Examination of Candidate Exonic Variants that Confer Susceptibility to Alzheimer Disease in the Amish

By

Laura Nicole D'Aoust

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

May, 2015

Nashville, TN

Approved:

Dana C. Crawford, Ph.D.

Bingshan Li, Ph.D.

Paul Newhouse, M.D.

Jonathan L. Haines, Ph.D.

Tricia Thornton-Wells, Ph.D.

ACKNOWLEDGEMENTS

I am grateful to all of those with whom I have had the pleasure to work with during this project and others throughout my time in graduate school. Each of the members of my dissertation committee has provided extremely valuable guidance and advice on this project and on advancing my training. I would especially like to thank Dr. Dana C. Crawford, the chairwoman of my committee. I would also like to especially thank my co-mentors Drs. Jonathan L. Haines and Tricia Thornton-Wells. As my main mentor, Dr. Haines has taught me the skills necessary for my future career as a researcher, mentor and leader. I thank the valuable members, past and present, of the Haines Laboratory who have provided me with daily assistance and support. Without the work conducted by previous researchers and the many collaborators, this project would not have been possible. For their time and effort to bring the study to this point, I thank them. I also thank the Amish communities of Ohio and Indiana for the altruism and willingness to participate in these studies. This work would not have been possible without the financial support of the Vanderbilt Medical Scientist Training Program. I am especially indebted to the members of the MSTP leadership team, who have been exceptionally supportive of my development as a future physician scientist.

Lastly, I would like to thank my family for their support and understanding as I have pursed this project and training. My parents are the greatest role models and imparted in me their desire for knowledge.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I
INTRODUCTION

Diagnosis, Pathophysiology and Treatment

Alzheimer disease (AD) is the most common cause of dementia, the global loss of cognitive ability beyond the normal changes associated with aging. AD is distinguished by multiple cognitive deficits manifested by memory impairment and mental disturbances, such as language, motor, sensory and executive functioning impairment. These impairments are characterized by gradual onset and continuing decline. The early behavior changes commonly include slight memory loss, decreases in initiative, and faulty judgment. As the impairment increases, memory loss becomes more significant, higher order functions are affected and behavioral and mood disturbances may occur. In the late stage of AD, memory loss is severe, there is little impulse control, patients may become paranoid and irrational, and severe language deficits may be present.

AD was traditionally diagnosed based upon the criteria and guidelines published by the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Association (ADRDA) (G. McKhann et al., 1984). Additionally, the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* lists the diagnostic criteria for major or mild neurocognitive disorder due to Alzheimer's disease (American Psychiatric Association. & American Psychiatric Association. DSM-5 Task Force., 2013). These criteria include an insidious onset and gradual progression of impairment in at least one cognitive domain of executive function, learning and memory, language, perceptual-motor or social cognition. These deficits must be evident through the concern of the individual, an informant, or the clinician, or

through standardized neuropsychological testing or clinical assessment. These deficits interfere with independence in everyday activities, do not occur due to delirium and are not better explained by another etiology.

According to the DSM-V criteria, probable AD is diagnosed when either there is evidence of a causative AD genetic mutation or when there is clear evidence of significant decline and impairment in memory and learning and at least one other cognitive domain, a steadily progressive and gradual decline in cognition, and no evidence of mixed etiology. Possible AD is diagnosed when there is clear evidence of steadily progressive and gradual decline in memory and learning with no evidence of another etiology.

The National Institute of Aging (NIA) and the Alzheimer's Association issued new criteria and guidelines for clinical use when diagnosing AD that revised the criteria from 1984 (G. M. McKhann et al., 2011). These guidelines focus on the different stages of AD; dementia, mild cognitive impairment (MCI) and preclinical or presymptomatic. These criteria proposed classifying individuals with AD as (1) probable, (2) possible, and (3) probable or possible AD with evidence of AD pathophysiological process. The first two classifications are for clinical use and the third is meant for research settings.

A diagnosis of dementia is made when there are cognitive or neuropsychiatric symptoms that meet five criteria. First, they interfere with the ability to function at work or at usual activities. Second, represent a decline from previous levels of functioning and

performing. Third, are not explained by delirium or major psychiatric disorder. Fourth, cognitive impairment is detected and diagnosed through a combination of (1) history-taking from the patient and a knowledgeable informant and (2) an objective cognitive assessment, either a bedside mental status examination or neuropsychiatric testing. Fifth, the cognitive or behavioral impairment involves a minimum of two of the following domains: impaired ability to acquire and remember new information, impaired reasoning and handling of complex tasks or judgment, impaired visuospatial abilities, impaired language functions, or changes in personality, behavior or comportment.

