent), YRI (African), CHB and JPT (Asians) datasets were used as references for the clustering. The CHB and JPT populations cluster together, and were therefore coded as the same population. In a second analysis we also ran Structure using all of the Amish individuals to assess population structure within the Amish. Individuals from Adams County, IN, were coded as a separate population from the rest of our Amish dataset from Elkhart and LaGrange Counties, IN, and Holmes County, OH. We did not include HapMap samples in this second analysis to see better distinction between the Amish communities.

***Pedigree errors***

Graphical representation of relationship errors (GRR) was used to compare reported pedigree relationships with genotype-estimated pedigree relationships. GRR calculates and plots the mean and variance of IBS allele sharing for each pair of individuals (106). GRR categorizes all pairs of individuals with genotypes into full sibling pairs, half sibling pairs, parent-offspring pairs, 'other' relatives pairs, and 'unrelated' pairs. A set of 1000 SNPs was randomly selected using PLATO. GRR was first run with the raw dataset, and then rerun after removing a duplicate sample and other samples that did not pass other previous quality control thresholds.

## Results

### *SNP quality control*

All SNPs with less than 98% genotyping efficiency were removed from analysis. This step eliminated 76,816 SNPs from the dataset.

### *Minor allele frequency*

Employing a 5% minor allele frequency cut-off removed 206,970 SNPs from the dataset. An additional 7,849 SNPs were removed using a 5% MQLS-adjusted minor allele frequency cut-off. Observing the manhattan and quantile-quantile plots displaying the MQLS-derived p-value results before and after the removal of the SNPs with low MQLS-adjusted minor allele frequencies suggests that removing the additional 7,849 SNPs greatly reduced the number of likely false positive association results (Figures 2.1a,b and 2.2a,b).

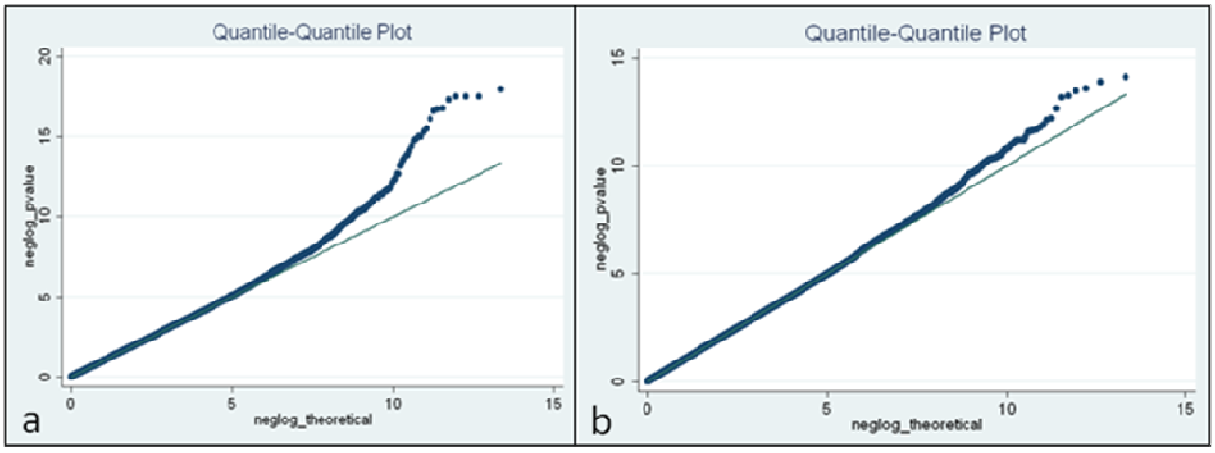



**Figure 2.1a,b. Quantile-quantile plots of MQLS p-values before (a) and after (b) removing additional SNPs with MQLS-adjusted minor allele frequencies <0.05.**

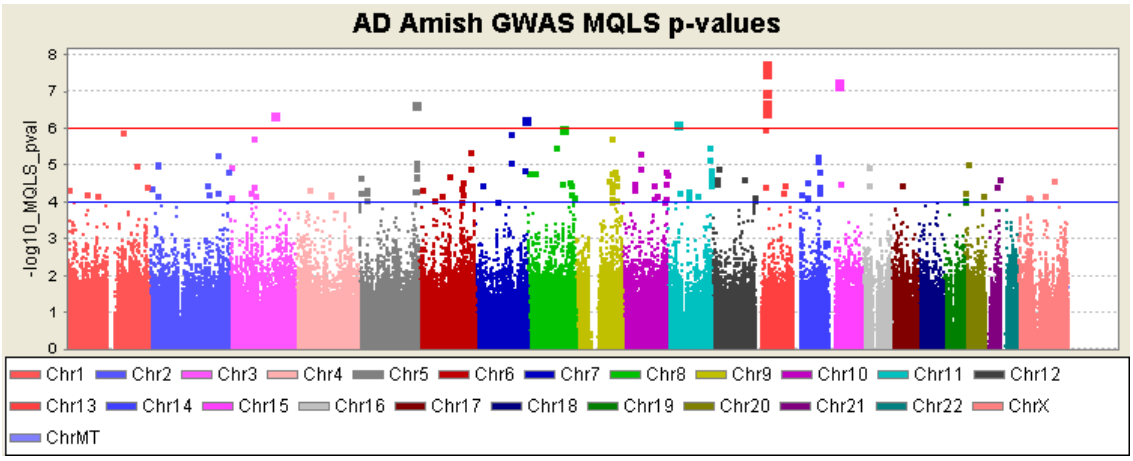



**Figure 2.2a. Manhattan plot of the MQLS results before removing additional SNPs with MQLS-adjusted minor allele frequencies <5%.** This plot shows the MQLS result for each SNP plotted as $-\log_{10}$P value on the y axis. Chromosomal locations are designated on the x axis. 206,970 SNPs with minor allele frequencies <5% based on raw counts were already removed. The red line is at $p=1x10^{-6}$. The blue line is at $p=1x10^{-4}$.

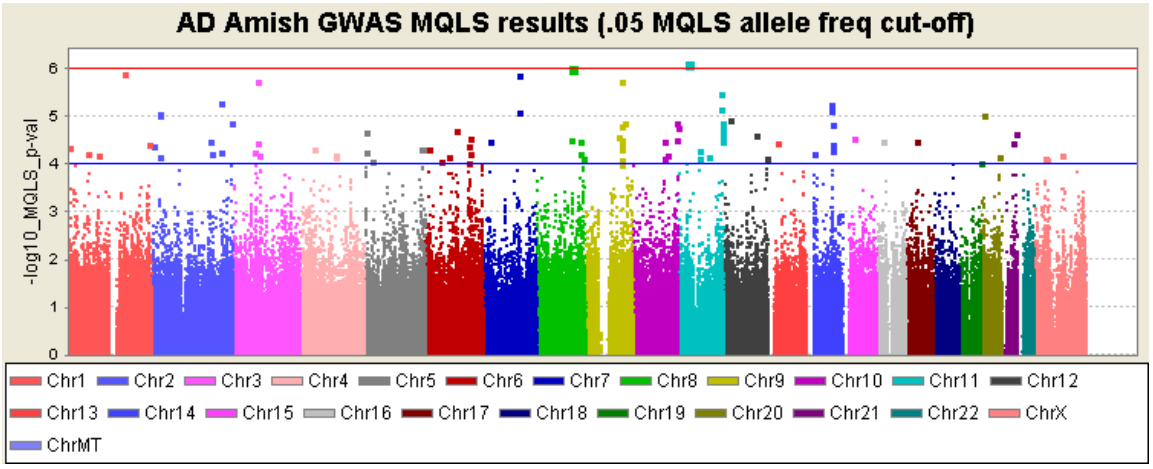



**Figure 2.2b. Manhattan plot of the MQLS results after removing additional SNPs with MQLS-adjusted minor allele frequencies <5%.** This plot shows the MQLS result for each SNP plotted as $-\log_{10}$P value on the y axis. Chromosomal locations are designated on the x axis. 7,849 additional SNPs were removed from the group of SNPs shown in Figure 2.2a before generating this plot. Red line is at $p=1x10^{-6}$. Blue line is at $p=1x10^{-4}$.

***Sample genotyping efficiencies***

Using a 95% per-sample genotyping efficiency threshold, nine samples were removed from analysis.

***Gender***

Two individuals, 1 and 44, were clearly gender errors, most likely due to a sample mix-up or swap. Another explanation could be sex chromosome anomalies such as Kleinfelter syndrome (XXY males) or Turner syndrome (X0 females). However these syndromes are rare and usually lead to infertility, which was not an issue for these individuals. Individual 1 was labeled as a female but the percentage of SNPs on the X chromosome that were heterozygous was only 0.33%. Individual 44 was listed as a male but had 23.92% heterozygous genotypes for SNPs on the X chromosome (see Table 2.1). The other individuals in red in Table 2.1 were not blatant errors, i.e. the females did not have heterozygosity rates low enough to be called male and the males did not have high enough heterozygosity rates to be called female, but were also removed from analysis since they were outliers. The outliers were determined by observing a clear cut-off between 'female' samples 3 and 4 and 'male' samples 31 and 32. We did not observe a correlation between the gender errors and genotyping efficiency or sample quality (as determined by DNA source, date of collection, and evidence of degradation). The genotyping efficiency was at least 98% for all individuals with a gender error.

**Table 2.1. Average percentage of heterozygosity for SNPs on the X chromosome for individuals whose reported and genetic sex are potentially discordant.** The table only displays the females with the lowest X chromosome SNP heterozygosity rates and the males with the highest X chromosome SNP heterozygosity rates. Individuals in red were removed from the dataset. Dummy ID's are displayed to protect the identity of study participants.

| Individual | Sex | % X chromosome heterozygous | Individual | Sex | % X chromosome heterozygous |
|---|---|---|---|---|---|
| **1** | **Female** | **0.33** | 23 | Male | **0.4029** |
| **2** | **Female** | **14.77** | 24 | Male | **0.4058** |
| **3** | **Female** | **14.78** | 25 | Male | **0.4062** |
| 4 | Female | **18.40** | 26 | Male | **0.4132** |
| 5 | Female | **18.68** | 27 | Male | **0.4181** |
| 6 | Female | **18.71** | 28 | Male | **0.4256** |
| 7 | Female | **19.49** | 29 | Male | **0.4292** |
| 8 | Female | **19.56** | 30 | Male | **0.4331** |
| 9 | Female | **19.70** | 31 | Male | **0.4647** |
| 10 | Female | **20.10** | **32** | **Male** | **1.752** |
| 11 | Female | **20.16** | **33** | **Male** | **1.967** |
| 12 | Female | **20.17** | **34** | **Male** | **2.446** |
| 13 | Female | **20.27** | **35** | **Male** | **2.506** |
| 14 | Female | **20.29** | **36** | **Male** | **2.522** |
| 15 | Female | **20.58** | **37** | **Male** | **2.641** |
| 16 | Female | **20.65** | **38** | **Male** | **2.914** |
| 17 | Female | **20.77** | **39** | **Male** | **3.115** |
| 18 | Female | **20.85** | **40** | **Male** | **3.481** |
| 19 | Female | **20.94** | **41** | **Male** | **3.705** |
| 20 | Female | **20.96** | **42** | **Male** | **3.84** |
| 21 | Female | **20.99** | **43** | **Male** | **5.789** |
| 22 | Female | **21.12** | **44** | **Male** | **23.92** |

***Mendelian errors***

PLATO detected 10,023 Mendelian errors in the entire dataset. The highest percentage of genotypes with Mendelian errors for any individual was 0.6%. Because no sample had a substantially high percentage of Mendelian errors, the 10,023 genotypes were removed from the dataset, but no samples were removed.

***Race/Ethnicity***

An examination of the population structure produced the following conclusions: 1) All samples in the dataset are Amish. No samples from another race/ethnicity were accidentally included in the project (Figure 2.3a); 2) The Amish are most similar racially/ethnically to the HapMap CEU samples compared to other Hapmap samples tested (Figure 2.3a); 3) Although relatively quite homogeneous, there is a low level of population substructure within the Amish communities in our dataset, i.e. Adams County is somewhat distinct from Elkhart and LaGrange Counties, IN, and Holmes County, OH (Figure 2.3b). No individuals were removed after examining race/ethnicity.

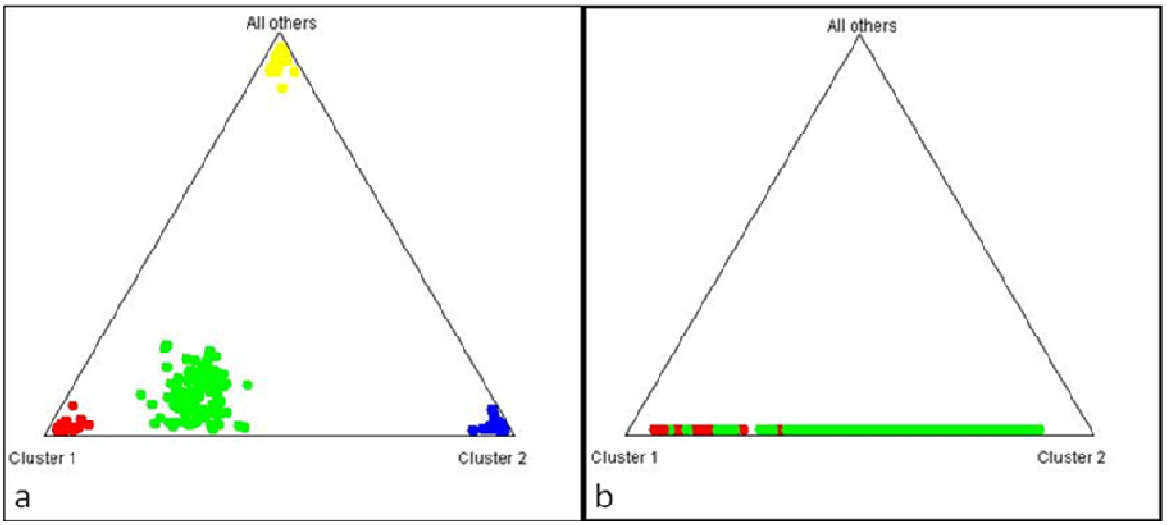



**Figure 2.3a,b. Population structure of the Amish.** a) The 124 most unrelated Amish (red) were used to compare with the HapMap CEU (green), JPT and CHB (blue), and YRI (Yellow) populations. b) Amish individuals from Adams County (red) compared to Elkhart, LaGrange, and Holmes Counties (green).

***Pedigree errors***

When using GRR to cluster by mean and standard deviation of IBS, no outliers were observed for half-sibling pairs, parent-offspring pairs, and unrelated pairs (Figure 2.4b,c,e). However the graphs for full sibling pairs and other relatives pairs both show outliers, including a relative pair with mean IBS=2.0 (Figure 2.4a,d). The two samples with IBS=2.0 either belong to a pair of identical twins or one of the samples was duplicated. We rejected the first scenario because the two individuals have different parents and are 11 years apart in age. Examination of other outliers from the sibling pairs and other relatives pairs clusters revealed that one of the individuals was causing several other outliers. Because we could determine which individual was the duplicate, we removed that sample and not both.

After removing the duplicated sample as well as other samples that did not meet previous quality thresholds, a few outliers remained for the sibling pairs and other relatives pairs (Figure 2.5 a,b). The mean IBS for all full siblings was 1.68. Thresholds were determined empirically with the observed clustering to take into account the uniquely higher relatedness in this dataset. We were most concerned about the sibling pairs that had an average IBS <1.5. Individuals in more than one erroneous pair were determined to be the problematic individuals and were eliminated from the dataset. Individual 45, part of the sibling pair with a mean IBS of 1.42, was also part of the one outlier pair (mean IBS=1.69) in the other relatives group, which had an overall IBS mean of 1.55 (Tables 2.2 and 2.3). The remaining sibling pairs with mean IBS <1.5 all contained individual 48, who was therefore removed. We also decided to eliminate individual 52 since that individual was part of three sibling pairs that did not cluster tightly with the rest of the sibling pairs (Table 2.2).

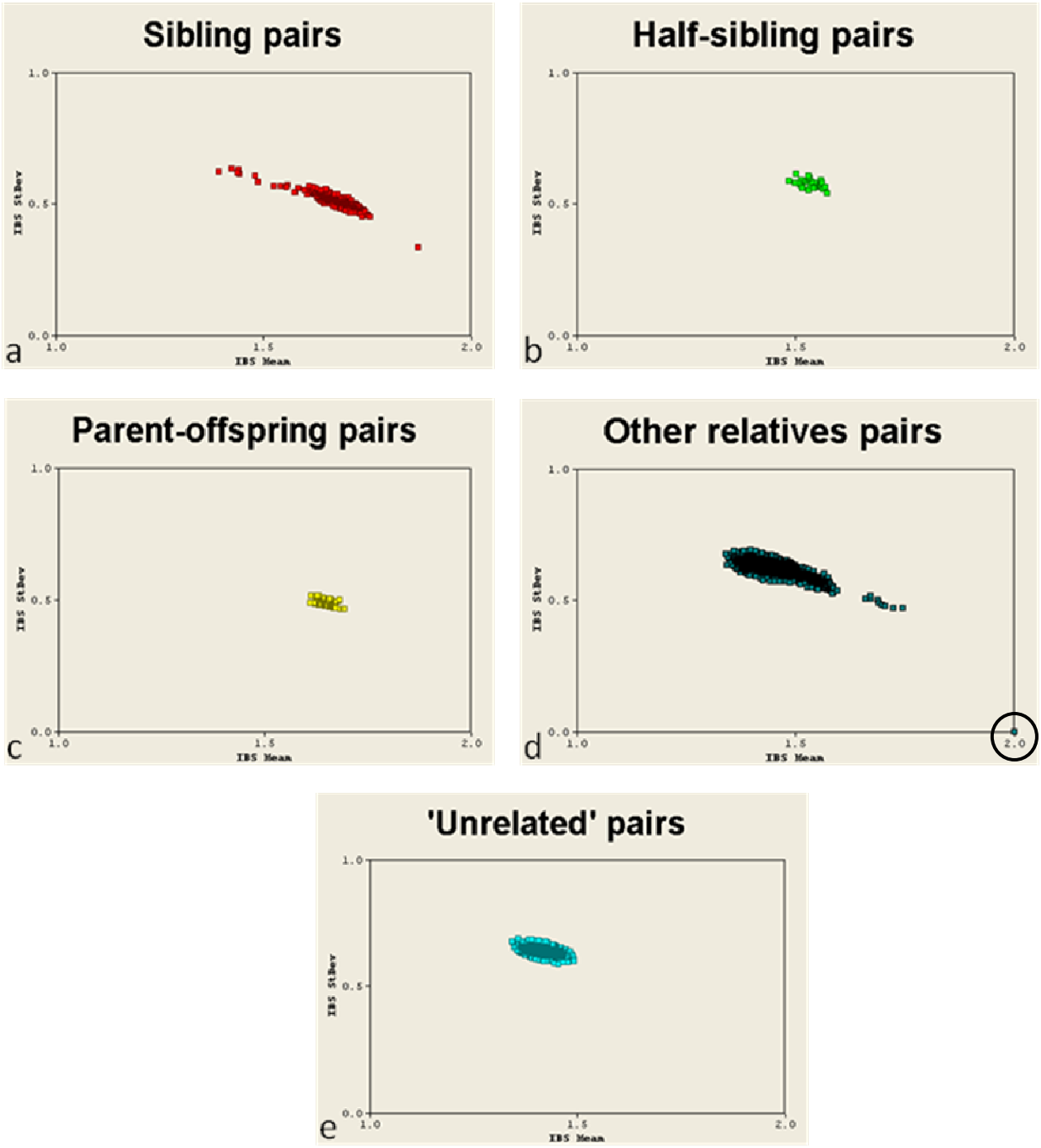



**Figure 2.4a-e. Output from Graphical Representation of Relationships using raw data.** X axis is IBS mean, Y axis is IBS standard deviation. Means and standard deviations are based on 1000 randomly chosen SNPs. Each point represents one pair of individuals. Pair of individuals with mean IBS=2 is circled.

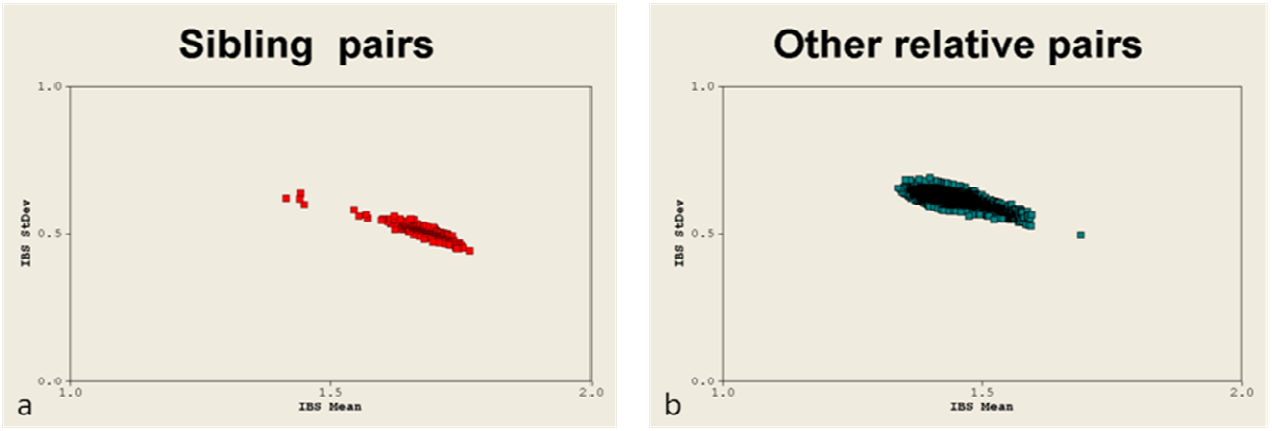



**Figure 2.5a,b. Output from Graphical Representation of Relationships after removing duplicate sample and samples not meeting other quality control thresholds.** X axis is IBS mean, Y axis is IBS standard deviation. Means and standard deviations are based on 1000 randomly chosen SNPs. Each point represents one pair of individuals.