Probable AD is diagnosed if this patient meets the above five criteria for dementia and has an insidious onset, clear-cut history of worsening of cognition, and initial and most prominent deficits of an amnestic presentation (impaired learning and recall of learned information) or nonamnestic presentation (language, visuospatial or executive dysfunction). Probable AD is not diagnosed when there is evidence of concomitant cerebrovascular disease, core features of dementia with Lewy bodies, features of behavioral variant frontotemporal dementia, features of semantic or nonfluent/agrammatic variant primary progressive aphasia, or evidence for another active neurological disease or medication use that affects cognition. There is increased certainty if there is documented decline or evidence of a causative genetic mutation.

Possible AD is diagnosed when (a) the course meets the cognitive deficits criteria but has either a sudden onset or there is insufficient documentation of decline or (b) has an etiologically mixed presentation of concomitant cerebrovascular disease, features of dementia with Lewy bodies, or evidence of a neurological disease, comorbidity or

medication use that affects cognition. Additionally, these criteria incorporate biomarkers for the pathophysiological process of AD. These biomarkers are of two classes: brain amyloid-beta (Aβ) protein desposition, low CSF $A\beta_{42}$ and positive PET amyloid imaging, and downstream neuronal degeneration, elevated CSF tau, fluorodeoxyglucose (FDG) uptake on PET, and disproportionate atrophy on MRI. If there is positive evidence of a biomarker, it will increase the certainty that the dementia is of AD pathophysiological process but it is not recommended to be used as routine diagnosis criteria.

AD is categorized as early onset at age 65 or below or late onset (LOAD) after the age of 65. Because the definitive diagnosis is only made post-mortem, when beta-amyloid plaques and neurofibrillary tangles are found upon autopsy, attempts must be made to rule out other etiologies of dementia. Other types of dementia include vascular dementia, dementia with Lewy bodies, frontotemporal lobar degeneration, mixed dementia, Parkinson's disease, Creutzfeldt-Jakob disease and normal pressure hydrocephalus.

The hallmark pathophysiologic change observed in AD is gross global cortical atrophy, especially in neocortical association, non-primary motor and non-primary sensory areas (Figure I-1). There is also a non-uniform loss of neurons. This loss is greatest in the neocortex, hippocampus, amygdala, nucleus basalis of Meynert, nucleus locus coeruleus, and the Raphe nuclei. These locations are essential for higher-order cognitive functioning, learning and memory, emotional behavior, sleep/wake cycles and mood regulation. Upon histological examination, amyloid plaques and neurofibrillary tangles can be seen (Figure I-1). Misfolded beta-amyloid (Aβ) proteins aggregate and are deposited in the extracellular space as plaques. Paired helical filaments and

hyperphosphorylated tau protein comprise the intracellular neurofibrillary tangles that are often found in the brains of AD patients.