**Table 2.2. Lowest mean IBS for sibling pairs.** Mean IBS was calculated from 1000 randomly chosen SNPs from the GWAS. Different colors are used to distinguish unique ID's for individuals who were removed from the dataset. 'Markers' refers to the number of markers compared between Person1 and Person2. Note that individual 45 (highlighted in yellow) is also in Table 2.3.

| Person1 | Person2 | Markers | **IBS Mean** | Mean StDev | Relationship |
|---|---|---|---|---|---|
| 45 | 46 | 999 | **1.42** | 0.62 | SIB-SIB |
| 47 | 48 | 999 | **1.44** | 0.62 | SIB-SIB |
| 49 | 48 | 999 | **1.44** | 0.64 | SIB-SIB |
| 48 | 49 | 972 | **1.45** | 0.6 | SIB-SIB |
| 50 | 51 | 1000 | **1.55** | 0.58 | SIB-SIB |
| 52 | 53 | 998 | **1.55** | 0.58 | SIB-SIB |
| 52 | 54 | 995 | **1.56** | 0.56 | SIB-SIB |
| 55 | 55 | 994 | **1.57** | 0.56 | SIB-SIB |
| 52 | 57 | 994 | **1.57** | 0.55 | SIB-SIB |
| 58 | 59 | 998 | **1.6** | 0.54 | SIB-SIB |
| 60 | 61 | 996 | **1.6** | 0.55 | SIB-SIB |
| 62 | 63 | 976 | **1.6** | 0.55 | SIB-SIB |
| 64 | 65 | 997 | **1.61** | 0.55 | SIB-SIB |
| 66 | 67 | 984 | **1.61** | 0.55 | SIB-SIB |
| 68 | 59 | 1000 | **1.61** | 0.54 | SIB-SIB |
| 69 | 70 | 982 | **1.61** | 0.54 | SIB-SIB |
| 71 | 72 | 996 | **1.61** | 0.54 | SIB-SIB |
| 73 | 74 | 993 | **1.62** | 0.56 | SIB-SIB |

**Table 2.3. Highest mean IBS for other relatives pairs.** Mean IBS was calculated from 1000 randomly chosen SNPs from the GWAS. Note that individual 45 (highlighted in yellow) is also in Table 2.2 and was removed from the dataset.

| Person1 | Person2 | Markers | **IBS Mean** | IBS StDev | Relationship |
|---|---|---|---|---|---|
| 45 | 75 | 999 | **1.69** | 0.5 | Related |
| 76 | 77 | 994 | **1.6** | 0.56 | Related |
| 78 | 79 | 979 | **1.59** | 0.53 | Related |
| 80 | 81 | 997 | **1.59** | 0.55 | Related |
| 82 | 68 | 997 | **1.59** | 0.53 | Related |
| 83 | 84 | 1000 | **1.59** | 0.56 | Related |
| 85 | 86 | 991 | **1.59** | 0.56 | Related |
| 87 | 88 | 997 | **1.59** | 0.58 | Related |
| 89 | 90 | 975 | **1.59** | 0.55 | Related |
| 91 | 58 | 995 | **1.58** | 0.53 | Related |
| 92 | 93 | 994 | **1.58** | 0.55 | Related |
| 94 | 93 | 989 | **1.58** | 0.55 | Related |
| 95 | 96 | 992 | **1.58** | 0.54 | Related |
| 97 | 98 | 996 | **1.58** | 0.57 | Related |
| 99 | 100 | 995 | **1.58** | 0.56 | Related |
| 101 | 102 | 999 | **1.58** | 0.57 | Related |
| 103 | 93 | 983 | **1.58** | 0.57 | Related |
| 104 | 105 | 998 | **1.58** | 0.58 | Related |
| 106 | 101 | 998 | **1.58** | 0.57 | Related |
| 107 | 93 | 999 | **1.58** | 0.58 | Related |
| 106 | 103 | 983 | **1.57** | 0.58 | Related |
| 108 | 109 | 984 | **1.57** | 0.56 | Related |

***Result of all QC***

After all quality control measures were taken, 291,635 SNPs were removed, leaving 614,963 SNPs for analysis (Table 2.4). An additional 10,023 individual genotypes were removed due to Mendelian errors.

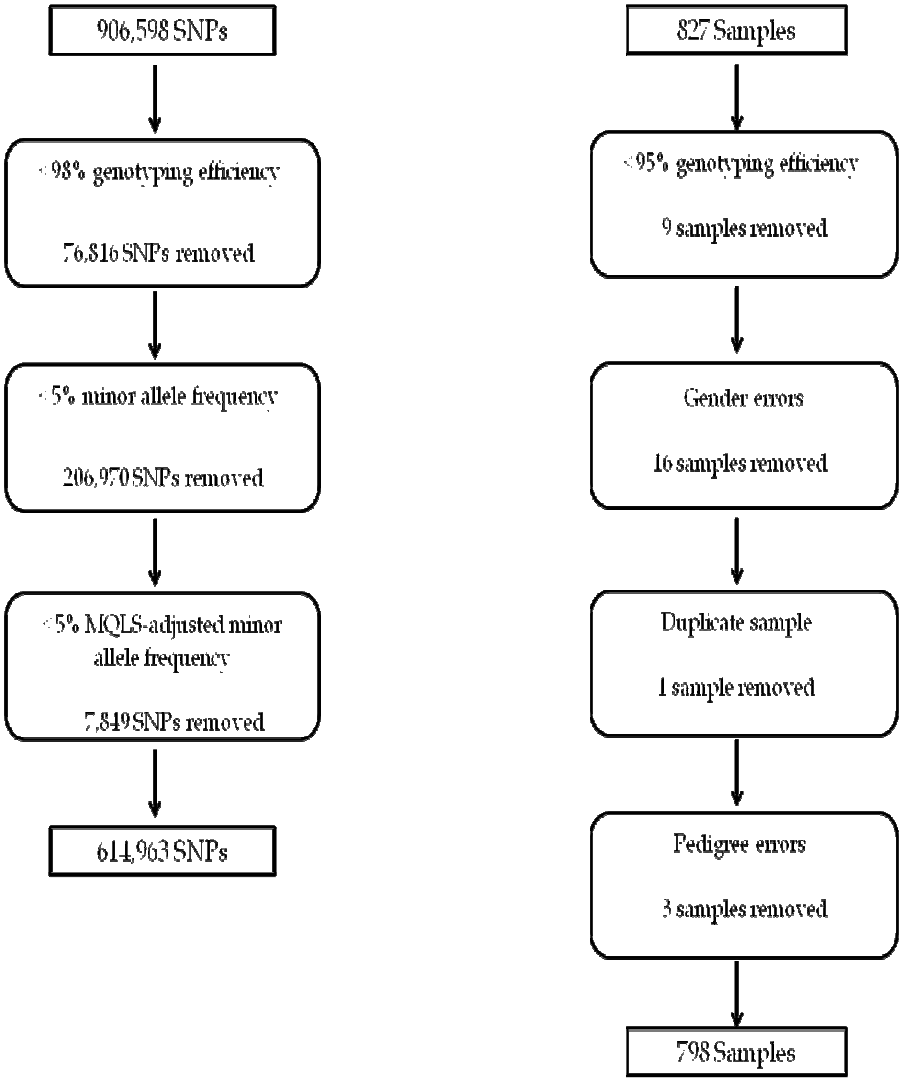



**Figure 2.6 Flowchart of SNP and sample quality control procedures.**

## Discussion

Following careful quality control procedures is extremely important, especially when handling a dataset with genotypes generated with high-throughput methods where overall quality is generally good but many errors could go undetected because of the high-throughput nature. A clean dataset will reduce the likelihood of spurious associations arising in the results and will increase the likelihood of detecting a true genetic association. Not only is a clean dataset important for the immediate analyses, i.e. genome-wide linkage and association analyses, but also for future studies that are based on the initial analyses. Ensuring the quality of the data at the beginning of the study provides for more efficient and effective future studies. In this study we started with 906,598 SNPs and 827 samples in the raw dataset.

We performed quality control procedures typical of most GWAS including eliminating samples and SNPs with low genotyping efficiency, eliminating SNPs with low minor allele frequency, and samples that were gender errors. Eliminating samples and SNPs with low genotyping efficiency removes poor performing and therefore potentially unreliable genotypes from analysis. SNPs with low minor allele frequencies can also indicate poor genotyping performance and are not useful since power to detect association is diminished. Unlike the usual case-control GWAS, we also incorporated kinships into our calculation of minor allele frequency to eliminate SNPs with a corrected minor allele frequency $<0.05$. Because of the family structure of our dataset, we removed genotypes, but no individuals, that appeared to be Mendelian errors. Checking gender errors was particularly important for detecting any sample handling mix-ups, which would result in genotypes matched to the wrong individual, and

therefore potentially the wrong phenotype and pedigree placement. A closer examination of the raw probe intensity data for SNPs the X chromosome could help explain some of the heterozygosity rates that were outliers. For instance, it would be interesting to see if there are specific areas of the chromosome affecting the heterozygosity rates or if the heterozygous SNPs are randomly distributed along the X chromosome. We also assessed the family structure to remove individuals whose allele sharing was not typical of the specified family relationship since improper specification of relatedness could dramatically change our linkage results. This procedure not only identified pedigree errors but also identified a sample duplicate which would not have been detected by any other quality control procedures. A closer examination of the allele sharing could reveal the correct pedigree placement for those individuals who were eliminated. Continuing to ascertain and genotype more individuals, which fills in more pedigree information, could help with that process.

As we expected, when comparing the Amish to other populations by clustering based on allele frequencies, the Amish are most similar to an outbred European-descent dataset compared to Yoruban, Japanese, and Chinese datasets. Although clustered closest to the European descent dataset, the Amish formed a completely distinct cluster, which is a product of the isolation of this population and the increased relatedness even between the most unrelated individuals. Even within the Amish we saw some population substructure between the Adams County community and Elkhart, LaGrange, and Holmes County communities, which is indicative of their history and cultural differences. Some settlers of the Elkhart County and LaGrange County community originally came from Holmes County Ohio, resulting in more relationships between these communities compared to the Adams County community which

remained distinct (95). The Adams County community has historically followed more Old Order ways and has been stricter with marriage choices. Overall, though, all Amish communities are much more historically and culturally similar to each other than the general population.

In total, after performing all of our quality checks, we removed 291,635 SNPs and 26 samples leaving 614,963 SNPs and 798 samples. These quality control procedures have provided a reliable dataset for genome-wide linkage and association analyses that are presented in Chapter III, and a follow-up sequence study presented in Chapter IV.

# CHAPTER III

## GENOME-WIDE LINKAGE AND ASSOCIATION STUDY FOR ALZHEIMER DISEASE IN AN AMISH POPULATION[1]

### Introduction

Late-onset Alzheimer disease (LOAD) is a neurodegenerative disorder causing the majority of dementia cases in the elderly. A complex combination of genetic and environmental components likely determines susceptibility to LOAD (19). The *APOE* E4 allele is a well-established genetic risk factor for LOAD. Additional risk genes have been difficult to detect and replicate until recent successes using large consortia-derived genome-wide association study (GWAS) datasets, which have added *CR1*, *CLU*, *PICALM*, *BIN1*, *EPHA1*, *MS4A*, *CD33*, *CD2AP*, and *ABCA7* to the list of confirmed LOAD susceptibility genes, each with modest effect (78-82).

Despite these recent successes the majority of the genetic risk for LOAD remains unknown. The remaining genetic risk may in part lie in additional loci with small effects at the population level, making most datasets underpowered. The use of a genetically isolated founder population, such as the Amish, represents an alternative to the use of

[1] Adapted from: Anna C. Cummings, Lan Jiang, Digna R. Velez Edwards, Jacob L. McCauley, Renee Laux, Lynne L. McFarland, Denise Fuzzell, Clare Knebusch, Laura Caywood, Lori Reinhart-Mercer, Laura Nations, John R. Gilbert, Ioanna Konidari, Michael Tramontana, Michael L. Cuccaro, William K. Scott, Margaret A. Pericak-Vance, Jonathan L. Haines. Genome-Wide Association and Linkage Study in the Amish Detects a novel Candidate Late-Onset Alzheimer Disease Gene. *Annals of Human Genetics*. 2012 Sep; 76(5):342-51.

large population based consortia-derived datasets in the search for genetic risk factors. In the case of a founder population, the number of disease variants is hypothesized to be fewer, thereby decreasing heterogeneity and increasing power.

We have taken this approach to discover at least one novel LOAD risk gene by studying the Amish communities of Holmes County, Ohio, and Adams, Elkhart and LaGrange Counties, Indiana (56;107). These communities are collectively part of a genetically isolated founder population originating from two waves of immigration of Swiss Anabaptists into the U.S in the 1700's and 1800's. The first wave of immigration brought the Anabaptists to Pennsylvania. In the early 1800's some of these immigrants moved to Holmes County, OH (94), while a second wave of immigration from Europe established more Amish communities in Ohio (including Wayne County but not Holmes County) and Indiana (including Adams County) (95). Starting in 1841, the Elkhart and LaGrange Counties Amish community was founded by Amish families primarily from Somerset County, PA, and from Holmes and Wayne Counties, OH, who were seeking new farmland to settle(93). The Amish marry within their faith, limiting the amount of genetic variation introduced to the population. Not only are the Amish more genetically homogeneous, but because of their strict lifestyle, environmental exposures are also more homogeneous. The Amish have large families and a well-preserved comprehensive family history that can be queried via the Anabaptist Genealogy Database (AGDB) (96;97), making the Amish a valuable resource for genetic studies.

Our current study undertook a genome-wide approach, in a population isolate, using complementary linkage and association analyses to further elucidate the complex genetic architecture of LOAD. We utilized linkage analysis to look for sharing of genomic regions among affected individuals, while also using association analysis to

look for differences in allele frequencies between affecteds and unaffecteds. We previously performed a genome-wide linkage study using microsatellites genotyped in only a small subset of the individuals included in this study(56). Here we use a much larger dataset with a much denser panel of markers using a genome-wide SNP chip. The results indicate that several novel regions likely harbor LOAD genes in the Amish, underscoring the genetic heterogeneity of this phenotype.

## Methods

### *Subjects*

Methods for ascertainment were reviewed and approved by the individual Institutional Review Boards of the respective institutions. Participants were identified from published community directories, referral from other community members or due to close relationship with other participants, as previously described (108). Informed consent was obtained from participants recruited from the Amish communities in Elkhart, LaGrange, and surrounding Indiana counties, and Holmes and surrounding Ohio counties with which we have had established working relationships for over 10 years.

### *Clinical Data*

For individuals who agreed to participate, demographic, family, and environmental information was collected, informed consent was obtained, and both a functional assessment and the Modified-Mini-Mental State Exam (3MS) were administered (109;110). Those scoring ≥ 87 on the 3MS were considered cognitively

normal and were considered unaffected in our study. Those scoring <87 were reexamined with further tests from the CERAD neuropsychological battery (111). Depression was also evaluated using the geriatric depression scale (GDS). Diagnoses for possible and probable AD were made according to the NINCDS-ADRDA criteria (17). A yearly consensus case conference was held to confirm all diagnoses.

***Genotyping***

SNPs for *APOE* were genotyped for 823 individuals (127 with LOAD). To identify the six *APOE* genotypes determined by the *APOE *E2*, **E3* and **E4* alleles, two single nucleotide polymorphisms (SNPs) were assayed using the TaqMan method [Applied Biosystems Inc. (ABI), Foster City, CA, USA]. SNP-specific primers and probes were designed by ABI (TaqMan genotyping assays) and assays were performed according to the manufacturer's instructions in 5 µl total volumes in 384-well plates. The polymorphisms distinguish the **E2* allele from the **E3* and **E4* alleles at amino acid position 158 (NCBI *rs7412*) and the **E4* allele from the **E2* and **E3* alleles at amino acid position 112 (NCBI *rs429358*).

Genome-wide genotyping was performed on 830 DNA samples using the Affymetrix 6.0 GeneChip ® Human Mapping 1 million array set (Affymetrix®, Inc Santa Clara, CA). DNA for this project was allocated by the respective DNA banks at both the Hussman Institute of Human Genomics (HIHG) at the University of Miami and the Center for Human Genetics Research (CHGR) at Vanderbilt University. Genomic DNA was quantitated via the ND-8000 spectrophotometer and DNA quality was evaluated via gel electrophoresis. The genomic DNA (250 ng/5ul) samples were processed according to standard Affymetrix procedures for processing of the Affymetrix 6.0

GeneChip assay. The arrays were then scanned using the GeneChip Scanner 3000 7G operated by the Affymetrix® GeneChip® Command Console® (AGCC) software. The data were processed for genotype calling using the Affymetrix® Power Tools (APT) software using the birdseed calling algorithm version 2.0 (Affymetrix®, Inc Santa Clara, CA) (112).

We applied a number of quality control (QC) procedures to both samples and SNPs to ensure the accuracy of our genotype data prior to linkage and association analyses. Specific sample QC included: 1) Each individual DNA sample was examined via agarose to ensure that the sample was of high quality prior to inclusion on the array; 2) CEPH samples were placed across multiple arrays to ensure reproducibility of results across the arrays; 3) Samples with call rates < 95% were re-examined individually to ensure quality of genotypes. 4) Ultimately if the sample call rate remained below 95% after further evaluation, attempts were made to rerun the array with a new DNA sample. If the sample still failed, it was dropped. Nine samples were dropped due to low genotyping efficiency. Three samples were excluded because they did not connect into a pedigree with the rest of the samples, and therefore, relationships of those individuals could not be accounted for. Sixteen samples with questionable gender based on X chromosome heterozygosity rates were eliminated. Three samples appearing to be aberrantly connected in the pedigree based on the genotype data were also excluded.

Specific SNP QC included: 1) Dropping 76,816 SNPs with call rates <98%. 2) Dropping 206,970 SNPs with minor allele frequencies (MAF) ≤0.05. We additionally excluded 7,849 SNPs with a MAF less than 0.05 after adjusting for pedigree relationships using MQLS (see below). Due to the relatedness in this dataset we did not check SNPs for Hardy-Weinberg equilibrium. Following this extensive quality control, 798 samples

(109 with LOAD, Table 3.1) and 614,963 SNPs were analyzed. Because APOE genotyping and QC were performed separately from genome-wide genotyping and QC, the sample sizes are different and the datasets are mostly, but not completely, overlapping. All 798 samples belong to one 4998-member pedigree with many consanguineous loops. The AGDB provided the pedigree information using an "all common paths" database query with all genotyped individuals (97).

**Table 3.1. Genome-wide dataset.** Ages of exam and onset averages and standard deviations were calculated for the 798 samples—Late-onset AD (LOAD) samples, cognitively normal (unaffected) samples, and unclear or unknown samples—which passed QC for genome-wide genotyping.

| | Males | Females | Total | Average Age of Exam (Standard Deviation) | Average Age of Onset (Standard Deviation) |
|---|---|---|---|---|---|
| LOAD Affected | 43 | 66 | 109 | 83 (7.57) | 79 (6.68) |
| Cognitively Normal | 192 | 258 | 450 | 78 (7.67) | - |
| Unclear or Unknown | 117 | 122 | 239 | 74 (15.52) | - |

## ***Statistical Analysis***

### *Association analysis*

We used the Modified Quasi-Likelihood Score (MQLS) test (software version 1.2) to correct for pedigree relationships (113). MQLS is analogous to a chi-square test, the most common approach for case-control data analysis with a binary trait, but MQLS incorporates kinship coefficients to correct for correlated genotypes of all the pedigree

relationships. This test allows all samples to be included without dividing the pedigree. The MQLS test cannot be applied to X chromosome data, which were, therefore, eliminated from analysis. Because we previously found that Adams County has a lower APOE-4 allele frequency than the general population (114), we did a stratified association analysis for APOE analyzing Adams County separately from the combined Elkhart, LaGrange, and Holmes Counties. Using the same stratification, we also re-analyzed our most significant SNPs from the GWAS analysis. To test the validity of the MQLS test in our pedigree, we performed simulation studies using this same pedigree structure to assess the type 1 error rate using MQLS for association. Type 1 error rates were not inflated (unpublished data).

*Linkage analysis*

Because of the large size and substantial consanguinity of the pedigree, we used PedCut (115) to find an optimal set of sub-pedigrees including the maximal number of subjects of interest within a bit-size limit (24 in this study) conducive to linkage analysis. This procedure resulted in 34 sub-pedigrees for analysis with an average of 7 genotyped individuals (3 genotyped affected) per sub-pedigree. Parametric heterogeneity two-point LOD (HLOD) scores were computed assuming affecteds-only autosomal dominant and recessive models using Merlin (116). Because the underlying genetic model is unknown, we tested both dominant and recessive models to maximize our ability to find a disease locus. A disease allele frequency of 1% was used for both the dominant and recessive models. For the dominant model penetrances of 0 for no disease allele and 0.0001 for one or two copies of the disease allele, and under the recessive model penetrances of 0 for zero or one disease allele and 0.0001 for two disease alleles were

used. SNPs on the X chromosome were analyzed using MINX (Merlin in X). Regions showing evidence for linkage, i.e. containing at least one two-point HLOD ≥ 3.0, were followed up with parametric multipoint linkage analysis (also using Merlin). For the multipoint analyses, SNPs were pruned for linkage disequilibrium (LD) in each region so that all pair-wise $r^2$ values were < 0.16 between all SNPs (117). The LD from the HapMap CEU samples (parents only) were used for pruning. Because the HapMap CEU samples may not be an exact representation of LD in our Amish population, we also tested pruning using the data from this Amish dataset, but linkage results did not change using this approach (data not shown). Because linkage analyses can be biased when breaking larger pedigrees into a series of smaller ones (118;119), we performed simulation studies assuming no linkage (e.g. null distribution) and using the same large pedigree structure and the same pedigree splitting method. We determined empirical thresholds for significance in our linkage studies to maintain a nominal type I error rate. We found that after 1000 replications of multipoint linkage for regions the size of the average size in this study, only 2.5% of the multipoint linkage scans generated a maximum HLOD >3.0 (unpublished data).

All computations were performed using either the Center for Human Genetics Research computational cluster or the Advanced Computing Center for Research and Education (ACCRE) cluster at Vanderbilt University.

*Evaluation of the MMSE and Word List Learning*

We examined a portion of the outcome of the CERAD neuropsychological battery of tests to determine if there were any specific correlations with significant genetic profiles in this Amish population. In a subset of 69 LOAD affected individuals,

we examined scores for the mini-mental state exam (MMSE) and the Word List Learning with delayed recall and recognition procedures portion of the CERAD battery of tests. We chose to evaluate the Word List Learning portion of the battery because, of all the tests in the battery, word list recall was found to distinguish AD cases from controls (120), and Word List Learning scores have been shown to be very heritable (92). The MMSE evaluates overall cognition by asking basic questions and requesting simple writing, drawing, and memory tasks. We chose to also include MMSE scores in our analyses so that we could adjust for overall cognitive abilities when specifically examining Word List Learning measures. The Word List Learning with delayed recall and recognition tests verbal episodic memory. During the Word List Learning and delayed recall and recognition tests, subjects first are given three tries to remember a list of ten words (word list memory trials 1 through 3) that is shown to them. After performing another non-verbal task, they are asked to recall the words again for a delayed recall score. A 'savings' score is calculated by dividing the delayed recall score by the word list memory trial 3 score. Then the same words interspersed with 10 additional words are shown to test if the subjects can recognize if the words were (recognition-yes) or were not (recognition-no) in the original list. Raw scores for the MMSE, the three word list memory trials, delayed recall, savings, recognition-yes, and recognition-no were converted to Z scores appropriate for age, gender, and years of education.

We assigned each of the 69 individuals to one of three LOAD risk groups based on *APOE* genotype. Individuals with the *APOE* 4/4 or 3/4 genotype were labeled high risk, individuals with the *APOE* 2/4 or 3/3 genotype were considered normal risk, and individuals with the *APOE* 2/3 genotype were assigned to the low risk group. There

were no individuals with the *APOE* 2/2 genotype. Because the Z scores were not normally distributed (Appendix C), we compared mean Z scores for each of the test results using the nonparametric Kruskal-Wallis test. Tests that resulted in a p-value ≤ 0.05 were further tested with the two-sample Wilcoxon rank-sum (Mann-Whitney) test to determine between which of the three *APOE* groups the difference in mean Z score was significantly different. We performed an analysis of covariance (ANCOVA) to test each of the Word List Learning Z scores while adjusting for MMSE Z scores to determine if any of the differences between in Word List Learning Z scores were simply a result of overall cognitive decline. Any results with $p<0.05$ were followed up with pairwise ANCOVA to adjust for MMSE Z scores while comparing each of the risk groups.

We also evaluated the correlation between the per-family lod scores of our most significant region of linkage on chromosome 2 and the neuropsychological test results. The per-family lod scores are the lod scores generated by each sub-pedigree in the linkage analyses. Both the dominant and recessive models generated high linkage peaks at 2p12, so we tested both the peak per-family lod scores under the dominant model and the peak per-family lod scores under the recessive model. The nonparametric Spearman correlation was used to test for correlations between the dominant or recessive per-family lod scores and each of the Z scores. We repeated the correlations for the Word List Memory, recall, savings, and recognition scores while adjusting for MMSE as a covariate.

All analyses were performed using STATA.

## Results

***APOE***

We found that LOAD was significantly associated with *APOE* (MQLS P=9.0x10-6) in our Amish population except for the Adams County, Indiana, community (MQLS P=0.55).The E4 frequency, adjusted for pedigree relationships, in LOAD individuals in Elkhart, LaGrange, and Holmes Counties was 0.18 for affected individuals compared to 0.11 for unaffected individuals (Table 3.2). This compares to an E4 allele frequency of 0.38 in Caucasian AD individuals (0.14 for controls) (alzgene.org). We also saw a progressively younger average age of onset with each additional copy of the E4 allele (Table 3.3), consistent with other populations. We did not see evidence for linkage with *APOE* in our sub-pedigrees (dominant HLOD=0.50, recessive HLOD=0.29).

**Table 3.2. MQLS-corrected *APOE* allele frequencies.** *APOE* allele frequencies of Late-onset AD (LOAD) affected individuals versus cognitively normal individuals (unaffecteds) were calculated using MQLS to correct for pedigree relationships. Frequencies were calculated in the Adams County individuals separately from Elkhart, LaGrange, and Holmes Counties.

| ***APOE* allele frequencies** | | | |
|---|---|---|---|
| **Elkhart, LaGrange, and Holmes Counties** | | | |
| | ***E*2** | ***E*3** | ***E*4** |
| LOAD Affected | 0.07 | 0.75 | 0.18 |
| Cognitively Normal | 0.08 | 0.82 | 0.11 |
| **Adams County** | | | |
| LOAD Affected | 0.00 | 0.94 | 0.06 |
| Cognitively Normal | 0.04 | 0.88 | 0.08 |

**Table 3.3. Age of onset and number of affected versus unaffected individuals by APOE genotype.** Average ages of onset and standard deviations by *APOE* genotype and number LOAD affected and unaffected by *APOE* genotype

| | ***APOE* Genotype** | | | | |
|---|---|---|---|---|---|
| | **4/4** | **¾** | **2/4** | **3/3** | **2/3** |
| Average Age of Onset (stdev) | 71 (7.59) | 76 (7.94) | 74 (3.54) | 80 (6.70) | 84 (7.42) |
| Number LOAD Affected | 10 | 34 | 2 | 69 | 12 |
| Number Cognitively normal | 6 | 90 | 6 | 308 | 35 |

***Genome-wide association***

In the GWAS, the most significant MQLS P-value (7.92x10-7), which did not surpass a Bonferroni-corrected genome-wide significance threshold of 8.13x10-8, was at rs12361953 on chromosome 11 in *LUZP2* (leucine zipper protein 2) (Table 3.4, Figure 3.1). The pedigree-adjusted minor allele frequency was 0.26 for affected individuals versus 0.15 for unaffected individuals. Fourteen additional SNPs had p-values <1.0x10-5 (Table 3.4). According to our simulation analyses, we have >80% power to detect a p-value ≤ 0.005 under an additive model with an odds ratio of 2.0 (data not shown). After stratifying, each of the fifteen most significant SNPs had a more significant p-value in the non-Adams County dataset. Although some of the SNPs have very different minor allele frequencies in the two strata, the less significant p-values for the Adams County dataset can be explained mostly by the lack of power in that stratum (9 LOAD affected). All SNPs showed the same direction of effect in the two strata except for rs472926, rs12361953, and rs472926 (Appendix A). These association results did not fall within a megabase of any of the other 9 previously verified LOAD genes (*CR1*, *CLU*, *PICALM*, *BIN1*, *EPHA1*, *MS4A*, *CD33*, *CD2AP*, and *ABCA7*). However, four SNPs (rs10792820,

rs11234505, rs10501608, and rs7131120) in *PICALM* generated nominally significant p-values ($P<0.05$). Rs11234505 is only ~3.0 kb from rs561655, the most significant SNP published by Naj et al(81), and rs10501608 is only ~10.5 kb from rs541458 the most significant SNP published by Harold et al and Lambert et al(78;79). We also have a nominally significant SNP, rs6591625, in the *MS4A10* gene. The SNP is ~0.5 Mb from rs4938933, the most significant SNP published by Naj et al(81).

**Table 3.4. Most significant genome-wide association results.** The fifteen most significant genome-wide association results calculated using MQLS. Minor allele frequencies (MAF) are MQLS-corrected for pedigree relationships. A gene is only listed if the SNP falls within specified gene. Megabase pair (Mpb) positions are based on NCBI Build 36.

| Chr | SNP | Position (Mbp) | Minor Allele | Affected MAF | Unaffected MAF | MQLS *P*-value | Gene |
|---|---|---|---|---|---|---|---|
| 1 | rs4145462 | 165.99 | T | 0.10 | 0.05 | 1.22x10-06 | *MPZL1* |
| 2 | rs41458646 | 23.09 | G | 0.27 | 0.17 | 8.44x10-06 | - |
| 2 | rs41476545 | 23.09 | G | 0.27 | 0.17 | 9.02x10-06 | - |
| 2 | rs6738181 | 204.84 | A | 0.35 | 0.19 | 4.97x10-06 | - |
| 3 | rs7638995 | 69.26 | A | 0.20 | 0.11 | 1.82x10-06 | - |
| 7 | rs679974 | 105.06 | C | 0.18 | 0.08 | 8.67x10-06 | *ATXN7L1* |
| 7 | rs11983798 | 105.07 | T | 0.17 | 0.08 | 1.49x10-06 | *ATXN7L1* |
| 8 | rs6468852 | 104.05 | G | 0.24 | 0.13 | 1.06x10-06 | - |
| 9 | rs9969729 | 107.67 | A | 0.14 | 0.08 | 1.94x10-06 | - |
| 11 | rs12361953 | 24.57 | C | 0.26 | 0.15 | 7.92x10-07 | *LUZP2* |
| 11 | rs472926 | 125.41 | C | 0.25 | 0.15 | 3.28x10-06 | *CDON* |
| 11 | rs4937314 | 127.69 | C | 0.27 | 0.17 | 7.00x10-06 | - |
| 14 | rs11848070 | 70.58 | C | 0.38 | 0.25 | 5.64x10-06 | *PCNX* |
| 14 | rs17767225 | 70.74 | T | 0.32 | 0.2 | 7.88x10-06 | - |
| 20 | rs6085820 | 6.93 | A | 0.17 | 0.09 | 9.31x10-06 | - |

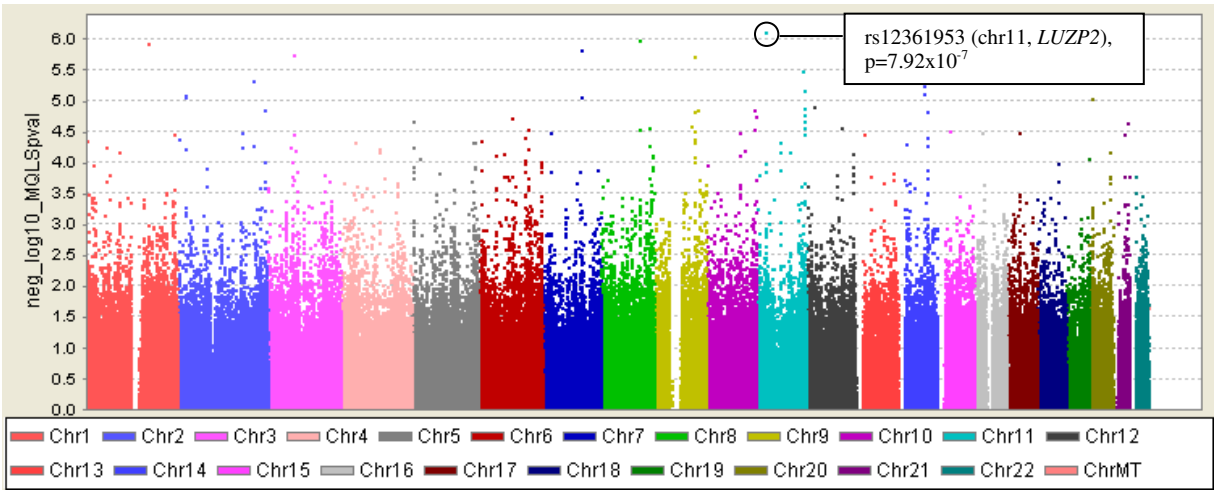



**Figure 3.1. MQLS Manhattan plot.** Genome-wide association results were calculated using MQLS for 798 individuals (109 Late-onset Alzheimer disease affected). The lowest P-value ($7.92 \times 10^{-7}$) was calculated on chromosome 11 at rs12361953 which is located in the Leucine zipper protein 2 (LUZP2) gene.

***Genome-wide linkage***

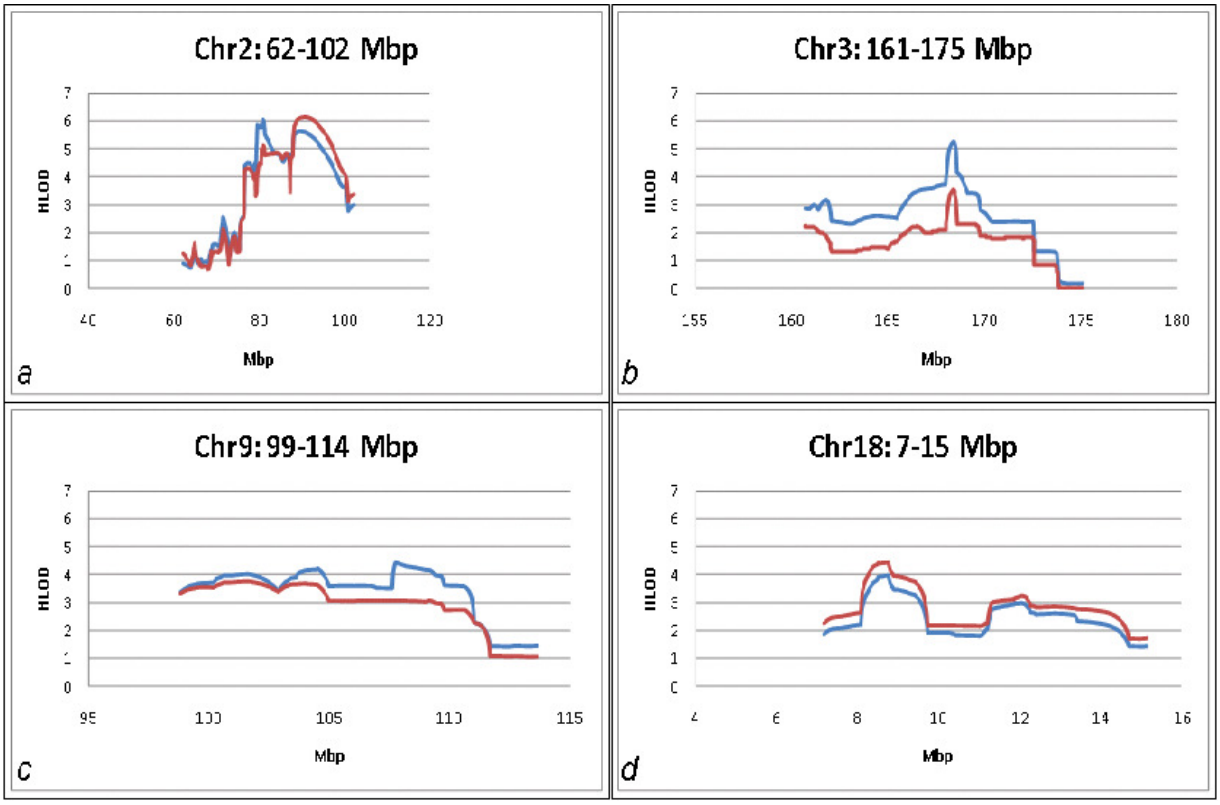
In the genome-wide analysis, forty five regions, among all chromosomes except 17, 21, and X, had at least one two-point HLOD ≥ 3.0 (Appendix B). Multipoint linkage analysis for these regions resulted in four regions, one each on chromosomes 2, 3, 9, and 18 with a multipoint peak HLOD > 3 (Table 3.5, Figure 3.2). The highest peak occurred on chromosome 2 with a recessive peak HLOD of 6.14 (90.91 Mbp) and a dominant peak HLOD of 6.05 (81.03 Mbp). The most significant association results within the recessive and dominant ±1-LOD-unit support interval were rs1258411 (P=5.29x10-2) and rs2974151 (P=1.29x10-4), respectively. Rs1258411 is not located in a gene, but rs2974151 is located in an intron of *CTNNA2* (catenin, alpha 2). In addition to rs2974151, 10 other SNPs in this gene had P-values <0.05. While this is less than 5% of the analyzed SNPs in *CTNNA2*, it still warrants attention.

The next highest multipoint result was on chromosome 3 with a dominant HLOD of 5.27 and a recessive HLOD of 3.53. The peak for both models is at 168.43 Mbp, and the most significant association result in the ±1-LOD-unit support interval was at rs9812366 (P=4.00x10-2), which is intergenic. The linkage peak on chromosome 9 reached an HLOD of 4.44 (107.76 Mbp) under the dominant model and 3.77 (101.7 Mbp) under the recessive model. This peak overlaps with the suggestive linkage peak found in the joint linkage analysis published by Hamshere et al(121), however this region has not been consistently replicated in other studies. For both models the most significant association result in the ±1-LOD-unit support interval was at rs9969729 (P=1.94x10-6), which is intergenic. On chromosome 18 the dominant and recessive results both peaked at 8.77 Mbp with HLOD=3.97 for the dominant model and HLOD=4.43 for the recessive model. The most significant association result in this ±1-LOD-unit support interval was at rs632912 (P=8.80x10-4), which is intergenic. None of these regions overlap the linkage peaks found in our previous genome-wide microsatellite linkage study, which used only a subset of the individuals in the current dataset (56). As with our association results, these multipoint peaks did not encompass the previously known LOAD genes.

**Table 3.5. Most significant multipoint linkage results.** Parametric dominant (Dom) and recessive (Rec) multipoint maximum heterogeneity (HLOD) scores were calculated using Merlin. Regions are determined by ±1-LOD-unit support intervals.

| Chr | Dom Peak HLOD | Dom peak alpha | Position (Mbp) | Region | Lowest MQLS p-value in Region | Rec Peak HLOD | Rec peak alpha | Position (Mbp) | Region | Lowest MQLS p-value in Region |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 6.05 | 0.48 | 81.03 | 79.46-82.95 | 1.29E-4 (rs2974151) | 6.14 | 0.39 | 90.81 | 87.97-97.46 | 5.29E-2 (rs1258411) |
| 3 | 5.27 | 0.49 | 168.43 | 168.06-168.60 | 4.00E-2 (rs9812366) | 3.53 | 0.26 | 168.43 | 168.06-168.60 | 4.00E-2 (rs9812366) |
| 9 | 4.44 | 0.34 | 107.77 | 102.94-110.80 | 1.94E-6 (rs9969729) | 3.77 | 0.23 | 101.7 | 98.86-109.79 | 1.94E-6 (rs9969729) |
| 18 | 3.97 | 0.27 | 8.77 | 8.12-9.59 | 8.80E-4 (rs632912) | 4.43 | 0.21 | 8.77 | 8.12-9.59 | 8.80E-4 (rs632912) |

**Figure 3.2. Strongest multipoint linkage peaks.** Parametric dominant (blue) and recessive (red) multipoint linkage peaks with HLOD scores >3 were calculated on chromosomes 2 (a), 3 (b), 9 (c), and 18 (d). Red=recessive, Blue=dominant




***Evaluation of the MMSE and Word List Learning***

After observing the association with *APOE* and the high linkage peak on chromosome 2, we tested whether the standardized scores (Z scores) on the Word List Learning portion of the neuropsychological battery of tests differentiated between low, normal, and high risk groups according to *APOE* genotype and between individuals in subpedigrees showing linkage between LOAD and 2p12 versus the other subpedigrees.

Comparing mean Z scores between the three LOAD risk groups assigned by APOE genotype, a decrease was seen in mean Z score for almost all Z score categories. The only exceptions were normal risk versus low risk for trial 2, and recognition yes

(high risk and normal risk were equal, mean for low risk was greater than high risk or normal risk) (Table 3.6).