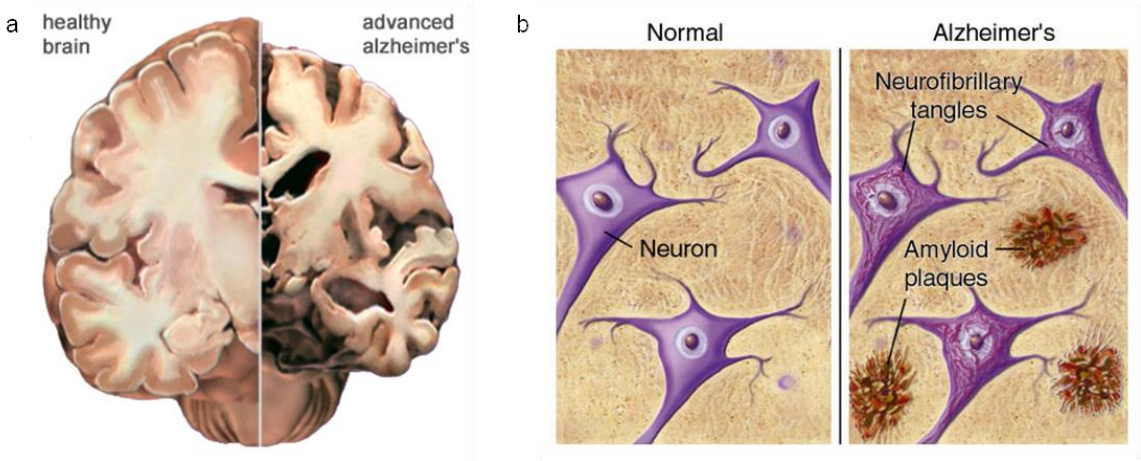

Figure I-1. Hallmark pathophysiology and histological changes associated with AD. (a) Gross cortical atrophy can be observed in the AD brain on the right compared to the healthy brain on the left. (b) Intracellular neurofibrillary tangles and misfolded beta-amyloid plaques can be seen in the schematic of histological examination of a brain from an AD patient. Images adapted from www.alz.org and www.ahaf.org.

According to the amyloid cascade hypothesis of AD pathogenesis, these plaques may stimulate surrounding cells, resulting in chronic inflammation. Aβ is the product of the sequential cleavage of amyloid prescursor protein (APP) by beta and gamma secretases. $A\beta_{42}$ is more likely to aggregate in plaques and is considered the more pathogenic form of the peptides produced by cleavage (Jakob-Roetne & Jacobsen, 2009). How these plaques cause cell injury and induce inflammation is not completely understood, but it is hypothesized that soluble amyloid may behave similar to ion channels and alter regulation of calcium flow into the neurons or alter the integrity of the cell membrane (Demuro et al., 2005; Sokolov et al., 2006). Changes in blood vessels and the blood-brain barrier can also be seen as amyloid is deposited in arteries leading

to leakage and hemorrhage. Selective breakdown of the blood-brain barrier may compromise the effectiveness of amyloid removal. The accumulation of plaques may interfere with neuron communication and contribute to cell death. The accumulation of plaques induces neuronal toxicity and synapse death, which then results in tangle formation (King et al., 2006). The three genetic risk factors for early-onset AD, the precursor protein and two proteins involved in the enzymatic cleavage of the protein, are all involved in the Aβ cascade and many of the LOAD risk loci (*CLU*, *PICALM*, *SORL*) may interact in this pathway. This evidence supports the above hypothesis that Aβ homeostasis is a factor in AD pathogenesis.

The intracellular tangles block transport within the neuron and may also contribute to cell death. Studies investigating the correlations between tangle load and clinical symptoms have found that these tangles may be a better predictor of disease severity than amyloid burden (Arriagada, Growdon, Hedley-Whyte, & Hyman, 1992; Giannakopoulos et al., 2003; Gold et al., 2001). Tau is a microtubule-associated protein and is well established in maintaining structural integrity of the neuron. Phosphorylated tau is more likely to aggregate and leads to depolymerization of microtubules. This evidence suggests that the tangles cause a loss of normal function and are toxic to the neuron (Ballatore, Lee, & Trojanowski, 2007; Buee, Bussiere, Buee-Scherrer, Delacourte, & Hof, 2000). A larger role for tau in the pathogenesis of AD is supported by the failure of anti-amyloid drug therapies in Phase III trials (Berkrot, 2012; ClinicalTrials.gov; ClinicalTrials.gov; "Eli Lilly and Company Announces Top-Line Results on Solanezumab Phase 3 Clinical Trials in Patients with Alzheimer's Disease," 2012).

As more research investigates the underlying biological process that contributes to development of AD, it is becoming clearer that many processes (plaque accumulation, tangle formation, cerebrovascular dysregulation, etc.) result in disease, as opposed to a single factor (Savelieff, Lee, Liu, & Lim, 2013). A population-based study investigating correlations between pathological findings upon autopsy and clinical diagnosis of dementia or AD found extensive overlap between individuals with and without dementia (Corrada, Berlau, & Kawas, 2012). In the brains of non-demented subjects, 22% had high stages of tangles and over 50% had neuritic plaques. Furthermore, 49% of the non-demented subjects met pathological criteria for AD but only 57% of the demented patients met these criteria. The non-demented subjects who did have pathological evidence had similar average MMSE scores as did the controls without evidence. Recent studies provide strong evidence to suggest the etiology of LOAD may be due to cerebrovascular dysregulation and that the neuronal degeneration is secondary to this dysregulation, but more work is needed to support this theory (Bomboi et al., 2010).