The most significant difference in mean Z scores between the three risk groups was for recognition-no (p=0.001). Follow up analysis with the two-sample Wilcoxon rank sum test revealed that comparison of means between normal risk and high risk (p=0.001) was significant and between low risk and high risk groups (p=0.004) were significant (Table 3.7). The high risk group was on average 8.61 standard deviations below the mean for their ability to recognize that a word was not previously shown to them, and the normal and low risk groups were on average 3.79 and 1.57 standard deviations below the mean, respectively (Table 3.6). Nominally significant results were also seen for delayed recall (0.03) and savings (0.02) Z scores, which are directly related scores since savings is calculated using the delayed recall result and the Word List Memory Trial 3 result. Applying the Wilcoxon rank sum test showed that differences of means for normal risk versus high risk and low risk versus high risk explain the nominally significant results (p=0.02 for both comparisons in both groups) (Table 3.7). After applying a strict Bonferroni correction and multiplying all p-values by 17 (the total number of tests conducted which are presented in Table 3.7), the only significant p-values are for the overall ANCOVA of recognition-no Z scores and the pairwise ANCOVA between the normal risk group versus the high risk group.

After adjusting for MMSE Z scores, the only nominal significant difference in mean Z scores for any of the Word List learning tasks was for recognition-no (p=0.02). The adjusted pairwise comparisons again showed that the differences in means was driven by the highest risk group since low risk versus high risk and normal versus high

risk were both significant, but not low risk versus normal risk (Table 3.8). None of these results would withstand a Bonferroni correction to the p-values (multiplying by 10). However, we had very small samples sizes in these analyses.

**Table 3.6. MMSE and Word list learning Z scores per LOAD risk group defined by *APOE*.** Affected individuals with the APOE 3/4 or 4/4 genotype were assigned to the high risk group, those with APOE 2/4 or 3/3 were assigned to the normal risk group, and those with the APOE 2/3 genotype were assigned to the low risk group. MMSE=Mini-Mental State Exam. Min=Minimum Z score. Max=Maximum Z score.

| | High Risk (APOE 3/4 and 4/4) | | | | Normal Risk (APOE 2/4 and 3/3) | | | | Low Risk (APOE 2/3) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # individuals | Mean | Min | Max | # individuals | Mean | Min | Max | # individuals | Mean | Min | Max |
| MMSE Z | 22 | -7.15 | -12.55 | -0.73 | 39 | -6.54 | -12.55 | -0.27 | 7 | -5.66 | -12.55 | -1.18 |
| Word List Memory Trial 1 Z | 22 | -2.06 | -3.00 | 0.13 | 40 | -2.02 | -3.00 | 0.13 | 7 | -1.93 | -3.00 | -0.50 |
| Word List Memory Trial 2 Z | 22 | -2.30 | -3.88 | -0.94 | 40 | -2.13 | -4.42 | -0.35 | 7 | -2.20 | -3.88 | -1.53 |
| Word List Memory Trial 3 Z | 22 | -2.37 | -4.00 | -1.37 | 40 | -2.28 | -4.43 | -0.32 | 7 | -1.97 | -2.95 | -0.84 |
| Delayed Recall Z | 22 | -3.29 | -3.53 | -1.42 | 40 | -2.85 | -4.06 | -0.37 | 7 | -2.70 | -3.53 | -1.42 |
| Savings Z | 22 | -4.48 | -5.18 | 0.60 | 40 | -3.49 | -5.18 | 1.75 | 7 | -2.72 | -5.18 | -0.60 |
| Recognition-Yes Z | 22 | -2.33 | -7.75 | 0.58 | 38 | -2.33 | -7.75 | 0.58 | 7 | -2.87 | -6.08 | -0.25 |
| Recognition-No Z | 22 | -8.61 | -29.67 | 0.33 | 38 | -3.79 | -33.00 | 0.33 | 7 | -1.57 | -3.00 | 0.33 |

**Table 3.7. Kruskal Wallis test results with follow-up two-sample Wilcoxon rank sum test results.** Mean Z scores were compared between the three LOAD risk groups based on APOE genotype (low=APOE 2/3, normal=APOE 2/4 or 3/3, high=APOE 3/4 or 4/4) using the Kruskal-Wallis test. Only significant results were followed up with the two-sample Wilcoxon rank sum test.

| | | Two-sample Wilcoxon rank sum test p-value | | |
|---|---|---|---|---|
| | Kruskal-Wallis p-value | low vs. normal | normal vs. high | low vs. high |
| MMSE Z | 0.65 | - | - | - |
| Word List Memory Trial 1 Z | 0.89 | - | - | - |
| Word List Memory Trial 2 Z | 0.71 | - | - | - |
| Word List Memory Trial 3 Z | 0.71 | - | - | - |
| **Delayed Recall Z** | **0.03** | **0.52** | **0.02** | **0.02** |
| **Savings Z** | **0.02** | **0.37** | **0.02** | **0.02** |
| Recognition-Yes Z | 0.85 | - | - | - |
| **Recognition-No Z** | **0.001** | **0.56** | **0.001** | **0.004** |

**Table 3.8. Analysis of covariance test results with follow-up pairwise test results.** Mean Z scores were compared between the three LOAD risk groups, based on *APOE* genotype (low=*APOE 2/3*, normal=*APOE 2/4* or *3/3*, high=*APOE 3/4* or *4/4*), adjusting for MMSE Z scores using the analysis of covariance test in STATA. Only significant results were followed up with pairwise tests.

| | | pairwise ANCOVA p-value | | |
|---|---|---|---|---|
| | ANCOVA p-value | low vs. normal | normal vs. high | low vs. high |
| Word List Memory Trial 1 Z | 0.99 | - | - | - |
| Word List Memory Trial 2 Z | 0.89 | - | - | - |
| Word List Memory Trial 3 Z | 0.79 | - | - | - |
| Delayed Recall Z | 0.08 | - | | |
| Savings Z | 0.06 | - | - | - |
| Recognition Yes Z | 0.74 | - | - | - |
| Recognition No Z | 0.02 | 0.37 | 0.02 | 0.03 |

Significant ($p \leq 0.05$) correlations between Z scores and per-family lod scores were seen for Word List Memory trials 2 and 3, delayed recall, and savings. The correlation coefficient for each of the significant correlations was positive but weak, ranging from 0.25 to 0.31 (Table 3.9). All other correlations were very weak and not significant (see Appendix D for scatter plots). Using MMSE Z scores as a covariate to test for the correlation between the Word List memory, delayed recall, and recognition procedures produced almost identical results (Appendix E). All in all, we do not see a strong correlation between our linkage result on 2p12 and these specific neuropsychological test results.

**Table 3.9 Spearman's correlation between 2p12 lod scores and Z scores of MMSE and Word List learning.** Recessive LOD refers to the per-family lod scores at the peak of the 2p12 linkage region calculated under a recessive model, and dominant LOD refers to per-family lod scores at the peak of the 2p12 linkage region calculated under a dominant model. Spearman's rho is the correlation coefficient.

| | Recessive LOD | | Dominant LOD | |
|---|---|---|---|---|
| | Spearman's rho | p-value | Spearman's rho | p-value |
| MMSE Z | -0.15 | 0.22 | 0.10 | 0.43 |
| Word List Memory Trial 1 Z | -0.07 | 0.55 | 0.14 | 0.24 |
| Word List Memory Trial 2 Z | 0.15 | 0.22 | 0.31 | **0.01** |
| Word List Memory Trial 3 Z | 0.08 | 0.49 | 0.25 | **0.04** |
| Delayed Recall Z | 0.15 | 0.21 | 0.27 | **0.02** |
| Savings Z | 0.2 | 0.11 | 0.27 | **0.03** |
| Recognition-Yes Z | 0.04 | 0.75 | 0.06 | 0.61 |
| Recognition-No Z | 0.002 | 0.99 | -0.02 | 0.87 |

## Discussion

*APOE* was clearly associated with dementia in our population; however, it did not explain the majority of affected individuals. In the Adams County communities, there were only 8/74 individuals who carried at least one *APOE-E4* allele. In the remaining Amish communities, the *APOE-E4* allele was more common, but still less common than in the general population. In addition, the majority of affected individuals (81/127, 64% for all counties; 45/115, 39% for non-Adams counties) did not carry an *APOE-E4* allele. The specific deficit of the *APOE-E4* allele in Adams County as well as differences in allele frequencies for some of the top GWAS SNPs indicates at least some level of locus heterogeneity underlying LOAD in the Amish population.

Additional support for locus heterogeneity arises from the linkage results. Examination of the subpedigree-specific lod scores for the four significant loci indicates that 13 of the 34 subpedigrees generate no lod scores >0.50 for any of the loci, and 14/21(67%) of the remaining subpedigrees generate lod scores >0.50 for only one of the 4 loci. In addition, the vast majority of the remaining SNPs across the genome generated HLOD scores with alpha values (proportion of linked pedigrees) <1.0. Finally, the suggestion of locus heterogeneity is consistent with the societal differences across church districts, which can further restrict marriages even within the Amish.

Because of the relatedness of individuals in our dataset we could take advantage of both linkage and association approaches to identify potential LOAD loci. In our examination, we found that our most significant association results did not fall under any of the four linkage peaks. However, under the linkage peaks we did see some evidence of association. Within our most significant region of linkage lies *CTNNA2*,

which also had suggestive evidence for association. In addition to the result at rs2974151 (P=1.29x10-4), multiple SNPs in *CTNNA2* had P-values < 0.05, decreasing the likelihood of a false positive association for this gene. However, because of the relatedness in our dataset it was difficult to get an accurate measurement of LD structure to determine if the SNPs in this region were more highly correlated due to a founder effect.

*CTNNA2* encodes the catenin alpha 2 protein, which is a neuronal-specific catenin. Catenins are cadherin-associated proteins and are thought to link cadherins to the cytoskeleton to regulate cell-cell adhesion. Catenin alpha 2 can form complexes with other catenins such as beta-catenin, which interacts with presenilin. Mutations in presenilin lead to destabilization of beta-catenin which potentiates neuronal apoptosis (122). Catenin alpha 2 is also thought to regulate morphological plasticity of synapses and cerebellar and hippocampal lamination during development in mice (123). It also functions in the control of startle modulation in mice (123).

It was not completely unexpected to see some discordance between the linkage and association results, as was demonstrated in our *APOE* results where we saw evidence for association but not for linkage. Because we needed to divide the pedigree to facilitate linkage analysis and because we used an affecteds-only analysis, only a subset of the individuals analyzed in association analysis were analyzed in linkage analysis. The breaking of the pedigree likely reduces the observed genomic sharing between relatives as the tracking of the natural flow of alleles was somewhat disrupted, as we saw when we tested *APOE* for linkage. Also, the very nature of association analysis versus linkage analysis will provide some different results. Linkage analysis locates shared genomic regions between affected individuals in the same pedigree by

testing for co-segregation of a chromosomal segment from a common ancestor. Association using MQLS tests for differences in allele frequencies between affected and unaffected individuals while correcting for the pedigree relationships. Association analysis is more powerful in detecting protective effects as well as smaller effects in the population compared to affecteds-only linkage analysis but is underpowered when sample sizes are small and genetic heterogeneity is present. Conversely, linkage analysis is more suitable for finding large effects in a small number of related individuals and is more robust to allelic heterogeneity.

Genetic locus heterogeneity in complex diseases is likely tied to phenotypic heterogeneity. Word List Learning scores have also been shown to be very heritable (92) and well differentiate LOAD cases from controls (120). Therefore, we tested whether *APOE* genotypes or linkage to 2p12 are correlated with MMSE scores or Word List Learning with delayed recall and recognition scores. With just a couple of exceptions, mean performance on each of the tests decreased with higher risk for LOAD due to *APOE* genotype. Therefore, one could make an argument that *APOE* specifically affects learning and memory. However, after adjusting for overall cognitive impairment indicated by MMSE scores and applying a Bonferroni correction, the only significant difference was seen between the normal risk group and high risk group for recognition-no. A similar study was performed in the Cache County cohort where they tested the affect of *APOE* on neurocognitive measures including the MMSE and Word List Learning; however, only cognitively normal individuals were included in the analysis (124). In their analysis the only significant effect observed of *APOE* on any neurocognitive test result was for delayed recall, but they only saw the effect in the group of individuals with greater than 12 years of education. We also saw nominal

significance for delayed recall, but most of the individuals in our dataset have fewer than 12 years of education. Other studies using different neuropsychlogical tests also found significant associations between *APOE* and episodic memory (125-128) showing that *APOE* not only affects AD status but also specific neurocognitive deficits such as episodic memory.

The MMSE and the Word List Learning with delayed recall and recognition procedures are not sufficient to explain our linkage peak on 2p12. In the future other test scores from the CERAD battery could be analyzed to test for a possible correlation between the lod scores at 2p12 and one or more of these other specific deficits. Or, it is possible that none of the specific test results are correlated with lod scores and the genetic heterogeneity at this locus is not detected by heterogeneity in these test results.

Our results confirmed the complex genetic architecture of LOAD even in this more homogeneous set of individuals. Multiple loci appeared to be significantly contributing to LOAD risk in the Amish. We replicated the affect of APOE, replicated the evidence for linkage on 9q22, and also found modest evidence for association of both *PICALM* and *MS4A* in this population. Most importantly, this unique population allowed us to find additional candidate loci, particularly in the *CTNNA2* region in which we saw strong evidence for both linkage and association. The role of *CTNNA2* in the brain also makes this gene a promising candidate. The *CTNNA2* region, in addition to other potential risk regions, need to be more closely examined to identify the underlying responsible variants and their functional consequences. We, therefore, have performed a sequence analysis of *CTNNA2* that is presented in Chapter IV.

**Acknowledgements**

We thank the family participants and community members for graciously agreeing to participate, making this research possible. This study is supported by the National Institutes of Health grants AG019085 (to JLH and MAP-V) and AG019726 (to WKS). Some of the samples used in this study were collected while WKS, JRG, and MAP-V were faculty members at Duke University. The authors would like to thank Gene Jackson of Scott & White for his effort and support on this project. Additional work was performed using the Vanderbilt Center for Human Genetics Research Core facilities: the Genetic Studies Ascertainment Core, the DNA Resources Core, and the Computational Genomics Core.

# CHAPTER IV

# SEQUENCE ANALYSIS OF A NOVEL ALZHEIMER DISEASE CANDIDATE GENE: CTNNA2

## Introduction

Up to 75% of the heritability of late-onset Alzheimer disease (LOAD) remains unexplained (52). To explain at least part of this missing heritability, we have been studying the genetically isolated Amish populations of Ohio and Indiana. Chapters II and III discuss our application of a genome-wide SNP association and linkage study, which led to the identification of the *CTNNA2* (catenin alpha 2) gene, which encodes the CTNNA2 (catenin alpha 2) protein, as a strong candidate for LOAD. *CTNNA2* is located in our most strongly linked region (2p12) where the HLOD (heterogeneity lod) score reached 6.14 under a recessive model and 6.05 under a dominant model. Several SNPs in the gene were also associated with LOAD with $P<.05$ (lowest $P=1.29x10\text{-}4$). Within the seventh intron of *CTNNA2* resides the small, one exon gene, *LRRTM1* (leucine rich repeat neuronal transmembrane protein 1).

Catenin proteins bind to cadherins and link the cadherins to the cytoskeleton, which regulates cell-to-cell adhesion. The catenin alpha 2 protein is a catenin specific to neurons and also binds to and works in conjunction with other catenins, such as beta-catenin. Beta-catenin interacts with presenilin, and mutations in presenilin have been linked to early-onset AD (7) and lead to destabilization of beta-catenin, which

potentiates neuronal apoptosis (122). It is possible that CTNNA2 plays a role in the amyloid beta pathway, which is central to Alzheimer disease (AD) pathophysiology.

In mice, CTNNA2 plays important roles in brain development by regulating morphological plasticity of synapses and cerebellar and hippocampal lamination (123). The hippocampal and cerebellar regions of the brain are important for learning and memory and are the first areas to show physical changes in Alzheimer patients. Although these mouse studies focus on the developing brain, it is possible that the same gene also plays a role in the aging brain, as is suggested by gene expression studies in adult rhesus monkey brains that show high levels of expression of CTNNA2 in the dorsolateral prefrontal cortex and in the hippocampus (129).

SNPs in a related gene, *CTNNA3* (also known as VR22) were previously associated with Aβ levels (130) and AD status (131). *CTNNA3* was of particular interest because it is located in the region on chromosome 10 linked to Alzheimer disease (62;132-134), and, like beta-catenin, CTNNA3 binds to presenilin. *CTNNA3* contains *LRRTM3* in its seventh intron, similar to the *CTNNA2*/*LRRTM1* relationship. Some of the significant SNPs reported by Martin et al in their analysis of *CTNNA3* were SNPs in the exon of *LRRTM3* (131). However, additional reports have had mixed results for this association between *CTNNA3/LRRTM3* and AD (75;77;135-139), and it has not been associated in any of the large scale GWAS studies (78-82).

Neither *CTNNA2* nor *LRRTM1* have previously been proposed as an AD candidate gene in the comprehensive listing of AD candidate genes in the AlzGene database (alzgene.org). The evidence from our genome-wide linkage and association study, suggestive roles of other related proteins and genes (beta-catenin and *CTNNA3*)

in AD, and the known functional roles of *CTNNA2* in the brain, prompted us to perform a sequence study of *CTNNA2* and its nested gene *LRRTM1*.

The true functional disease associated variant is rarely genotyped directly using GWAS data, but rather is hopefully tagged through linkage disequilibrium. Therefore, followup with more genotyping and/or sequencing is necessary to fine map a putative region. Also, the idea that rare variants could be playing a role in common diseases is becoming a topic of interest in the field of human genetics as next-generation sequencing has become more accessible in terms of cost and efficiency of protocols. Historically, large pedigrees have been successful for finding rare disease variants, especially for Alzheimer disease (4-7). To assess the validity of *CTNNA2/LRRTM1*'s involvement in LOAD risk, we have examined the exonic sequence and partial nonexonic sequence of this gene pair in 101 (47 LOAD) Amish individuals.

## Methods

### ***Study population***

A subset of 100 individuals from the genome-wide dataset (presented in Chapter III) was selected for whole exome sequencing. Ascertainment and clinical assessment details are presented in Chapter III. To maximize the potential to find a novel LOAD genetic variant, the following prioritizations were used to select individuals for whole exome sequencing: 1) LOAD affected and unaffected individuals in the sub-pedigrees with LOD >1.5 at 2p12 in our previous genome-wide linkage analysis; 2) LOAD affected individuals with the *APOE 2/3* or *APOE 3/3* genotype and their unaffected siblings; 3) sibships with at least two LOAD individuals and two unaffected individuals; 4) LOAD

affected individuals and unaffected individuals with kinship coefficients ≥0.0625 (kinship coefficient for first cousins) with any of the sibships in the previous category. We also included one additional LOAD individual and fifteen controls which were selected for an overlapping study of Parkinson disease in the Amish.

Because standard quality control measures for next-generation sequence data have not been well established, we determined thresholds based on what appeared to be outliers for each quality metric. Fifteen individuals were eliminated after employing the following quality control thresholds for the whole exome sequences: 1) <53,000,000 total reads, 2) <52,000,000 uniquely mapped reads, 3) <45,000,000 on target reads, 4) >50% duplication rate, 5) <50% reads with at least 10X coverage, and 6) <65% capture efficiency in the whole exome sequence or in the exonic sequence of *CTNNA2*. The final dataset used for analysis is shown in Tables 4.1 and 4.2.

**Table 4.1 Sequencing dataset characteristics including total number of individuals, *APOE* genotype, and mean and range of ages of exam and onset.**

| | LOAD | Cognitively normal | Total |
|---|---|---|---|
| Total | 47 | 54 | 101 |
| Number Female (%) | 27 (57%) | 32 (59%) | 59 (58%) |
| Number Male (%) | 20 (43%) | 22 (41%) | 42 (42%) |
| APOE 4/4 | 0 | 0 | 0 |
| APOE 3/4 | 14 | 6 | 20 |
| APOE 2/4 | 0 | 2 | 2 |
| APOE 3/3 | 27 | 41 | 68 |
| APOE 2/3 | 6 | 5 | 11 |
| APOE 2/2 | 0 | 0 | 0 |
| Mean (range) age at exam | 82 (64-97) | 77 (62-91) | 79 (62-97) |
| Mean (range) age at onset | 79 (58-93) | - | - |

**Table 4.2 Whole exome sequence quality of dataset used for analysis.**

| | Mean | Range |
|---|---|---|
| Total Reads | 102,260,136 | 57,920,032 – 204,943,436 |
| Reads Uniquely Map to HG19 | 100,178,092 | 57,415,402 – 202,542,103 |
| Percent Reads Uniquely Mapped to HG 19 | 98.01% | 96.00% - 99.14% |
| Reads on Target | 75,108,125 | 46,338,263 – 167,691,437 |
| Capture Efficiency | 74.86% | 69.55% - 86.00% |
| Duplication Rate | 18.19% | 4.00% - 46.89% |
| Percent Reads Above 5x | 78.80% | 68.00% - 89.00% |
| Percent Reads Above 10x | 65.90% | 54.00% - 77.00% |
| Percent Reads Above 15x | 57.27% | 45.00% - 69.00% |
| Percent Reads Above 20x | 50.54% | 37.00% - 63.00% |
| Percent Reads Above 30x | 40.04% | 26.00% - 53.00% |

***Sequencing***

Paired-end sequences for *CTNNA2/LRRTM1* were obtained via paired-end whole exome sequencing from two sequencing sites: The Genome Sciences Resource at the Vanderbilt University Medical Center and the sequencing core of the Center for Genome Technology at the Hussman Institute for Human Genomics at the University of Miami Miller School Of Medicine. Exons from genomic DNA were captured with the Agilent SureSelect Human All Exon 50 Mb capture kit, which was designed to capture coding exons annotated by the GENCODE plus any additional exons annotated in the consensus CDS database. Ten basepairs of flanking sequence for each targeted exon is included, and small non-coding RNAs from the miRBase (version 13) and Rfam were also included. The DNA sample of interest is first sheared to create a library of whole genome DNA from which the targeted regions are enriched by capturing them with 120-mer biotinylated cRNA baits. The exome DNA library was then sequenced with read lengths of 75 basepairs on the Illumina HiSeq 2000.