Since understanding the etiology of a disease is essential to the development of effective diagnostics and therapeutics, addressing which of these hypotheses correctly explains the pathological pathway of AD is of utmost importance. As the true pathogenesis of disease remains unclear, treatment options are only available to manage memory and behavioral problems (Table I-1). Available treatment for AD is given to help alleviate symptoms, cholinesterase inhibitors and drugs targeted to glutamate receptors help some patients but results vary from patient to patient. Four cholinesterase inhibitors and one N-methyl-D-aspartate (NMDA) receptor antagonist have been approved by the United States Food and Drug Administration (FDA) for AD treatment. Clinical trials investigating immunotherapy agents that target beta amyloid

were halted after those receiving treatment did not have improved cognition or daily functioning compared to those who received placebo (Berkrot, 2012; ClinicalTrials.gov; ClinicalTrials.gov; "Eli Lilly and Company Announces Top-Line Results on Solanezumab Phase 3 Clinical Trials in Patients with Alzheimer's Disease," 2012). However, when analyzing only individuals with mild impairment, solanezumab did show a small benefit on cognition and trials continue to investigate this effect.

Table I-1. FDA approved AD therapies and drugs in clinical trials.

	Class	Drug	Stage of Disease
Approved	NMDA receptor antagonist	memantine	moderate to severe AD
	cholinesterase inhibitor	donepezil	mild to moderate AD
		tacrine	
		rivastigmine	
		galantamine	
Clinical Trials	immunotherapy	solanezumab	mild AD?
		~~bapineuzumab~~	~~moderate AD~~

Epidemiology and Risk Factors

The prevalence of AD in the United States for individuals aged 85 years and older is estimated to be around 30% based on census data and a population-based study of chronic diseases of older people. Furthermore, the number of people with AD is predicted to triple by 2050, the year in which the youngest baby boomers will be over 85 years old (Hebert, Weuve, Scherr, & Evans, 2013; Thies, Bleiler, & Alzheimer's, 2013). This estimate reflects the changing structure of the United States (US) as the population ages and as people have longer life expectancies. Epidemiological studies have found that more women than men have dementia, in the US it is estimated that 2/3 of all AD patients are women (Hebert et al., 2013). The estimated lifetime risk of AD for a female

at age 65 (without dementia) is 17%, for men at this age the risk is 9% (Seshadri et al., 2006). The World Health Organization lists AD as the 4th leading cause of death in high-income countries and it is the fifth leading cause of death in Americans over the age of 65 (Thies et al., 2013; WHO, 2011). These numbers may not be true estimates because of the way causes of death are recorded and tabulated. If AD is the underlying cause of death, a person is considered to have died from the disease and it is listed as the cause of death on the death certificate. However, AD patients can experience a wide variety of comorbidities that are affected by the individual's decreased cognition, executive functioning and reliance on caregivers. Pneumonia is a common cause of death among dementia patients and may be listed on the death certificate as the immediate cause (Brunnstrom & Englund, 2009). The total costs for caring for individuals with AD and other dementias are projected to be $1.2 trillion in 2050 (Hebert et al., 2013).

There are many risk factors associated with developing AD. It is an age-dependent disease as very few individuals below the age of 65 have AD, about 5-10% of all cases. But as age increases, so does the prevalence of the disease, from 3.0% in individuals aged 65-74 years, 17.6% 75-84 years and 32.3% aged 85 years or older (Hebert et al., 2013). Traumatic brain injury (TBI) has been associated with a higher risk of AD, two-times the risk for moderate TBI and 4.5x risk for severe TBI (Fleminger, Oliver, Lovestone, Rabe-Hesketh, & Giora, 2003; Guo et al., 2000; Mortimer et al., 1991; Plassman et al., 2000; Schofield et al., 1997). Evidence suggests cardiovascular risk factors, such as smoking, obesity, diabetes, high cholesterol and hypertension during middle age, and the metabolic syndrome, are also risk factors for AD as vascular integrity is important for beta-amyloid homeostasis (Table I-2) (Brenner et al., 1993; Doll, Peto, Boreham, & Sutherland, 2000; Hebert et al., 1992; Holden et al., 2009; Kivipelto et

al., 2001; Launer et al., 1999; Launer, Masaki, Petrovitch, Foley, & Havlik, 1995; Leibson et al., 1997; Lesser et al., 2001; Luchsinger, Tang, Stern, Shea, & Mayeux, 2001; Merchant et al., 1999; Michikawa, 2003; Ott et al., 1998; Ott et al., 1999; Raffaitin et al., 2009; Solfrizzi et al., 2010; Verghese, Lipton, Hall, Kuslansky, & Katz, 2003; Whitmer, Sidney, Selby, Johnston, & Yaffe, 2005; Wieringa, Burlinson, Rafferty, Gowland, & Burns, 1997). However, continued study of these factors is necessary to determine how each contributes to developing AD and to resolve discrepancies between study findings.