***Sequence processing***

Read mapping to hg19 was performed using BWA (Burrows-Wheeler Aligner) (140). Duplicates were marked using Picard (http://picard.sourceforge.net), and recalibration, realignment, SNV (single nucleotide variant) calling, and SNV filtering to increase the validity of variants called were performed using GATK (The Genome Analysis Toolkit) (141).

Single-sample variant calling and multiple-sample variant calling were both applied using GATK, and a second calling algorithm, GlfMultiples (http://genome.sph.umich.edu/wiki/GlfMultiples), was also applied to perform multiple-sample variant calling. Both algorithms use a probabilistic framework and should produce very similar results. For the single-sample call set of GATK-called variants, the following quality filters were applied, as suggested by one of GATK's 'Best Practices': QualByDepth (QD) > 2, RMSMappingQuality (MQ) > 40, FisherStrand (FS) < 60, Haplotype score < 13, MappingQualityRankSumTest (MQRanksum) > -12.5, ReadPosRankSum > -8. Variants that do not pass these quality filters are likely due to sequencing or alignment errors leading to false positive variant calls. QualbyDepth is the variant confidence divided by the unfiltered depth. The RMSMappingQuality is the root mean square of the mapping quality. FisherStrand is the Phred-scaled p-value using Fisher's Exact Test to detect strand bias (the variation being seen on only the forward or only the reverse strand) in the reads. A haplotype score assesses the likelihood of other variants within 10 basepairs of the variant of interest to detect possible alignment errors. A higher score indicates regions of bad alignment which can lead to false positive SNV calls. The MQRankSum is the u-based z-approximation from the Mann-Whitney Rank Sum Test for mapping qualities (reads with reference bases

versus those with alternate alleles). ReadPosRankSum is the u-based z-approximation from the Mann-Whitney Rank Sum Test for the distance from the end of the reads for reads with the alternate allele. If the alternate allele is only seen near the ends of reads it can indicate an error.

For the multiple-sample GATK-called variants, SNVs with base quality or map quality ≤20 were removed. Base and map qualities are given on a phred scale, so a quality of 20 equates to a 0.01 chance that the base call or the alignment is an error. No quality filtering was performed for the GLFMultiples call set to provide an additional point of comparison between the call sets.

Therefore, three different call sets (single-sample GATK, multi-sample GATK, and GLFMultiples) were available for increased confidence in the variants called. Annotation of all variants was obtained from SeattleSeq Annotation 131 (http://snp.gs.washington.edu/SeattleSeqAnnotation131/).

***Analysis***

Genotypes from the single-sample variant call set were used for all allele frequency calculations since more stringent quality control was applied to that call set. Allele frequencies for all discovered exonic SNVs were compared between all cases and controls and between cases and controls in the subset of individuals who contributed to the linkage signal on chromosome 2. We also tabulated all noncoding SNVs, including all intronic SNVs and any SNVs detected 150 Kb upstream and downstream of *CTNNA2*, and compared allele frequencies between affecteds and unaffected for all SNVs that were discovered in all three call sets. Because of the complicated relatedness

of the dataset and because the purpose for this study is to screen for variants that will be followed up in the broader dataset, no statistical tests were performed.

***Genotyping***

Rs72822556 was genotyped using Sequenom's iPLEX Gold assay on the MassARRAY platform (San Diego, CA) according to manufacturer's instructions (www.sequenom.com) on 142 LOAD affected and 542 cognitively normal Amish individuals. Genotypes for only three of the sequenced individuals could not be obtained. One of the three individuals was LOAD affected and in a subpedigree with lod >1.5 at 2p12. The other two were cognitively normal and not in a subpedigree with lod >1.5 at 2p12.

## Results

***Variation in the exons***

A total of nine exonic SNVs were discovered in *CTNNA2* and *LRRTM1* (Table 4.3). Of the nine SNVs identified, rs17019360 in *CTNNA2* was the only SNP genotyped in our genome-wide study (presented in Chapter III). The genotypes from the genome-wide study were 100% concordant with the genotypes determined from sequencing. Rs17019360 and rs61291641 were the only two SNVs detected in *CTNNA2* and were both synonymous variations. Two of the seven SNVs in *LRRTM1* were synonymous variations and one (basepair position 80529418) was not listed in dbSNP build 134, 1000 Genomes phase 1, nor the Exome Variant Server (NHLBI GO Exome Sequencing Project). The other five *LRRTM1* SNVs were missense variations. The PolyPhen

predictions, which predict functional effects of SNPs on a gene, were benign for rs6733871 and rs76300062 and unknown for rs141752316 and the two other missense mutations, which previously have not been documented in dbSNP, 1000 Genomes, nor the Exome Variant Server. Rs141752316 was not verified with either of the multiple-sample variant calling algorithms; therefore, it could be a sequencing artifact or the multiple sample variant calling might not have been sensitive enough to call this singleton. The fact that this variant is listed in the Exome Variant Server makes the latter explanation more plausible.

Almost no allele frequency differences were observed between all LOAD affected individuals and cognitively normal individuals. Four of the nine exonic variants were present in the individuals from the subpedigres showing linkage to 2p12 in the genome-wide study, but very few of the individuals actually had the variation. Rs17019360 was the most common variant with four of the ten LOAD individuals being heterozygous and one of the six cognitively normal individuals being heterozygous (Table 4.3).

**Table 4.3 Summary of all detected SNVs in the exons of *CTNNA2* and *LRRTM1*.** Basepair positions are HG19 positions. Function was obtained from SeattleSeq Annotation. LOAD=late-onset Alzheimer disease individuals. Het=heterozygous. Hom=homozygous for alternate allele.

| **All detected exonic SNVs** | | | | **Entire dataset**<br><br>**47 LOAD, 54 Cognitively normal** | | **Individuals in sub-pedigrees with LOD >1.5 at 2p12 in genome-wide linkage study**<br><br>**10 LOAD, 6 Cognitively normal** | |
|---|---|---|---|---|---|---|---|
| **Chr2 bp position** | **rs ID** | **Gene** | **Function** | **# LOAD individuals with SNV (allele frequency)** | **# Cognitively normal with SNV (allele frequency)** | **# LOAD individuals with SNV (allele frequency)** | **# Cognitively normal with SNV (allele frequency)** |
| 80101321 | rs61291641 | *CTNNA2* | coding-synonymous | 2 het, 1 hom (4.3%) | 2 het (1.9%) | 1 het (5.0%) | 0 |
| 80529418 | N/A | *LRRTM1* | coding-synonymous | 3 het (3.2%) | 3 het (2.8%) | 1 het (5.0%) | 0 |
| 80529655 | rs138747880 | *LRRTM1* | coding-synonymous | 4 het (4.3%) | 6 het (5.6%) | 0 | 0 |
| 80529956 | rs6733871 | *LRRTM1* | missense | 11 het, 2 hom (16.0%) | 12 het, 3 hom (16.7%) | 2 het (10.0%) | 1 het (8.3%) |
| 80530062 | rs76300062 | *LRRTM1* | missense | 1 het (1.1%) | 0 | 0 | 0 |
| 80530625 | rs141752316 | *LRRTM1* | missense | 0 | 1 het (0.9%) | 0 | 0 |
| 80530868 | N/A | *LRRTM1* | missense | 1 het (1.1%) | 1 het (0.9%) | 0 | 0 |
| 80530886 | N/A | *LRRTM1* | missense | 0 | 1 het (0.9%) | 0 | 0 |
| 80801346 | rs17019360 | *CTNNA2* | coding-synonymous | 17 het, 5 hom (28.7%) | 20 het, 5 hom (27.8%) | 4 het (20.0%) | 1 het (8.3%) |

***Extra-exonic variants***

Outside the exons, including intronic regions and regions 150 Kb upstream and downstream *CTNNA2*, 1,811 SNVs were found in all three call sets. Four SNV's, none of which were genotyped in our previous genome-wide study, had at least a 30% difference in alternate allele frequency between the LOAD affected and unaffected individuals in the chromosome 2 peak subpedigree individuals (Table 4.4). None of these variants were listed in the Exome Variant Server, most likely due to coverage threshold (8x) used in those projects. The greatest difference was seen for rs72822556, an intronic SNV. Of the ten LOAD affected individuals in the chromosome 2 peak subpedigrees, six individuals had the alternate allele, and five of the six individuals were homozygous for the alternate allele. None of the six cognitively normal individuals in the chromosome 2 peak subpedigrees had the alternate allele. After expanding the analysis to include all 47 LOAD and 54 cognitively normal individuals, four additional LOAD patients had the alternate allele, one of which was homozygous. Four of the cognitively normal individuals had the alternate allele. Therefore, we saw a higher concentration of this variant in the subpedigrees that showed a strong linkage signal for this chromosome 2 region.

Rs6719427, an intronic SNV, had a 50% versus 17% allele frequency in the LOAD affected versus unaffected individuals respectively. Expanding the analysis to all individuals, we saw a 43% versus 34.3% allele frequency in the LOAD patients versus unaffected individuals. Rs7595284 (upstream of *CTNNA2*) and an SNV at basepair position 80270223 (intronic) both had a 30% allele frequency difference between LOAD affected and unaffected individuals in the chromosome 2 peak subpedigrees. These

differences decreased to 8% and 0.3%for position 80270223 and rs7595284 respectively when including all individuals in the calculations.

We chose to genotype rs72822556 to attempt to verify the sequencing-generated genotypes since it showed such substantial differences in allele frequency between cases and controls. The genotyping did confirm that there is variation at this basepair position in our dataset. Unfortunately, 30% of the sequencing-generated genotypes were discordant with the Sequenom-generated genotypes resulting in a much smaller gap between alternate allele frequencies of cases and controls. In the 2p12 peak subpedigrees, no individuals homozygous for the alternate allele were observed. However, a genotype could not be obtained for one of the sequenced individuals who appeared to be homozygous for the alternate allele. Six out of the 11 genotyped cases were heterozygous (27.3% alternate allele frequency), and 2 out of the 14 genotyped controls were heterozygous (14.3% alternate allele frequency). In all 142 genotyped cases and 542 genotyped controls, the alternate allele frequencies were 20.8% and 17.7%, respectively (Table 4.4). The average depth of coverage at rs72822556 was only 2, so it is not surprising that several of the genotypes could not be reproduced.

**Table 4.4 Summary of selected non-exonic SNVs with at least a 30% difference in allele frequency between LOAD and cognitively normal individuals in the subpedigrees showing the most evidence for linkage at 2p12.** Basepair positions are HG19 positions. Function was obtained from SeattleSeq Annotation. LOAD=late-onset Alzheimer disease individuals. Het=heterozygous. Hom=homozygous for alternate allele. SNV (rs72822556) with the largest difference in allele frequency between LOAD and cognitively normal individuals is bolded and was followed up with genotyping.

| **Selected non-exonic SNVs** | | | | **Entire dataset**<br>**47 LOAD, 54 Cognitively normal** | | **Individuals in sub-pedigrees with LOD >1.5 at 2p12 in genome-wide linkage study**<br>**10 LOAD, 6 Cognitively normal** | |
|---|---|---|---|---|---|---|---|
| **Chr2 bp position** | **rs ID** | **Gene** | **Function** | **# LOAD individuals with SNV (allele frequency)** | **# Cognitively normal with SNV (allele frequency)** | **# LOAD individuals with SNV (allele frequency)** | **# Cognitively normal with SNV (allele frequency)** |
| 79603240 | rs7595284 | 5′ of *CTNNA2* | unknown (intergenic) | 6 het, 13 hom (34.0%) | 7 het, 15 hom (34.3%) | 1 het, 5 hom (55.0%) | 1 het, 1 hom (25.0%) |
| **79937921** | **rs72822556** | ***CTNNA2*** | **unknown (intronic)** | **5 het, 6 hom (18.1%)** | **3 het, 1 hom (4.6%)** | **1 het, 5 hom (55%)** | **0** |
| 80270223 | N/A | *CTNNA2* | unknown (intronic) | 7 hom (14.9%) | 1 het, 3 hom (6.5%) | 3 hom (30.0%) | 0 |
| 80801052 | rs6719427 | *CTNNA2* | unknown (intronic) | 10 het, 15 hom (42.6%) | 9 het, 14 hom (34.3%) | 2 het, 4 hom (50.0%) | 1 hom (16.7%) |

**Table 4.5 Sequenom-generated genotype results of rs72822556.** Basepair position is HG19. LOAD=late-onset Alzheimer disease individuals. Het=heterozygous. Hom=homozygous for alternate allele.

| | | **Entire genotyping dataset**<br><br>**142 LOAD, 542 Cognitively normal** | | **Individuals in sub-pedigrees with LOD >1.5 at 2p12 in genome-wide linkage study**<br><br>**11 LOAD, 7 Cognitively normal** | |
|---|---|---|---|---|---|
| **Chr2 bp position** | **rs ID** | **# LOAD individuals with SNV (allele frequency)** | **# Cognitively normal with SNV (allele frequency)** | **# LOAD individuals with SNV (allele frequency)** | **# Cognitively normal with SNV (allele frequency)** |
| 79937921 | rs72822556 | 6 het (27.3%) | 2 het (14.3%) | 49 het, 5 hom (20.8%) | 162 het, 15 hom (17.7%) |

## Discussion

We have detected nine exonic variants and at least 1,811 possible extra-exonic variants at the *CTNNA2/LRRTM1* locus. We did not discover any novel rare variants in the coding sequence of *CTNNA2*. We discovered five SNVs in LRRTM1 that previously have not been recorded in dbSNP or 1000 Genomes. None of the exonic SNVs seems to explain the high linkage peak we see on 2p12 in our genome-wide linkage analysis. Expanding our analysis to extra-exonic SNVs we saw a substantial difference in allele frequencies between LOAD affected and cognitively normal individuals at rs72822556, particularly in the individuals who contributed the most to our strong linkage signal on 2p12. However, verification genotyping, while confirming the presense of the variation, did not reproduce 30% of the genotypes. The discordance presumably is due to the extremely low coverage at this SNP, confirming the need for higher coverage to make reliable genotype calls. The genotypes revealed a much smaller alternate allele frequency difference between cases and controls, negating the predicted importance of this SNV. Therefore, SNVs in the small portion of noncoding sequence that was captured also does not seem to explain our linkage peak on 2p12.

This sequence analysis also demonstrates a difference between single sample and multiple sample variant calling, since we only saw the variant at basepair position 80530886 with single sample variant calling. Single-sample calling is more prone to false positive singleton variant calls, while multiple sample calling might be less sensitive to true singleton variants. Because the single-sample variant calling went through more stringent quality filters we can be more confident in the validity of the variant.

However, without followup genotyping or Sanger sequencing we cannot make any decisive conclusions.

These results suggest that any role that variation in *CTNNA2* and/or *LRRTM1* play in late-onset Alzheimer disease risk is due to variations other than SNVs in the coding sequence. The poor coverage of noncoding regions did not allow for a thorough examination of those regions of *CTNNA2*. Therefore, any further sequencing of this gene should include better coverage of the intronic and 5′ and 3′ regions surrounding *CTNNA2*. Future studies should also explore other types of variation such as insertions and deletions in *CTNNA2* and *LRRTM1*, which could be driving the linkage signal on 2p12. It is also very possible that the linkage peak at 2p12 is the result of other variants in other genes or intergenic regions, which should also be explored. Detection of additional variants in this gene could lead to better understanding of the functional role of this gene in AD pathophysiology.

**Acknowledgements**

Whole exome sequencing was done in the Genome Sciences Resource at Vanderbilt Medical Center under the management of Dr. Travis Clark and under the direction of Dr. Christopher Coldren and Dr. Mark Magnusun. Additional whole exome sequencing was performed at the sequencing core of the Center for Genome Technology, led by John L. Gilbert, at the Hussman Institue for Human Genomics at the University of Miami Miller School of Medicine. Genotyping was planned and implemented by Joshua Hoffman, Ping Mayo, Dr. Nathalie Schnetz-Boutaud, and Melissa Allen. Miguel Herrera generated the quality metrics for the sequences. Christian Shaffer developed the

sequence processing and variant calling pipeline for single-sample variant calling. Eric Torstenson in the laboratory of Dr. Chun Li processed sequences and generated the GATK multi-sample call set. Dr. Bingshan Li ran the GlfMultiples calling algorithm. This work was funded by NIH grants AG019085 to Jonathan L. Haines and AG019726 to William K. Scott, a Discovery Grant from Vanderbilt University, and Michael J Fox Foundation grants.

# CHAPTER V

# CONCLUSION

## Summary

Alzheimer disease is a neurodegenerative disease and is complex in many ways including its genetic etiology. Much of the complexity is due to the heterogeneity which weakens most genetic studies of complex diseases. Many genes contribute to Alzheimer disease risk, but the genes contributing to risk can vary from population to population and from person to person, and each gene often has a small effect. In addition to multiple genes at play, there can be multiple variants within a gene contributing to Alzheimer disease risk. Hence, the search for genes affecting Alzheimer disease risk has been long and difficult. With a couple of exceptions, the current list of known Alzheimer disease genes was discovered by either linkage analysis or genome-wide association studies (GWAS). Linkage studies were successful in large pedigrees with early-onset Alzheimer disease and also worked to localize *APOE* for late-onset Alzheimer disease (LOAD). GWAS worked for common variants with population-level effects. In this work we combined both of these approaches, linkage analysis in pedigrees and GWAS, with the added advantage of implementing these approaches in an isolated population, the Amish communities of Ohio and Indiana, to minimize heterogeneity.

An important first step of any study is to ensure quality of the data before any analyses begin. In Chapter II I presented the quality control procedures that were

implemented on 827 Amish individuals with genotypes for 906,598 SNPs (single nucleotide polymorphisms) from the Affymetrix Genome-wide SNP Array 6.0. After removing individuals with low genotyping efficiency, individuals with suspected gender errors, a duplicate sample, and individuals misconnected into the pedigree, 798 samples (109 with LOAD) remained for analysis. Elimination of SNPs with low genotyping efficiency and low minor allele frequency, including minor allele frequency adjusted for pedigree relationships, left 614,963 high quality SNPs for analysis.

Using the cleaned dataset we were able to perform genome-wide linkage and association analyses, which were presented in Chapter III. In Chapter III we also specifically genotyped *APOE* for the E2, E3, and E4 alleles and analyzed the results using both association and linkage analyses. We found that in Holmes, Elkhart, and LaGrange Counties, LOAD is significantly associated with *APOE* ($p=9x10^{-6}$). However, in Adams County, where the E4 allele is much more rare, LOAD is not associated with *APOE* (p=0.55). Surprisingly, we saw no evidence for linkage between LOAD and *APOE*. This result confirms the need for both linkage and association analyses when studying a complex phenotype in the Amish.

Genome-wide analyses resulted in several novel LOAD loci, with the most notable on 2p12 which reached an HLOD of 6.14 under the recessive multipoint linkage model and an HLOD of 6.05 under the dominant multipoint linkage model. Other loci reaching HLOD scores greater than 3 were detected on 3q26, 9q31, and 18p11. Converging linkage and association results, the most significantly associated SNP under the 2p12 peak was at rs2974151 ($p=1.24x10^{-4}$). This SNP is located in *CTNNA2*, which encodes catenin alpha 2, a neuronal-specific catenin known to play roles in the developing brain and possibly also in the aging brain.

Although locus heterogeneity is less of an issue in the Amish, the results from the genome-wide analyses demonstrated that heterogeneity was not completely avoided. Each of the linkage peaks only had a subset of subpedigrees showing evidence for the linkage result, and association with *APOE* was not seen in Adams County. Despite the complexity, multiple novel loci were implicated by the analyses, but the most striking result was for *CTNNA2*, which led us to analyze the sequence of the gene to attempt to identify the causal variant.

As discussed in Chapter IV, the sequence analysis of *CTNNA2* and its nested gene, *LRRTM1*, in 47 LOAD and 54 cognitively normal indivduals identified nine (two in *CTNNA2*, seven in *LRRTM1*) exonic single nucleotide variants (SNVs). The two *CTNNA2* exonic SNVs and two of the seven *LRRTM1* SNVs were previously recorded in dbSNP. None of the exonic SNVs showed notable allele frequency differences in LOAD individuals and cognitively normal individuals. Although the sequence of *CTNNA2* and *LRRTM1* were obtained via whole exome sequencing analysis, some noncoding sequence was also captured and was analyzed, which led to the identification of the intronic SNV rs72822556. This SNP previously had not been genotyped in our genome-wide study. Followup genotyping of rs72822556 disproved many of the whole-exome-generated genotypes and weakened the significance of this variant. Better coverage of the noncoding regions of *CTNNA2* is needed before involvement of *CTNNA2* in LOAD risk can be ruled out. These results attest to the the complexity of the genetic architecture of LOAD and the need for multiple approaches to uncover the missing heritability of LOAD.

## Future Directions

Coupled with the rapid development of genomic technologies, the recent expansion of our knowledge of Alzheimer disease genetics, including the work presented in this thesis, opens the door for future discoveries of Alzheimer disease genetics. The work presented in this thesis contributes substantially to our understanding of the underlying locus heterogeneity in LOAD. Multiple approaches will need to be taken to capture the various genetic effects contributing to Alzheimer disease risk. Various study populations, genetic variations, and analysis techniques will need to be incorporated in future studies.

Additional genes in the 2p12 region should be explored using the existing whole exome sequence data to identify other variants that could be contributing to the significant linkage peak. The genes and variants in the other regions of linkage from the genome wide study (Chapter III), 3q26, 9q31, and 18p11, should also be investigated with the exome data. In addition, the rest of the exomes could also be screened for rare variants that might not have been well-tagged by the genome-wide SNP data. Of course, the exome data excludes most of the noncoding regions of the genome, and it is likely that at least some of the genetic risk to LOAD lies in the noncoding regions. In particular, the noncoding regions of *CTNNA2* should be screened. Whole genome sequencing is becoming more accessible and should be employed to study noncoding variation in the Amish genomes. In addition to single-nucleotide variations, other genetic variations such as insertions and deletions need to be queried in both whole exome and whole genome sequence data, starting with *CTNNA2* and in the entire genome. Sequence data also provides opportunities to characterize the genomes of the

Amish individuals to better understand variation and patterns of linkage disequilibrium in the population. For instance a comparison of the variation and linkage disequilibrium in the Adams County community verses the communities in other counties would be a worthwhile endeavor since we have already seen such a difference in our *APOE* results. This type of characterization could be expanded to other Amish communities and other related populations.

Any variants found to be significantly associated with or linked to Alzheimer disease in our Amish dataset, or a subset of the Amish dataset, could be examined in other Amish populations, such as the Amish communities in Pennsylvania, to see how frequent the variant is and if the association with or linkage to Alzheimer disease is characteristic of other Amish communities. Then the attempt to replicate the effect of the variant should be carried out in other populations, starting with populations of European descent and expanding to other populations to see how generalizable the effect of the variant is. The replication studies should include other variants in the gene and even other genes in the same pathway since the same gene or pathway could be affecting multiple populations even if the same variant is not present.

Additional technological advances will also aid in the search for Alzheimer disease genes. While we cannot get autopsy data from the Amish, advances in brain imaging technology could allow for more accurate diagnoses in the future to make genetic studies even more productive. Imaging data can provide more confident Alzheimer disease diagnoses and identify controls and individuals with 'unclear' diagnoses who appear to be heading toward Alzheimer disease status. For example, we could identify MCI (mild cognitive impairment) patients who have pathological evidence of Alzheimer disease and include them as cases in some of the analyses.

Imaging analysis can help create even more homogeneous case and control sets which could help tease out some of the genetic heterogeneity. These advancements will be helpful for studying all populations, not just the Amish. Replicating genetic effects between populations could be more productive with the better case and control definitions. Imaging analysis can also lead to interesting longitudinal studies, in which younger control subjects are followed over time to compare the genetic profiles of those who progress to Alzheimer disease and those who do not.

More work is also needed in the area of analysis methods. The complexity of the Amish family structure complicates analyses. Our current methods could be limiting our ability to detect significant genetic variation contributing to Alzheimer disease. The work presented in this thesis was limited to single-variant analyses. Therefore, finding ways to include gene-gene and gene-environment interactions would strengthen future studies. We also know that splitting the pedigree prior to linkage analysis could affect power and type 1 error. Simulations of the pedigree structure to evaluate these possible affects can help direct future analyses, for instance, by optimizing parameters used. Better ways of splitting the pedigree could be explored, and ideally, better linkage programs to able to deal with bigger and more complicated pedigree structures would greatly enhance this research.

The currently known list of Alzheimer disease risk genes plus other soon-to-be identified genes and other genetic variations will need to be functionally studied to understand the impact on gene expression and the downstream Alzheimer disease pathophysiology that could lead to possible therapeutic interventions to prevent or slow the progression of Alzheimer disease. As geneticists we should strive for our work to go from computer to bench and from bench to bedside.

**Appendix A. Most significant genome-wide association results, stratified.** Most significant genome-wide association results (see Table 4) calculated for Adams county individuals separately from non-Adams County (Elkhart, Lagrange, and Holmes Counties)

| SNP | Adams Minor Allele | Adams Affected MAF | Adams Unaffected MAF | Adams MQLS *P*-value | non-Adams Minor Allele | non-Adams Affected MAF | non-Adams Unaffected MAF | non-Adams MQLS *P*-value |
|---|---|---|---|---|---|---|---|---|
| rs4145462 | T | 0.17 | 0.01 | 1.06E-02 | T | 0.14 | 0.09 | 1.54E-03 |
| rs41458646 | G | 0.33 | 0.19 | 1.49E-01 | G | 0.33 | 0.22 | 2.92E-04 |
| rs41476545 | G | 0.33 | 0.19 | 1.49E-01 | G | 0.33 | 0.22 | 3.14E-04 |
| rs6738181 | A | 0.11 | 0.17 | 9.35E-01 | A | 0.40 | 0.23 | 4.61E-06 |
| rs7638995 | A | 0.06 | 0.05 | 8.15E-01 | A | 0.25 | 0.15 | 2.10E-05 |
| rs679974 | C | 0.28 | 0.08 | 2.90E-02 | C | 0.19 | 0.10 | 1.17E-03 |
| rs11983798 | T | 0.28 | 0.08 | 3.00E-02 | T | 0.19 | 0.10 | 2.13E-04 |
| rs6468852 | G | 0.11 | 0.05 | 8.75E-01 | G | 0.30 | 0.17 | 8.97E-06 |
| rs9969729 | A | 0.11 | 0.05 | 8.70E-01 | A | 0.21 | 0.13 | 7.74E-04 |
| rs12361953 | C | 0.11 | 0.13 | 9.15E-01 | C | 0.32 | 0.19 | 2.61E-05 |
| rs472926 | C | 0.06 | 0.14 | 8.92E-01 | C | 0.31 | 0.19 | 2.11E-05 |
| rs4937314 | C | 0.28 | 0.16 | 3.76E-02 | C | 0.34 | 0.24 | 1.91E-04 |
| rs11848070 | C | 0.39 | 0.16 | 3.44E-02 | C | 0.43 | 0.30 | 1.76E-04 |
| rs17767225 | T | 0.39 | 0.24 | 1.63E-01 | T | 0.36 | 0.24 | 1.19E-04 |
| rs6085820 | A | 0.06 | 0.01 | 4.67E-01 | A | 0.22 | 0.13 | 2.37E-04 |

**Appendix B. Regions with at least one SNP with a two-point HLOD ≥ 3. These regions were analyzed with multipoint linkage analysis.**

| Region | chr | Mbp | SNP | 2pt DOM | 2pt REC | 2pt Npall | 2pt Nppairs | MQLS p-value |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 83.66 | rs17466903 | 3.60 | 3.63 | 1.07 | 0.78 | 0.85 |
| | | | | | | | | |
| 2 | 1 | 161.11 | rs1986957 | 2.55 | 3.05 | 0.78 | 0.57 | 0.92 |
| | | | | | | | | |
| 3 | 1 | 237.33 | rs16837761 | 2.68 | 3.03 | 1.04 | 0.80 | 0.37 |
| | 1 | 237.63 | rs10925877 | 2.83 | 3.61 | 1.26 | 0.99 | 0.19 |
| | 1 | 237.90 | rs7541783 | 2.76 | 3.03 | 1.74 | 1.38 | 0.42 |
| | 1 | 238.21 | rs1415277 | 2.19 | 3.16 | 1.28 | 1.08 | 0.95 |
| | 1 | 238.24 | rs2065914 | 4.10 | 2.36 | 1.41 | 0.99 | 0.95 |
| | 1 | 238.24 | rs10926054 | 3.86 | 2.18 | 1.32 | 0.93 | 0.98 |
| | 1 | 238.24 | rs12130657 | 4.09 | 2.36 | 1.41 | 0.99 | 0.94 |
| | 1 | 238.24 | rs10926061 | 4.11 | 2.37 | 1.41 | 0.99 | 0.98 |
| | 1 | 238.24 | rs16839333 | 4.11 | 2.37 | 1.41 | 0.99 | 0.96 |
| | 1 | 238.88 | rs16840459 | 2.67 | 3.47 | 0.98 | 0.76 | 0.38 |
| | | | | | | | | |
| 4 | 2 | 52.53 | rs4588226 | 1.51 | 3.01 | 0.08 | 0.08 | 0.38 |
| | 2 | 52.55 | rs12713204 | 1.63 | 3.19 | 0.19 | 0.17 | 0.22 |
| | | | | | | | | |
| 5 | 2 | 62.38 | rs7578484 | 2.25 | 3.81 | 0.84 | 0.67 | 0.09 |
| | 2 | 69.01 | rs10197208 | 3.00 | 2.56 | 0.87 | 0.58 | 0.52 |
| | 2 | 69.83 | rs10177224 | 3.68 | 2.78 | 1.39 | 1.03 | 0.25 |
| | 2 | 71.70 | rs11897583 | 1.95 | 3.39 | 0.78 | 0.64 | 0.24 |
| | 2 | 73.71 | rs4852939 | 3.02 | 1.25 | 1.21 | 0.89 | 0.17 |
| | 2 | 73.82 | rs11894953 | 3.10 | 1.07 | 0.78 | 0.61 | 0.27 |
| | 2 | 74.16 | rs12991192 | 3.19 | 3.77 | 1.04 | 0.76 | 0.01 |
| | 2 | 74.17 | rs7593050 | 3.05 | 2.38 | 0.65 | 0.55 | 0.03 |
| | 2 | 80.66 | rs216616 | 2.85 | 3.37 | 1.09 | 0.86 | 0.60 |
| | 2 | 85.80 | rs6759087 | 3.79 | 3.26 | 0.93 | 0.63 | 0.18 |
| | 2 | 85.80 | rs11126997 | 3.79 | 3.26 | 0.93 | 0.63 | 0.18 |
| | 2 | 86.15 | rs6735014 | 1.99 | 3.06 | 0.87 | 0.74 | 0.13 |
| | 2 | 86.23 | rs4569473 | 2.16 | 3.29 | 0.90 | 0.76 | 0.07 |
| | 2 | 89.88 | rs13003799 | 2.32 | 3.59 | 0.99 | 0.81 | 0.66 |
| | 2 | 101.26 | rs6713930 | 2.05 | 3.10 | 0.49 | 0.37 | 0.13 |
| | 2 | 101.29 | rs17190412 | 2.43 | 3.52 | 0.83 | 0.65 | 0.35 |
| | 2 | 102.22 | rs11692230 | 3.36 | 2.41 | 1.10 | 0.82 | 0.82 |
| | | | | | | | | |
| 6 | 2 | 111.51 | rs3789085 | 2.70 | 3.81 | 0.97 | 0.73 | 0.18 |
| | | | | | | | | |

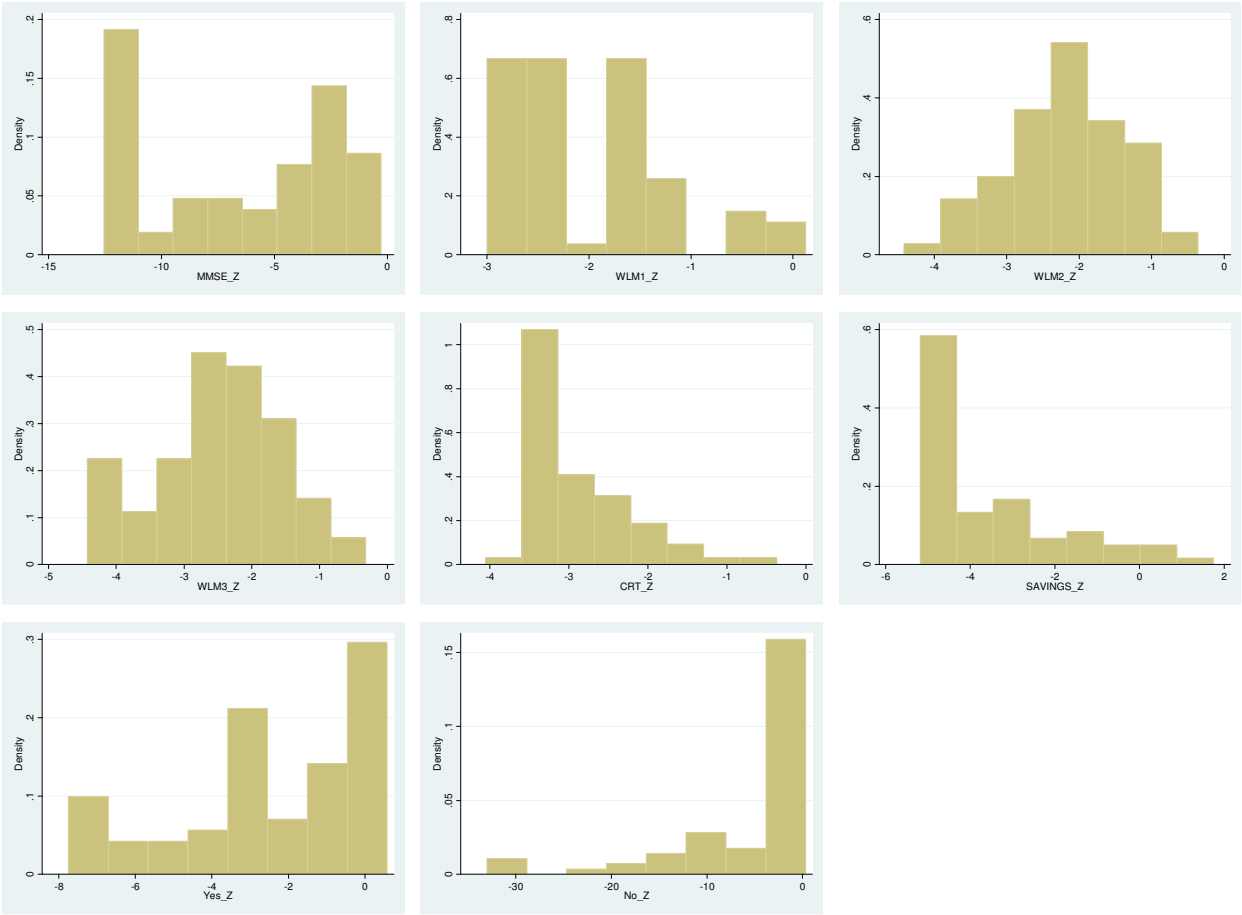
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 7 | 2 | 124.90 | rs17725619 | 2.74 | 3.42 | 1.14 | 0.89 | 0.35 |
| | 2 | 129.17 | rs6431018 | 1.90 | 3.02 | 0.60 | 0.44 | 0.67 |
| | | | | | | | | |
| 8 | 2 | 174.29 | rs12329372 | 2.51 | 3.73 | 0.80 | 0.62 | 0.79 |
| | 2 | 174.29 | rs13022741 | 2.51 | 3.63 | 0.76 | 0.58 | 0.36 |
| | 2 | 174.30 | rs6433434 | 2.58 | 3.13 | 0.65 | 0.42 | 0.67 |
| | 2 | 174.31 | rs16823293 | 2.66 | 3.23 | 0.74 | 0.48 | 0.68 |
| | 2 | 178.24 | rs13028757 | 2.98 | 3.39 | 1.31 | 0.97 | 0.65 |
| | 2 | 178.25 | rs7586934 | 2.98 | 3.39 | 1.31 | 0.97 | 0.65 |
| | 2 | 178.25 | rs7560737 | 2.98 | 3.39 | 1.31 | 0.97 | 0.65 |
| | 2 | 178.25 | rs6433688 | 2.98 | 3.39 | 1.31 | 0.97 | 0.65 |
| | 2 | 180.80 | rs16867269 | 1.97 | 3.12 | 1.28 | 0.96 | 0.84 |
| | | | | | | | | |
| 9 | 2 | 195.74 | rs7423326 | 2.44 | 3.09 | 0.85 | 0.62 | 0.81 |
| | 2 | 195.75 | rs6434742 | 2.47 | 3.13 | 0.85 | 0.61 | 0.79 |
| | 2 | 195.75 | rs7593916 | 2.44 | 3.09 | 0.85 | 0.62 | 0.82 |
| | 2 | 195.75 | rs7422772 | 2.44 | 3.09 | 0.85 | 0.62 | 0.81 |
| | 2 | 195.75 | rs7424669 | 2.46 | 3.12 | 0.86 | 0.62 | 0.77 |
| | 2 | 195.75 | rs6434746 | 2.44 | 3.09 | 0.85 | 0.62 | 0.81 |
| | 2 | 195.81 | rs1858305 | 2.35 | 3.05 | 0.90 | 0.65 | 0.65 |
| | | | | | | | | |
| 10 | 2 | 225.46 | rs2304335 | 3.49 | 4.21 | 0.90 | 0.65 | 0.46 |
| | | | | | | | | |
| 11 | 3 | 1.77 | rs4432626 | 2.37 | 3.36 | 0.96 | 0.72 | 0.69 |
| | 3 | 1.90 | rs4685457 | 3.87 | 3.13 | 1.02 | 0.71 | 0.33 |
| | 3 | 1.91 | rs1844171 | 3.60 | 2.87 | 0.96 | 0.66 | 0.41 |
| | 3 | 1.99 | rs2171596 | 3.08 | 1.88 | 0.89 | 0.69 | 0.54 |
| | 3 | 2.00 | rs10510224 | 4.50 | 3.13 | 0.98 | 0.70 | 0.63 |
| | 3 | 2.01 | rs7611355 | 2.89 | 3.17 | 1.02 | 0.77 | 0.06 |
| | 3 | 2.01 | rs6767479 | 2.90 | 3.18 | 1.02 | 0.77 | 0.05 |
| | 3 | 3.28 | rs4685622 | 2.19 | 3.21 | 0.54 | 0.46 | 0.15 |
| | 3 | 3.63 | rs1601875 | 1.47 | 3.01 | 0.71 | 0.57 | 0.25 |
| | | | | | | | | |
| 12 | 3 | 57.68 | rs6790054 | 2.59 | 3.06 | 1.03 | 0.81 | 0.67 |
| | | | | | | | | |
| 13 | 3 | 79.68 | rs9820160 | 1.96 | 3.13 | 0.87 | 0.65 | 0.12 |
| | 3 | 80.17 | rs4635670 | 2.38 | 3.33 | 0.77 | 0.58 | 0.00 |
| | 3 | 80.20 | rs12636593 | 1.33 | 3.08 | 0.80 | 0.70 | 0.38 |
| | 3 | 80.83 | rs1437042 | 2.82 | 3.20 | 1.01 | 0.76 | 0.09 |
| | 3 | 80.92 | rs17018312 | 2.82 | 3.20 | 1.01 | 0.76 | 0.09 |
| | 3 | 81.02 | rs13060424 | 2.79 | 3.17 | 1.01 | 0.75 | 0.08 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 3 | 85.59 | rs13323436 | 2.12 | 3.04 | 0.91 | 0.66 | 0.25 |
| | 3 | 85.74 | rs1449399 | 2.29 | 3.14 | 1.06 | 0.85 | 0.04 |
| | 3 | 85.77 | rs11926266 | 2.80 | 3.72 | 1.27 | 1.00 | 0.03 |
| | 3 | 85.77 | rs4507269 | 2.79 | 3.72 | 1.25 | 0.99 | 0.03 |
| | | | | | | | | |
| 14 | 3 | 173.13 | rs4894786 | 3.39 | 2.62 | 0.83 | 0.58 | 0.18 |
| | 3 | 173.14 | rs9862319 | 3.16 | 2.49 | 0.80 | 0.56 | 0.19 |
| | | | | | | | | |
| 15 | 4 | 139.99 | rs17268257 | 3.17 | 3.51 | 1.31 | 0.91 | 0.01 |
| | 4 | 140.08 | rs13125601 | 3.16 | 2.14 | 0.91 | 0.58 | 0.43 |
| | 4 | 140.08 | rs11734771 | 3.16 | 2.14 | 0.91 | 0.58 | 0.43 |
| | 4 | 140.09 | rs6844552 | 3.16 | 2.14 | 0.91 | 0.58 | 0.43 |
| | 4 | 140.09 | rs4076773 | 3.16 | 2.14 | 0.91 | 0.58 | 0.43 |
| | | | | | | | | |
| 16 | 4 | 158.14 | rs1443230 | 2.26 | 3.06 | 0.93 | 0.71 | 0.62 |
| | 4 | 158.18 | rs7668059 | 2.51 | 3.33 | 1.03 | 0.78 | 0.61 |
| | 4 | 158.21 | rs6812324 | 2.51 | 3.21 | 1.06 | 0.79 | 0.90 |
| | | | | | | | | |
| 17 | 5 | 25.81 | rs1479675 | 3.49 | 2.11 | 1.03 | 0.77 | 0.01 |
| | 5 | 25.81 | rs12517113 | 3.23 | 1.91 | 0.90 | 0.68 | 0.02 |
| | | | | | | | | |
| 18 | 5 | 152.92 | rs716517 | 2.11 | 3.20 | 0.83 | 0.68 | 0.75 |
| | | | | | | | | |
| 19 | 5 | 173.62 | rs17077144 | 3.21 | 2.99 | 1.07 | 0.76 | 0.01 |
| | 5 | 173.62 | rs1368273 | 3.20 | 2.99 | 1.06 | 0.75 | 0.01 |
| | | | | | | | | |
| 20 | 6 | 7.47 | rs1408482 | 4.17 | 2.89 | 0.98 | 0.60 | 0.10 |
| | 6 | 7.48 | rs11243192 | 3.67 | 1.60 | 0.72 | 0.41 | 0.12 |
| | 6 | 7.48 | rs9328430 | 3.74 | 1.75 | 0.74 | 0.43 | 0.11 |
| | 6 | 7.48 | rs1322219 | 3.78 | 1.77 | 0.75 | 0.43 | 0.09 |
| | | | | | | | | |
| 21 | 6 | 80.77 | rs466335 | 2.03 | 3.47 | 0.81 | 0.67 | 0.99 |
| | | | | | | | | |
| 22 | 6 | 92.63 | rs9294513 | 2.96 | 4.11 | 0.96 | 0.70 | 0.05 |
| | | | | | | | | |
| 23 | 6 | 114.42 | rs717389 | 2.28 | 3.20 | 0.97 | 0.74 | 0.06 |
| | | | | | | | | |
| 24 | 6 | 151.32 | rs6557106 | 3.20 | 1.56 | 0.96 | 0.69 | 0.99 |
| | | | | | | | | |
| 25 | 7 | 4.34 | rs4425656 | 3.35 | 1.38 | 0.85 | 0.59 | 0.03 |
| | | | | | | | | |