Table I-2. Published effect sizes for cardiovascular risk factors that have been associated with dementia or AD. Effect sizes are either odds ratios, relative risks, or hazard ratios.

Risk Factor	Effect	Study
Smoking	0.61	Brenner et al., 1993
	0.99	Doll et al., 2000
	0.7	Hebert et al., 1992
	1.97-3.17 male/1.08-1.50 female	Launer et al., 1999
	1.9	Merchant et al., 1999
	2.3	Ott et al., 1998
	1.26	Whitmer et al., 2005
Obesity	0.66	Holden et al., 2009
	1.74	Whitmer et al., 2005
Diabetes	2.27 male/1.37 female	Leibson et al., 1997
	1.3	Luchsinger et al., 2001
	1.6	Cheng et al., 2011
	1.9	Ott et al., 1999
	1.46	Whitmer et al., 2005
High cholesterol	2.1	Kivipelto et al., 2001
	1.26	Lesser et al., 2001
	1.42	Whitmer et al., 2005
Hypertension	2.3	Kivipelto et al., 2001
	1.24	Whitmer et al., 2005

Certain other diseases or syndromes, such as Down's syndrome and cerebrovascular injury, are risk factors (Korenberg et al., 1994; Pendlebury & Rothwell, 2009). Down's

syndrome is caused by trisomy 21, and individuals with this syndrome typically develop AD due an extra copy of *APP* on chromosome 21 which increases production of mRNA and protein (Oyama et al., 1994). It is estimated that 70% of individuals with Down's syndrome will develop dementia by age 70 (Evenhuis, 1990). Upon postmortem analysis, nearly all Down's syndrome patients will have plaques and meet pathological criteria for AD (Mann, 1988). Again, this evidence supports the conclusion that amyloid can contribute to AD but the fact that not all Down's syndrome patients develop AD symptoms, despite evidence of amyloid plaques, suggests that multiple processes may be interacting.

Some evidence suggests high physical activity is a protective factor against AD (Abbott et al., 2004; Fratiglioni, Paillard-Borg, & Winblad, 2004; Podewils et al., 2005; Rovio et al., 2005; Scarmeas, Levy, Tang, Manly, & Stern, 2001; Verghese, Lipton, Katz, et al., 2003). It is hypothesized that education or a high cognitive brain reserve and cognitive training are also protective, but the specific mechanism and which cognitive domains are affected need to be further studied to provide a consensus (Acevedo & Loewenstein, 2007; Ball et al., 2002; Carlson et al., 2008; Fratiglioni & Wang, 2007; Unverzagt et al., 2007).

Known Genetic Risk Factors

In addition to the numerous risk factors detailed previously, family history and inherited genetic variations are also associated with AD. The early-onset form of AD resembles Mendelian disorders, and therefore familial studies investigating large, multigenerational

studies by linkage analysis and positional were successful in identifying *APP* (chromosome 21q), *PSEN1* (14q), and *PSEN2* (1q) (Goate et al., 1991; Levy-Lahad et al., 1995; Rogaev et al., 1995; Sherrington et al., 1995). Over 200 disease-causing mutations in these genes have been identified for EOAD.

The heritability of LOAD is estimated at 60-80% (Gatz et al., 2006). *APOE* is the strongest genetic risk factor for LOAD, but accounts for far less than 30% of the expected genetic effects (Corder et al., 1994; Corder et al., 1993; Goldstein et al., 2001; Graff-Radford et al., 2002; Henderson et al., 1995; Hsiung, Sadovnick, & Feldman, 2004; Lambert et al., 2013; Myers et al., 1996; Naj et al., 2011; Polvikoski et al., 1995; Skoog et al., 1998; Slooter et al., 2004).

Large genome-wide studies have identified risk loci in or very near *CR1*, *CLU*, *PICALM*, *BIN1*, *EPHA1*, *MS4A*, *CD33*, *CD2AP*, *ABCA7*, *HLA-DRB5*/*HLA-DRB1*, *PTK2B*, *SLC24A4*/*RIN3*, *DSG2*, *INPP5D*, *MEF2C*, *NME8*, *ZCWPW1*, *CELF1*, *FERMT2*, and *CASS4* (Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Lambert et al., 2013; Naj et al., 2011; Seshadri et al., 2010) (Table I-3). These loci are involved in complement pathway activation, nervous system development, inflammation, synaptic transmission, and beta-amyloid regulation. However, the common variants in these loci confer very modest risk. Evidence from candidate gene studies investigating *SORL1* has been supported by additional studies and even suggest additional genes associated with *SORL1*, such as the *SORCS* family (Lambert et al., 2013; J. H. Lee et al., 2007; Reitz, Cheng, et al., 2011; Reitz, Tokuhiro, et al., 2011; Reitz et al., 2013; Rogaeva et al., 2007). Recent sequencing studies have identified rare variants in *APP* and *TREM2*

(Guerreiro et al., 2013; Jonsson et al., 2012; Jonsson et al., 2013). Variants in *MAPT* cause a variety of neurodegenerative disorders, including frontotemporal dementia-spectrum disorders and PSP. A rare variant in this gene identified in a patient with PSP was found to also be associated with increased risk of AD (MAF = 0.20%, n = 3345 cases, OR = 2.3, p = 0.004) (Coppola et al., 2012). By genotyping the rare variant identified from sequencing a subset of individuals in a larger dataset, the researchers were able to show an association with LOAD and report larger effect sizes than those from the GWAS studies.

Table I-3. Summary of genes and SNPs/alleles associated with LOAD disease risk.

Gene	Study	SNP/allele	OR
APOE	Corder et al., 1993	E4/E2	3.78
ABCA7	Lambert et al., 2013	rs4147929	1.15
APP	Jonsson et al., 2012	rs63750847	0.19
BIN1	Lambert et al., 2013	rs6733839	1.22
CASS4	Lambert et al., 2013	rs7274581	0.88
CD2AP	Lambert et al., 2013	rs10948363	1.1
CD33	Lambert et al., 2013	rs3865444	0.94
CELF1	Lambert et al., 2013	rs10838725	1.08
CLU	Lambert et al., 2013	rs9331896	0.86
CR1	Lambert et al., 2013	rs6656401	1.18
DSG2	Lambert et al., 2013	rs8093731	0.73
EPHA1	Lambert et al., 2013	rs11771145	0.9
FERMT2	Lambert et al., 2013	rs17125944	1.14
HLA-DRB5/HLA-DRB1	Lambert et al., 2013	rs9271192	1.11
INPP5D	Lambert et al., 2013	rs35349669	1.08
MAPT	Coppola et al., 2012	p.A152T	2.3
MEF2C	Lambert et al., 2013	rs190982	0.93
MS4A	Lambert et al., 2013	rs983392	0.9
NME8	Lambert et al., 2013	rs2718058	0.93
PICALM	Lambert et al., 2013	rs10792832	0.87
PTK2B	Lambert et al., 2013	rs28834970	1.1
SLC24A2/RIN3	Lambert et al., 2013	rs10498633	0.91
SORL1	Lambert et al., 2013	rs11218343	0.77
ZCWPW1	Lambert et al., 2013	rs1476679	0.91
TREM2	Jonsson et al., 2013	rs75932628	2.92

Rare variants in early-onset genes, *APP, PSEN1* and *PSEN2*, were identified by sequencing 439 probands from multiplex (multiple affected siblings) LOAD families (Cruchaga et al., 2012). These results suggest that rare variants found in genes known to cause an early-onset version of a disease or disorder may contribute disease risk to a late-onset version of the same disease. This study also identified known causative mutations in *PSEN1* and *GRN*. Additional novel variants in these genes as well as a novel variant in *MAPT* are likely to be causative or highly penetrant risk alleles. As of April 2011, an online database of genetic association studies performed for AD,

AlzGene, included 1,395 studies reporting associations for 2,973 polymorphisms in 695 genes (Bertram, McQueen, Mullin, Blacker, & Tanzi, 2007).