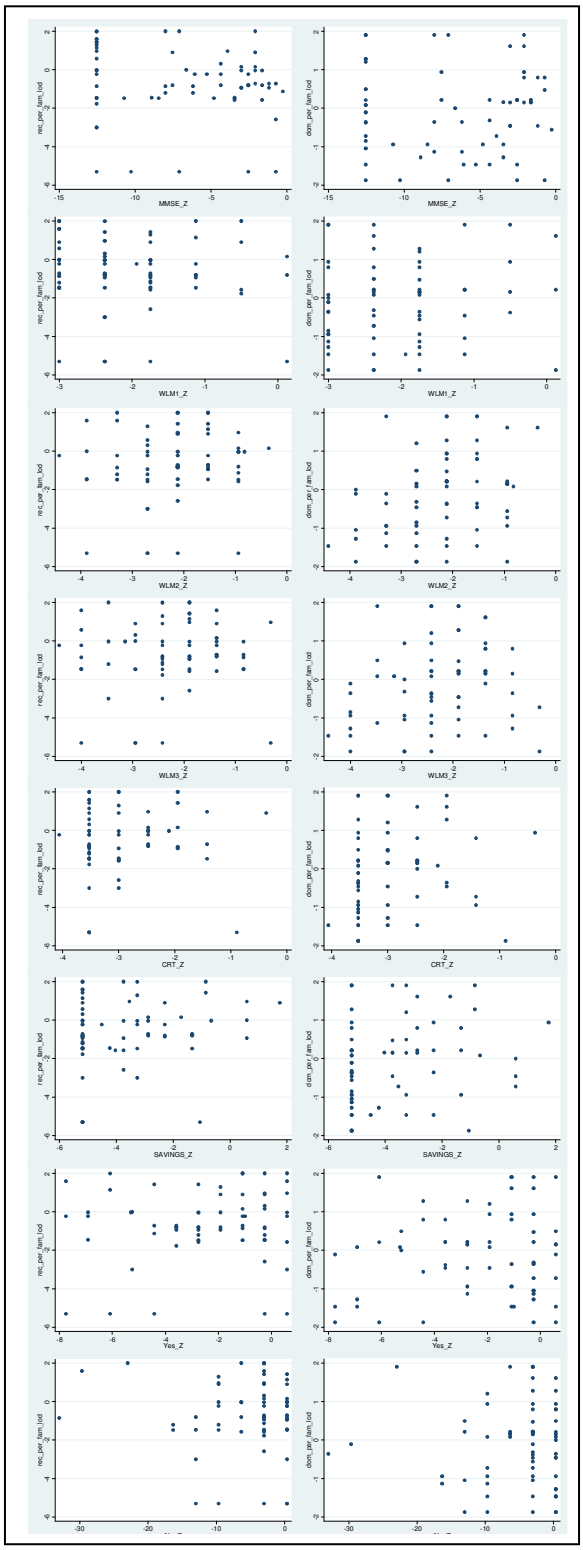
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 26 | 7 | 48.04 | rs2686782 | 2.02 | 3.19 | 1.03 | 0.81 | 0.89 |
| | 7 | 48.05 | rs1879830 | 2.06 | 3.28 | 1.04 | 0.82 | 0.67 |
| | 7 | 48.05 | rs10233232 | 2.09 | 3.31 | 1.05 | 0.83 | 0.73 |
| | 7 | 48.06 | rs2686792 | 2.05 | 3.27 | 1.05 | 0.83 | 0.76 |
| | | | | | | | | |
| 27 | 7 | 104.13 | rs10269217 | 1.71 | 3.36 | 0.70 | 0.59 | 0.03 |
| | | | | | | | | |
| 28 | 8 | 9.19 | rs747751 | 2.56 | 3.71 | 0.88 | 0.68 | 0.76 |
| | 8 | 12.96 | rs607499 | 3.21 | 2.85 | 1.05 | 0.80 | 0.74 |
| | 8 | 13.62 | rs10094983 | 2.65 | 3.57 | 0.97 | 0.79 | 0.45 |
| | 8 | 15.90 | rs17580109 | 3.00 | 3.19 | 0.82 | 0.58 | 0.85 |
| | 8 | 18.08 | rs17126329 | 1.76 | 3.47 | 0.78 | 0.65 | 0.13 |
| | | | | | | | | |
| 29 | 9 | 26.05 | rs957252 | 1.91 | 4.03 | 1.25 | 1.02 | 0.35 |
| | | | | | | | | |
| 30 | 9 | 78.49 | rs669296 | 3.08 | 2.20 | 0.72 | 0.51 | 0.06 |
| | 9 | 81.92 | rs7850306 | 3.90 | 2.09 | 1.08 | 0.77 | 0.13 |
| | | | | | | | | |
| 31 | 9 | 95.80 | rs10993017 | 3.14 | 2.78 | 1.00 | 0.69 | 0.39 |
| | | | | | | | | |
| 32 | 9 | 100.80 | rs10988521 | 2.57 | 3.73 | 0.92 | 0.69 | 0.14 |
| | 9 | 101.79 | rs1852865 | 3.10 | 4.16 | 0.85 | 0.68 | 0.98 |
| | 9 | 101.79 | rs10760710 | 3.10 | 4.18 | 0.85 | 0.68 | 0.98 |
| | 9 | 101.89 | rs1529192 | 3.10 | 4.17 | 0.85 | 0.68 | 0.97 |
| | 9 | 101.92 | rs7846794 | 3.03 | 3.08 | 0.74 | 0.57 | 0.97 |
| | 9 | 101.92 | rs7861003 | 2.99 | 3.87 | 0.84 | 0.67 | 0.94 |
| | 9 | 101.97 | rs2476441 | 2.86 | 3.89 | 0.83 | 0.67 | 0.88 |
| | 9 | 101.98 | rs9886877 | 2.52 | 3.08 | 0.75 | 0.60 | 0.88 |
| | 9 | 102.02 | rs1014652 | 2.75 | 3.59 | 0.81 | 0.66 | 0.86 |
| | 9 | 102.03 | rs2787365 | 2.69 | 3.72 | 0.80 | 0.65 | 0.89 |
| | 9 | 102.04 | rs2787397 | 2.51 | 3.42 | 0.78 | 0.64 | 1.00 |
| | 9 | 102.07 | rs2806693 | 2.86 | 3.89 | 0.83 | 0.67 | 0.86 |
| | 9 | 102.07 | rs2485745 | 2.86 | 3.89 | 0.83 | 0.67 | 0.86 |
| | 9 | 105.76 | rs10512321 | 2.88 | 3.50 | 0.77 | 0.50 | 0.80 |
| | 9 | 108.58 | rs4625092 | 1.97 | 3.22 | 0.64 | 0.50 | 0.45 |
| | 9 | 110.09 | rs1570504 | 2.26 | 3.18 | 0.75 | 0.56 | 0.71 |
| | 9 | 111.67 | rs10816902 | 3.12 | 3.53 | 0.66 | 0.50 | 0.22 |
| | | | | | | | | |
| 33 | 9 | 128.50 | rs3861878 | 3.00 | 0.95 | 0.73 | 0.49 | 0.51 |
| | 9 | 128.51 | rs2417033 | 3.26 | 1.22 | 1.04 | 0.74 | 0.07 |
| | 9 | 128.52 | rs10760452 | 3.24 | 1.20 | 1.03 | 0.74 | 0.11 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 34 | 10 | 29.81 | rs913034 | 3.32 | 1.29 | 0.84 | 0.58 | 0.47 |
| | 10 | 29.81 | rs7096453 | 3.41 | 1.42 | 0.85 | 0.59 | 0.49 |
| | | | | | | | | |
| 35 | 10 | 127.91 | rs10466250 | 2.34 | 3.32 | 0.94 | 0.72 | 0.38 |
| | | | | | | | | |
| 36 | 11 | 102.14 | rs11225417 | 3.02 | 3.19 | 0.84 | 0.59 | 0.67 |
| | 11 | 104.55 | rs1503389 | 2.88 | 3.29 | 0.62 | 0.40 | 0.61 |
| | | | | | | | | |
| 37 | 12 | 126.65 | rs12816855 | 3.05 | 1.48 | 0.62 | 0.45 | 0.81 |
| | | | | | | | | |
| 38 | 13 | 90.18 | rs1417853 | 3.19 | 1.77 | 0.56 | 0.41 | 0.29 |
| | | | | | | | | |
| 39 | 14 | 46.06 | rs8017002 | 3.29 | 2.20 | 1.17 | 0.90 | 0.15 |
| | 14 | 47.32 | rs2022567 | 2.38 | 3.25 | 0.99 | 0.78 | 0.15 |
| | | | | | | | | |
| 40 | 15 | 85.07 | rs4386109 | 2.65 | 3.28 | 1.01 | 0.74 | 0.44 |
| | | | | | | | | |
| 41 | 16 | 86.67 | rs11117362 | 2.39 | 3.28 | 0.85 | 0.65 | 0.94 |
| | | | | | | | | |
| 42 | 18 | 9.18 | rs7506291 | 2.58 | 3.09 | 0.94 | 0.76 | 0.12 |
| | 18 | 11.37 | rs1455237 | 2.07 | 3.02 | 1.03 | 0.80 | 0.60 |
| | 18 | 13.17 | rs4797730 | 3.36 | 2.49 | 1.27 | 0.95 | 0.22 |
| | | | | | | | | |
| 43 | 18 | 59.85 | rs1400569 | 3.42 | 3.10 | 1.18 | 0.88 | 0.10 |
| | 18 | 59.86 | rs11872249 | 3.21 | 2.83 | 1.21 | 0.90 | 0.15 |
| | 18 | 59.87 | rs213100 | 3.23 | 2.84 | 1.21 | 0.91 | 0.13 |
| | 18 | 59.87 | rs213097 | 3.23 | 2.84 | 1.22 | 0.91 | 0.15 |
| | 18 | 59.87 | rs213093 | 3.21 | 2.83 | 1.21 | 0.90 | 0.15 |
| | 18 | 59.87 | rs213086 | 3.20 | 2.82 | 1.21 | 0.90 | 0.15 |
| | | | | | | | | |
| 44 | 19 | 15.57 | rs8112423 | 3.20 | 2.18 | 0.68 | 0.53 | 0.67 |
| | 19 | 21.75 | rs10414913 | 2.25 | 3.05 | 0.54 | 0.39 | 0.01 |
| | | | | | | | | |
| 45 | 19 | 49.83 | rs203717 | 2.07 | 3.31 | 0.73 | 0.58 | 0.90 |
| | | | | | | | | |
| 46 | 20 | 8.54 | rs6118268 | 3.13 | 1.32 | 0.87 | 0.65 | 0.80 |

**Appendix C. Distributions of Z scores from the Mini-Mental State Exam (MMSE_Z)), Word List Memory trials 1-3, delayed recall, delayed recall, savings, recognition-yes, and recognition-no**



**Appendix D. Scatter plots of recessive (left) and dominant (right) per-family lod scores versus Z scores from the Mini-Mental State Exam (MMSE_Z)), Word List Memory trials 1-3, delayed recall, delayed recall, savings, recognition-yes, and recognition-no**



**Appendix E. Spearman's correlation between 2p12 lod scores and Z scores of Word List learning with delayed recall and recognition procedure with MMSE Z scores as a covariate.** Recessive LOD refers to the per-family lod scores at the peak of the 2p12 linkage region calculated under a recessive model, and dominant LOD refers to per-family lod scores at the peak of the 2p12 linkage region calculated under a dominant model. Spearman's rho is the correlation coefficient.

| | Recessive LOD | | Dominant LOD | |
|---|---|---|---|---|
| | Spearman's rho | p-value | Spearman's rho | p-value |
| Word List Memory Trial 1 Z | -0.08 | 0.53 | 0.14 | 0.26 |
| Word List Memory Trial 2 Z | 0.14 | 0.26 | 0.31 | **0.01** |
| Word List Memory Trial 3 Z | 0.1 | 0.43 | 0.25 | **0.04** |
| Delayed Recall Z | 0.14 | 0.24 | 0.28 | **0.02** |
| Savings Z | 0.19 | 0.13 | 0.27 | **0.03** |
| Recognition Yes Z | 0.05 | 0.67 | 0.06 | 0.62 |
| Recognition No Z | -0.008 | 0.95 | -0.02 | 0.86 |

REFERENCES


(1) Hebert LE, Scherr PA, Bienias JL, Bennett DA, Evans DA. Alzheimer disease in the US population: prevalence estimates using the 2000 census. Arch Neurol 2003 Aug;60(8):1119-22.

(2) Minino AM, Xu J, Kochanek KD. Deaths: Preliminary Data for 2008. National Vital Statistics Reports 2011;59.

(3) 2012 Alzheimer's disease facts and figures. Alzheimers Dement 2012 Mar;8(2):131-68.

(4) Goate AM, Chartier-Harlin MC, Mullan MC, Brown J, Crawford F, Fidani L, et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature 1991;33:53-6.

(5) Rogaev EI, Sherrington R, Rogaeva EA, Levesque G, Ikeda M, Liang G, et al. Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. Nature 1995;376(6543):775-8.

(6) Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, Pettingell WH, et al. Candidate gene for the chromosome 1 familial Alzheimer's disease locus. Science 1995;269:973-7.

(7) Sherrington R, Rogaev E, Liang Y, Rogaeva EA, Levesque G, Ikeda M, et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature 1995;375:754-60.

(8) Pericak-Vance MA, Yamaoka LH, Haynes CS, Speer MC, Haines JL, Gaskell PC, et al. Genetic linkage studies in Alzheimer's disease families. Experimental Neurology 1988;102(3):271-9.

(9) Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science 2002 Jul 19;297(5580):353-6.

(10) Li G, Higdon R, Kukull WA, Peskind E, Van Valen MK, Tsuang D, et al. Statin therapy and risk of dementia in the elderly: a community-based prospective cohort study. Neurology 2004 Nov 9;63(9):1624-8.

(11) McLean CA, Cherny RA, Fraser FW, Fuller SJ, Smith MJ, Beyreuther K, et al. Soluble pool of Abeta amyloid as a determinant of severity of neurodegeneration in Alzheimer's disease. Ann Neurol 1999 Dec;46(6):860-6.

(12) Santa-Maria I, Haggiagi A, Liu X, Wasserscheid J, Nelson PT, Dewar K, et al. The MAPT H1 haplotype is associated with tangle-predominant dementia. Acta Neuropathol 2012 Jul 17.

(13) Tiraboschi P, Sabbagh MN, Hansen LA, Salmon DP, Merdes A, Gamst A, et al. Alzheimer disease without neocortical neurofibrillary tangles: "a second look". Neurology 2004 Apr 13;62(7):1141-7.

(14) Hampel H, Frank R, Broich K, Teipel SJ, Katz RG, Hardy J, et al. Biomarkers for Alzheimer's disease: academic, industry and regulatory perspectives. Nat Rev Drug Discov 2010 Jul;9(7):560-74.

(15) Henry MS, Passmore AP, Todd S, McGuinness B, Craig D, Johnston JA. The development of effective biomarkers for Alzheimer's disease: a review. Int J Geriatr Psychiatry 2012 Jun 4.

(16) Jack CR, Jr., Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, et al. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011 May;7(3):257-62.

(17) McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 1984 Jul;34(7):939-44.

(18) Plassman BL, Khachaturian AS, Townsend JJ, Ball MJ, Steffens DC, Leslie CE, et al. Comparison of clinical and neuropathologic Diagnoses of Alzheimer's disease in 3 epidemiologic samples. Alzheimer's & Dementia 2006;2(1):2-11.

(19) Bertram L, Lill CM, Tanzi RE. The genetics of Alzheimer disease: back to the future. Neuron 2010 Oct 21;68(2):270-81.

(20) Chandra V, Pandav R, Laxminarayan R, Tanner C, Manyam B, Rajkumar S, et al. Neurological Disorders. 2006.

(21) Kivipelto M, Ngandu T, Fratiglioni L, Viitanen M, Kareholt I, Winblad B, et al. Obesity and vascular risk factors at midlife and the risk of dementia and Alzheimer disease. Arch Neurol 2005 Oct;62(10):1556-60.

(22) Solomon A, Kivipelto M, Wolozin B, Zhou J, Whitmer RA. Midlife serum cholesterol and increased risk of Alzheimer's and vascular dementia three decades later. Dement Geriatr Cogn Disord 2009;28(1):75-80.

(23) Raji CA, Ho AJ, Parikshak NN, Becker JT, Lopez OL, Kuller LH, et al. Brain structure and obesity. Hum Brain Mapp 2010 Mar;31(3):353-64.

(24) Yaffe K, Lindquist K, Schwartz AV, Vitartas C, Vittinghoff E, Satterfield S, et al. Advanced glycation end product level, diabetes, and accelerated cognitive aging. Neurology 2011 Oct 4;77(14):1351-6.

(25) Scarmeas N, Stern Y, Tang MX, Mayeux R, Luchsinger JA. Mediterranean diet and risk for Alzheimer's disease. Ann Neurol 2006 Jun;59(6):912-21.

(26) Potter GG, Plassman BL, Burke JR, Kabeto MU, Langa KM, Llewellyn DJ, et al. Cognitive performance and informant reports in the diagnosis of cognitive impairment and dementia in African Americans and whites. Alzheimers Dement 2009 Nov;5(6):445-53.

(27) Gurland BJ, Wilder DE, Lantigua R, Stern Y, Chen J, Killeffer EH, et al. Rates of dementia in three ethnoracial groups. Int J Geriatr Psychiatry 1999 Jun;14(6):481-93.

(28) Anstey KJ, von SC, Salim A, O'Kearney R. Smoking as a risk factor for dementia and cognitive decline: a meta-analysis of prospective studies. Am J Epidemiol 2007 Aug 15;166(4):367-78.

(29) Rusanen M, Kivipelto M, Quesenberry CP, Jr., Zhou J, Whitmer RA. Heavy smoking in midlife and long-term risk of Alzheimer disease and vascular dementia. Arch Intern Med 2011 Feb 28;171(4):333-9.

(30) Buchman AS, Boyle PA, Yu L, Shah RC, Wilson RS, Bennett DA. Total daily physical activity and the risk of AD and cognitive decline in older adults. Neurology 2012 Apr 24;78(17):1323-9.

(31) McDowell I, Xi G, Lindsay J, Tierney M. Mapping the connections between education and dementia. J Clin Exp Neuropsychol 2007 Feb;29(2):127-41.

(32) Roe CM, Xiong C, Miller JP, Morris JC. Education and Alzheimer disease without dementia: support for the cognitive reserve hypothesis. Neurology 2007 Jan 16;68(3):223-8.

(33) Stern Y. Cognitive reserve and Alzheimer disease. Alzheimer Dis Assoc Disord 2006 Jul;20(3 Suppl 2):S69-S74.

(34) Plassman BL, Havlik RJ, Steffens DC, Helms MJ, Newman TN, Drosdick D, et al. Documented head injury in early adulthood and risk of Alzheimer's disease and other dementias. Neurology 2000 Oct 24;55(8):1158-66.

(35) Lye TC, Shores EA. Traumatic brain injury as a risk factor for Alzheimer's disease: a review. Neuropsychol Rev 2000 Jun;10(2):115-29.

(36) Katzman R, Galasko DR, Saitoh T, Chen X, Pay MM, Booth A, et al. Apolipoprotein-epsilon4 and head trauma: Synergistic or additive risks? Neurology 1996 Mar;46(3):889-91.

(37) Tang MX, Maestre G, Tsai WY, Liu XH, Feng L, Chung WY, et al. Effect of age, ethnicity, and head injury on the association between APOE genotypes and Alzheimer's disease. Ann N Y Acad Sci 1996 Dec 16;802:6-15.

(38) Breitner JC, Silverman JM, Mohs RC, Davis KL. Familial aggregation in Alzheimer's disease: comparison of risk among first degree relatives of early- and late-onset cases, and among male and female relatives in a successive generation. Neurology 1988;38:207-12.

(39) Hirst C, Sadovnick AD, Yee IML. Familial risks for Alzheimer disease: data from an Alzheimer clinic population. Genetic Epidemiology 1994;11:365-74.

(40) Sadovnick AD, Irwin ME, Baird PA, Beattle BL. Genetic studies on an Alzheimer clinic population. Genetic Epidemiology 1989;6:663-43.

(41) Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry 2006 Feb;63(2):168-74.

(42) Bergem AL. Heredity in dementia of the Alzheimer type. Clin Genet 1994 Jul;46(1 Spec No):144-9.

(43) Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 1993 Aug 13;261(5123):921-3.

(44) Namba Y, Tamonaga M, Kawasaki H, Otomo E, Ikeda K. Apolipoprotein E immunoreactivity in cerebral amyloid deposits and neurofibrillary tangles in Alzheimer's disease and cru plaque amyloid in Creutzfeldt-Jakob disease. Brain Research 1991;541:163-6.

(45) Pericak-Vance MA, Bebout JL, Gaskell PC, Yamaoka LH, Hung WY, Alberts MJ, et al. Linkage studies in familial Alzheimer's disease: evidence for chromosome 19 linkage. Am J Hum Genet 1991;48:1034-50.

(46) Saunders AM, Strittmatter WJ, Breitner JC, Schmechel D, St George-Hyslop PH, Pericak-Vance MA, et al. Association of apolipoprotein E allele 4 with late-onset familial and sporadic Alzheimer's disease. Neurology 1993;43:1467-72.

(47) Strittmatter WJ, Saunders AM, Schmechel D, Pericak-Vance MA, Enghild J, Salvesen GS, et al. Apolipoprotein E: high avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer's disease. Proceedings of the National Academy of Sciences of the United States of America 1993;90:1977-81.

(48) Wisniewski T, Frangione B. Apolipoprotein E: a pathological chaperone protein in patients with cerebral and systemic amyloid. Neurosci Lett 1992 Feb 3;135(2):235-8.

(49) Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Gaskell PC, Jr., et al. Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. Nat Genet 1994 Jun;7(2):180-4.

(50) Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. JAMA 1997 Oct 22;278(16):1349-56.

(51) Locke PA, Conneally PM, Tanzi RE, Gusella JF, Haines JL. Apolipoprotein E4 allele and Alzheimer disease: examination of allelic association and effect on age at onset in both early- and late-onset cases. Genet Epidemiol 1995;12(1):83-92.

(52) So HC, Gui AH, Cherny SS, Sham PC. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol 2011 Jul;35(5):310-7.

(53) Ashley-Koch AE, Shao Y, Rimmler JB, Gaskell PC, Welsh-Bohmer KA, Jackson CE, et al. An autosomal genomic screen for dementia in an extended Amish family. Neurosci Lett 2005 May 13;379(3):199-204.

(54) Blacker D, Bertram L, Saunders AJ, Moscarillo TJ, Albert MS, Wiener H, et al. Results of a high-resolution genome screen of 437 Alzheimer's disease families. Hum Mol Genet 2003 Jan 1;12(1):23-32.

(55) Farrer LA, Bowirrat A, Friedland RP, Waraska K, Korczyn AD, Baldwin CT. Identification of multiple loci for Alzheimer disease in a consanguineous Israeli-Arab community. Hum Mol Genet 2003 Feb 15;12(4):415-22.

(56) Hahs DW, McCauley JL, Crunk AE, McFarland LL, Gaskell PC, Jiang L, et al. A genome-wide linkage analysis of dementia in the Amish. Am J Med Genet B Neuropsychiatr Genet 2006 Mar 5;141(2):160-6.

(57) Kehoe P, Wavrant-De VF, Crook R, Wu WS, Holmans P, Fenton I, et al. A full genome scan for late onset Alzheimer's disease. Hum Mol Genet 1999 Feb;8(2):237-45.

(58) Mayeux R, Lee JH, Romas SN, Mayo D, Santana V, Williamson J, et al. Chromosome-12 mapping of late-onset Alzheimer disease among Caribbean Hispanics. Am J Hum Genet 2002 Jan;70(1):237-43.

(59) Mayeux R. Dissecting the relative influences of genes and the environment in Alzheimer's disease. Ann Neurol 2004 Feb;55(2):156-8.

(60) Myers A, Wavrant De-Vrieze F, Holmans P, Hamshere M, Crook R, Compton D, et al. Full genome screen for Alzheimer disease: stage II analysis. Am J Med Genet 2002 Mar 8;114(2):235-44.

(61) Pericak-Vance MA, Bass MP, Yamaoka LH, Gaskell PC, Scott WK, Terwedow HA, et al. Complete genomic screen in late-onset familial Alzheimer disease: evidence for a new locus on chromosome 12. JAMA 1997 Oct 15;278(15):1237-41.

(62) Pericak-Vance MA, Grubber J, Bailey LR, Hedges D, West S, Santoro L, et al. Identification of novel genes in late-onset Alzheimer's disease. Exp Gerontol 2000 Dec;35(9-10):1343-52.

(63) Hiltunen M, Mannermaa A, Thompson D, Easton D, Pirskanen M, Helisalmi S, et al. Genome-wide linkage disequilibrium mapping of late-onset Alzheimer's disease in Finland. Neurology 2001 Nov 13;57(9):1663-8.

(64) Bertram L, Tanzi RE. Alzheimer's disease: one disorder, too many genes? Hum Mol Genet 2004 Apr 1;13 Spec No 1:R135-R141.

(65) International hapmap consortium. The International HapMap Project. Nature 2003 Dec 18;426(6968):789-96.

(66) International hapmap consortium. A haplotype map of the human genome. Nature 2005 Oct 27;437(7063):1299-320.

(67) Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007 Oct 18;449(7164):851-61.

(68) Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, Morgan A, et al. Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum Mol Genet 2007 Apr 15;16(8):865-73.

(69) Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, Lince DH, et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. J Clin Psychiatry 2007 Apr;68(4):613-8.

(70) Abraham R, Moskvina V, Sims R, Hollingworth P, Morgan A, Georgieva L, et al. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. BMC Med Genomics 2008;1:44.

(71) Beecham GW, Martin ER, Li YJ, Slifer MA, Gilbert JR, Haines JL, et al. Genome-wide association study implicates a chromosome 12 risk locus for late-onset Alzheimer disease. Am J Hum Genet 2009 Jan;84(1):35-43.

(72) Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, Hogan MF, et al. Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. Am J Hum Genet 2008 Nov;83(5):623-32.

(73) Carrasquillo MM, Zou F, Pankratz VS, Wilcox SL, Ma L, Walker LP, et al. Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nat Genet 2009 Feb;41(2):192-8.

(74) Feulner TM, Laws SM, Friedrich P, Wagenpfeil S, Wurst SH, Riehle C, et al. Examination of the current top candidate genes for AD in a genome-wide association study. Mol Psychiatry 2010 Jul;15(7):756-66.

(75) Li H, Wetten S, Li L, St Jean PL, Upmanyu R, Surh L, et al. Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. Arch Neurol 2008 Jan;65(1):45-53.

(76) Poduslo SE, Huang R, Huang J, Smith S. Genome screen of late-onset Alzheimer's extended pedigrees identifies TRPC4AP by haplotype analysis. Am J Med Genet B Neuropsychiatr Genet 2009 Jan 5;150B(1):50-5.

(77) Reiman EM, Webster JA, Myers AJ, Hardy J, Dunckley T, Zismann VL, et al. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. Neuron 2007 Jun 7;54(5):713-20.

(78) Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet 2009 Oct;41(10):1088-93.

(79) Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet 2009 Oct;41(10):1094-9.

(80) Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, et al. Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA 2010 May 12;303(18):1832-40.

(81) Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet 2011 Apr 3.

(82) Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo MM, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet 2011 May;43(5):429-35.

(83) Lee JH, Cheng R, Schupf N, Manly J, Lantigua R, Stern Y, et al. The association between genetic variants in SORL1 and Alzheimer disease in an urban, multiethnic, community-based cohort. Arch Neurol 2007 Apr;64(4):501-6.

(84) Reitz C, Cheng R, Rogaeva E, Lee JH, Tokuhiro S, Zou F, et al. Meta-analysis of the association between variants in SORL1 and Alzheimer disease. Arch Neurol 2011 Jan;68(1):99-106.

(85) Reitz C, Tokuhiro S, Clark LN, Conrad C, Vonsattel JP, Hazrati LN, et al. SORCS1 alters amyloid precursor protein processing and variants may increase Alzheimer's disease risk. Ann Neurol 2011 Jan;69(1):47-64.

(86) Rogaeva E, Meng Y, Lee JH, Gu Y, Kawarai T, Zou F, et al. The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. Nat Genet 2007 Feb;39(2):168-77.

(87) Lander ES, Botstein D. Homozygostiy mapping: a way to map human recessive traits with the DNA of inbred children. Science 1987;236:1567-70.

(88) Sheffield V, Carmi R, Kwitek-Black A, Rokhlina T, Nishimura D, Duyk GM, et al. Identification of a Bardet-Biedl syndrome locus on chromosome 3 and evaluation of an efficient approach to homozygosity mapping. Human Molecular Genetics 1994;3(8):1331-5.

(89) Vance JM, Jonasson F, Lennon F, Sarrica J, Damji KF, Stauffer J, et al. Linkage of a gene for macular corneal dystrophy to chromosome 16. Am J Hum Genet 1996;58(4):757-62.

(90) Hastbacka J, de la Chappelle A, Kaitila I, Sistonen P, Weaver A, Lander E. Linkage disequilibrium mapping in isolated founder populations:diastrophic dysplasia in Finland. Nature Genetics 1992;2(november):204-11.

(91) Liu F, rias-Vasquez A, Sleegers K, Aulchenko YS, Kayser M, Sanchez-Juan P, et al. A genomewide screen for late-onset Alzheimer disease in a genetically isolated Dutch population. Am J Hum Genet 2007 Jul;81(1):17-31.

(92) Lee JH, Flaquer A, Stern Y, Tycko B, Mayeux R. Genetic influences on memory performance in familial Alzheimer disease. Neurology 2004 Feb 10;62(3):414-21.

(93) Amish Heritage Committee. Amish and Mennonites in Eastern Elkhart & LaGrange Counties, Indiana 1841-1991. 2nd printing ed. Goshen, Indiana: Amish Heritage Committee; 2009.

(94) Beachy L. Unser Leit: The Story of the Amish. Millersburg, OH: Goodly Heritage Books; 2011.

(95) Hostetler J. Amish Society, 4th ed. Baltimore, MD: Johns hopkins University Press; 1993.

(96) Agarwala R, Biesecker LG, Tomlin JF, Schaffer AA. Towards a complete North American Anabaptist genealogy: A systematic approach to merging partially overlapping genealogy resources. Am J Med Genet 1999 Sep 10;86(2):156-61.

(97) Agarwala R, Biesecker LG, Schaffer AA. Anabaptist genealogy database. Am J Med Genet C Semin Med Genet 2003 Aug 15;121(1):32-7.

(98) Pollin TI, Damcott CM, Shen H, Ott SH, Shelton J, Horenstein RB, et al. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. Science 2008 Dec 12;322(5908):1702-5.

(99) Johnson CC, Rybicki BA, Brown G, D'Hondt E, Herpolsheimer B, Roth D, et al. Cognitive impairment in the Amish: a four county survey. International Journal of Epidemiology 1997;26:387-94.

(100) Johnson CC, Rybicki BA, Brown G, Jackson CE. Prevalence of Dementia in the Amish: a three county servey. American Journal of Epidemiology 1993;138:645.

(101) Rocca WA, Hofman A, Brayne C, Breteler MM, Clarke M, Copeland JR, et al. Frequency and distribution of Alzheimer's disease in Europe: a collaborative study of 1980-1990 prevalence findings. The EURODEM- Prevalence Research Group. Ann Neurol 1991 Sep;30(3):381-90.

(102) Haines JL, Crunk A.E., Gaskell P.C., Scott W.K., van der Walt J, Johnson S.R., et al. Studies of Dementia In the Midwestern Amish. 7th International AD/PD Conference 2005 Mar;25.

(103) Holder J, Warren AC. Prevalence of Alzheimer's disease and apolipoprotein E allele frequencies in the Old Order Amish. J Neuropsychiatry Clin Neurosci 1998;10(1):100-2.

(104) Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001 Sep;17(9):502-10.

(105) Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000 Jun;155(2):945-59.

(106) Abecasis GR, Cherny SS, Cookson WO, Cardon LR. GRR: graphical representation of relationship errors. Bioinformatics 2001 Aug;17(8):742-3.

(107) McCauley JL, Hahs DW, Jiang L, Scott WK, Welsh-Bohmer KA, Jackson CE, et al. Combinatorial Mismatch Scan (CMS) for loci associated with dementia in the Amish. BMC Med Genet 2006;7:19.

(108) Edwards DR, Gilbert JR, Jiang L, Gallins PJ, Caywood L, Creason M, et al. Successful aging shows linkage to chromosomes 6, 7, and 14 in the amish. Ann Hum Genet 2011 Jul;75(4):516-28.

(109) Teng EL, Chui HC. The Modified Mini-Mental State (3MS) examination. J Clin Psychiatry 1987 Aug;48(8):314-8.

(110) Tschanz JT, Welsh-Bohmer KA, Plassman BL, Norton MC, Wyse BW, Breitner JC. An adaptation of the modified mini-mental state examination: analysis of demographic influences and normative data: the cache county study. Neuropsychiatry Neuropsychol Behav Neurol 2002 Mar;15(1):28-38.

(111) Morris JC, Heyman A, Mohs RC, Hughes JP, van BG, Fillenbaum G, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. Neurology 1989 Sep;39(9):1159-65.

(112) Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 2008 Oct;40(10):1253-60.

(113) Thornton T, McPeek MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet 2007 Aug;81(2):321-37.

(114) Pericak-Vance MA, Johnson CC, Rimmler JB, Saunders AM, Robinson LC, D'Hondt EG, et al. Alzheimer's disease and apolipoprotein E-4 allele in an Amish population. Ann Neurol 1996 Jun;39(6):700-4.

(115) Liu F, Kirichenko A, Axenovich TI, van Duijn CM, Aulchenko YS. An approach for cutting large and complex pedigrees for linkage analysis. Eur J Hum Genet 2008 Jul;16(7):854-60.

(116) Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002 Jan;30(1):97-101.

(117) Boyles AL, Scott WK, Martin ER, Schmidt S, Li YJ, shley-Koch A, et al. Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. Hum Hered 2005;59(4):220-7.

(118) Liu F, Elefante S, van Duijn CM, Aulchenko YS. Ignoring distant genealogic loops leads to false-positives in homozygosity mapping. Ann Hum Genet 2006 Nov;70(Pt 6):965-70.

(119) Liu F, Arias-Vasquez A, Sleegers K, Aulchenko YS, Kayser M, Sanchez-Juan P, et al. A genomewide screen for late-onset Alzheimer disease in a genetically isolated Dutch population. Am J Hum Genet 2007 Jul;81(1):17-31.

(120) Welsh KA, Butters N, Hughes JP, Mohs RC, Heyman A. Detection and staging of dementia in Alzheimer's disease: Use of the neuropsychological measures developed for the Consortium to Establish a Registry for Alzheimer's disease. Archives of Neurology 1992;49:448-52.

(121) Hamshere ML, Holmans PA, Avramopoulos D, Bassett SS, Blacker D, Bertram L, et al. Genome-wide linkage analysis of 723 affected relative pairs with late-onset Alzheimer's disease. Hum Mol Genet 2007 Nov 15;16(22):2703-12.

(122) Zhang Z, Hartmann H, Do VM, Abramowski D, Sturchler-Pierrat C, Staufenbiel M, et al. Destabilization of beta-catenin by mutations in presenilin-1 potentiates neuronal apoptosis. Nature 1998 Oct 15;395(6703):698-702.

(123) Park C, Falls W, Finger JH, Longo-Guess CM, Ackerman SL. Deletion in Catna2, encoding alpha N-catenin, causes cerebellar and hippocampal lamination defects and impaired startle modulation. Nat Genet 2002 Jul;31(3):279-84.

(124) Welsh-Bohmer KA, Ostbye T, Sanders L, Pieper CF, Hayden KM, Tschanz JT, et al. Neuropsychological performance in advanced age: influences of demographic factors and Apolipoprotein E: findings from the Cache County Memory Study. Clin Neuropsychol 2009 Jan;23(1):77-99.

(125) Caselli RJ, Reiman EM, Osborne D, Hentz JG, Baxter LC, Hernandez JL, et al. Longitudinal changes in cognition and behavior in asymptomatic carriers of the APOE e4 allele. Neurology 2004 Jun 8;62(11):1990-5.

(126) Deary IJ, Whiteman MC, Pattie A, Starr JM, Hayward C, Wright AF, et al. Apolipoprotein e gene variability and cognitive functions at age 79: a follow-up of the Scottish mental survey of 1932. Psychol Aging 2004 Jun;19(2):367-71.

(127) Small BJ, Rosnick CB, Fratiglioni L, Backman L. Apolipoprotein E and cognitive performance: a meta-analysis. Psychol Aging 2004 Dec;19(4):592-600.

(128) Zehnder AE, Blasi S, Berres M, Monsch AU, Stahelin HB, Spiegel R. Impact of APOE status on cognitive maintenance in healthy elderly persons. Int J Geriatr Psychiatry 2009 Feb;24(2):132-41.

(129) Smith A, Bourdeau I, Wang J, Bondy CA. Expression of Catenin family members CTNNA1, CTNNA2, CTNNB1 and JUP in the primate prefrontal cortex and hippocampus. Brain Res Mol Brain Res 2005 Apr 27;135(1-2):225-31.

(130) Ertekin-Taner N, Ronald J, Asahara H, Younkin L, Hella M, Jain S, et al. Fine mapping of the alpha-T catenin gene to a quantitative trait locus on chromosome 10 in late-onset Alzheimer's disease pedigrees. Hum Mol Genet 2003 Dec 1;12(23):3133-43.

(131) Martin ER, Bronson PG, Li YJ, Wall N, Chung RH, Schmechel DE, et al. Interaction between the alpha-T catenin gene (VR22) and APOE in Alzheimer's disease. J Med Genet 2005 Oct;42(10):787-92.

(132) Myers A, Holmans P, Marshall H, Kwon J, Meyer D, Ramic D, et al. Susceptibility locus for Alzheimer's disease on chromosome 10. Science 2000 Dec 22;290(5500):2304-5.

(133) Bertram L, Blacker D, Mullin K, Keeney D, Jones J, Basu S, et al. Evidence for genetic linkage of Alzheimer's disease to chromosome 10q. Science 2000 Dec 22;290(5500):2302-3.

(134) Li YJ, Scott WK, Hedges DJ, Zhang F, Gaskell PC, Nance MA, et al. Age-of-onset in two common neurodegenerative diseases is genetically controlled. Am J Hum Genet 2002 Apr;70(4):985-93.

(135) Blomqvist ME, Andreasen N, Bogdanovic N, Blennow K, Brookes AJ, Prince JA. Genetic variation in CTNNA3 encoding alpha-3 catenin and Alzheimer's disease. Neurosci Lett 2004 Apr 1;358(3):220-2.

(136) Busby V, Goossens S, Nowotny P, Hamilton G, Smemo S, Harold D, et al. Alpha-T-catenin is expressed in human brain and interacts with the Wnt signaling pathway but is not responsible for linkage to chromosome 10 in Alzheimer's disease. Neuromolecular Med 2004;5(2):133-46.

(137) Cellini E, Bagnoli S, Tedde A, Nacmias B, Piacentini S, Sorbi S. Insulin degrading enzyme and alpha-3 catenin polymorphisms in Italian patients with Alzheimer disease. Alzheimer Dis Assoc Disord 2005 Oct;19(4):246-7.

(138) Morgan AR, Hamilton G, Turic D, Jehu L, Harold D, Abraham R, et al. Association analysis of 528 intra-genic SNPs in a region of chromosome 10 linked to late onset Alzheimer's disease. Am J Med Genet B Neuropsychiatr Genet 2008 Sep 5;147B(6):727-31.

(139) Edwards TL, Pericak-Vance M, Gilbert JR, Haines JL, Martin ER, Ritchie MD. An association analysis of Alzheimer disease candidate genes detects an ancestral risk haplotype clade in ACE and putative multilocus association between ACE, A2M, and LRRTM3. Am J Med Genet B Neuropsychiatr Genet 2009 Jul 5;150B(5):721-35.

(140) Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009 Jul 15;25(14):1754-60.

(141) McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010 Sep;20(9):1297-303.