As with many complex diseases, the identified variants do not explain the total expected genetic risk due to heritability in populations with similar ancestry to those in which they were identified. Prior to the most recent meta-analysis, a study was conducted to evaluate the heritability explained by the then known susceptibility variants for LOAD in European-descent populations (So, Gui, Cherny, & Sham, 2011). This study estimated that four risk loci (*APOE*, *CR1*, *CLU*, and *PICALM*) may explain 18% of the total variance, or might explain 23% of the 79% heritability of LOAD. A recent study investigated the proportion of total variation tagged by all genome-wide SNPs for three common diseases (S. H. Lee et al., 2013). For Alzheimer disease, this study used 499,757 SNPs genotyped in 3290 cases and 3849 controls. The researchers estimated that these SNPs explain 24% of the total variation. The most recent meta-analysis performed on genome-wide SNP data for LOAD demonstrated that the most strongly associated SNPs at each of the 21 risk loci, except for APOE, had population-attributable fractions (PAFs) or preventive fractions between 1.0-8.0% (Lambert et al., 2013). The PAF is the percentage of AD cases that could be prevented if the risk factors were removed. The unexplained genetic risk suggests additional variants in these known genes or previously unassociated genes may confer susceptibility. Through the identification of additional risk variants or loci, more can be learned about the underlying biology and pathogenesis of AD that can inform future studies about diagnosis and treatment targets.

Additional rare variants with larger effects may explain some of the unknown genetic risk for LOAD according to the common disease multiple rare variant hypothesis. This hypothesis states common diseases are caused by many causal rare variants in a single gene that have large effect sizes. Previous association studies have been limited by technology and methods and have been unable to effectively interrogate rare variation. The identification of rare functional variants may have more clinical utility in identifying at-risk individuals who might benefit from early treatment or increased screening. By studying variants identified from whole-exome sequencing, this study aimed to interrogate portions of the human genome previously unstudied under the common disease common variants hypothesis.

The Amish of Ohio & Indiana

Most genetic studies evaluating LOAD risk are performed in complex heterogeneous populations, introducing analysis and interpretation problems due to heterogeneity. To further the understanding of this disease, the genetically isolated Amish communities of Ohio and Indiana have been studied to identify additional genetic variants that contribute to disease risk.

There were two significant waves of immigration that established the Amish communities in the United States. The first wave occurred in the 1700s with Swiss Anabaptists settling in Pennsylvania. Then, in the 1800s, additional Swiss Anabaptists and individuals from the Pennsylvania settlements immigrated to Ohio and Indiana (Beachy, 2011) (Figure I-2).The severe population bottleneck that occurs when a small group of individuals establishes a separate subpopulation is known as a founder effect. The

random drift that occurs in this new subset of the total variation sampled from the population may change disease prevalence, reduce effective population size, alter allele frequencies and change patterns of linkage disequilibrium. These populations are more genetically homogenous because of this founder effect and because members of these communities marry within their culture, thus limiting the amount of new genetic variation introduced from the general population. Additionally, due to their strict lifestyle, environmental exposures are more homogenous. For example, the older Amish have generally led an agricultural lifestyle, achieved similar levels of education, and consumed similar diets. These factors make the Amish populations advantageous for genetic studies by controlling for both genetic and environmental heterogeneity.

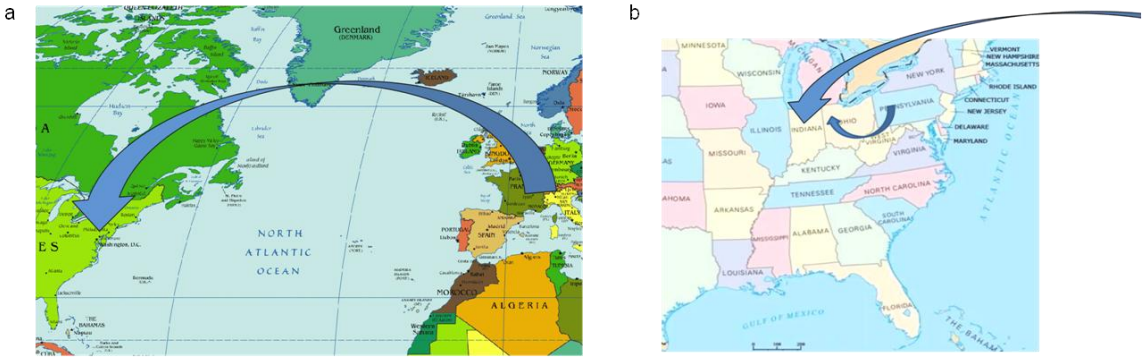

Figure I-2. Amish immigration patterns. (a) First wave of immigration during the 1700's which founded the community in Pennsylvania. (b) Second wave of immigration during the 1800's in which additional founders came from Europe and were joined in Ohio & Indiana by settlers from the older Pennsylvania community.

The Anabaptist Genealogy Database (AGDB) is a database that can be easily queried to assist in genetic studies. The developers of this database combined and digitized genealogy records and books that include multiple individuals and families. This database can be queried to generate a pedigree that connects all individuals of interest,

such as all affected individuals in a study population, to a common ancestor or couple (Agarwala, Biesecker, & Schaffer, 2003). This resource generated a pedigree consisting of over 5000 Amish members that spans 13 generation that connects all of the individuals sampled from the communities in Ohio and Indiana. From this pedigree, the relationship status and degree of relatedness can be determined for all individuals in the full dataset.

Previous Work in the Amish

Numerous studies investigating complex neurodegenerative diseases and aging have been conducted in this isolated founder population for over 10 years. These studies have investigated the genetic structure and variation underlying LOAD disease risk in the Amish communities from Adams, Elkhart and LaGrange Counties in Indiana and Holmes County in Ohio (Figure I-3). The first study investigated the contribution of the *APOE* ε4 risk allele in this population by studying six AD affected individuals and their unaffected siblings. However, none of the individuals genotyped for *APOE* carried the ε4 allele. When Amish controls were chosen at random from the population (n = 106 chromosomes, mean age = 55 years, age range 20-87 years), it was found that the ε4 allele frequency was 0.037. These Amish controls had a significantly lower frequency of the risk allele when compared to three different sets of Caucasian controls populations of grandparental Centre d'Etude du Polymorhpism Humaine (CEPH, n = 182 chromosomes, ε4 = 0.16, $p < 2 \times 10^{-4}$), AD spouse controls from Alzheimer's Disease Research Center (ADRC, n = 444 chromosomes, ε4 = 0.15, $p < 6 \times 10^{-5}$), and data from Menzel and coworkers (n = 2,000 chromosomes, ε4 = 0.14, $p < 2 \times 10^{-6}$) (PericakVance et al., 1996). More recent evidence from the genotypes from around 900 Amish individuals supports this early observation (Cummings et al., 2012). This study found a

significant association with *APOE* ($p = 9.0 \times 10^{-6}$) in the Amish population except for Adams County. In LOAD cases from the three other counties the ε4 allele frequency was 0.18 while it was only 0.11 in cognitively normal controls from the same counties. This work supported the hypothesis that a genetic etiology independent of *APOE* was likely to be underlying the AD observed in the Amish.

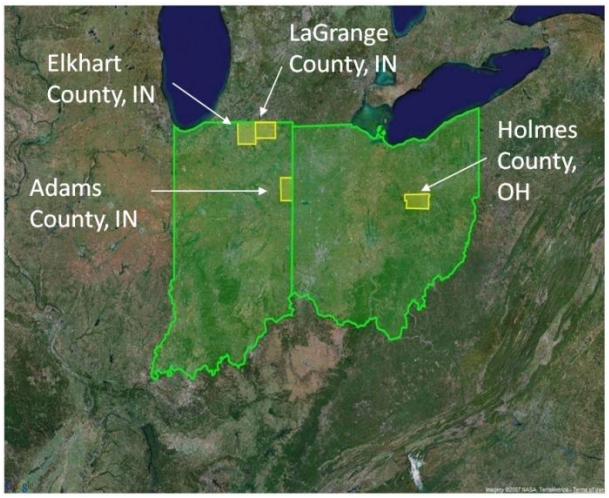

Figure I-3. Map showing the four Amish communities studied for over 10 years for LOAD, dementia, Parkinson's disease, age-related macular degeneration and autism.

To identify genetic variation independent of *APOE*, a family-based study of a large multiplex pedigree screened the autosomal genome for evidence of linkage (Ashley-Koch et al., 2005). Novel loci were identified that had not previously been implicated by genomic screens of outbred populations. Based on evidence from the literature, it was hypothesized that mitochondrial dysfunction was contributing to dementia in the Amish. However, haplotype analysis and maternal lineage tracing did not identify haplotypes more common in cases compared to controls (van der Walt et al., 2005). Using microsatellites, a genome-wide linkage analysis was performed to study five Amish

families (Hahs et al., 2006). This study replicated dementia loci identified from other populations, but additionally identified two novel loci. This work was expanded using Combinatorial Mismatch Scanning to perform association testing (McCauley et al., 2006). Expanding on this prior work, a genome-wide linkage and association study was performed using 798 individuals (109 LOAD cases) and over 600,000 single nucleotide polymorphisms (SNPs) (Cummings et al., 2012). This analysis identified four novel linkage regions. Under the most significant multipoint linkage peak on chromosome 2p12 with a maximum LOD of 6.14, one SNP was associated with LOAD with a p-value of 1.29×10^{-4}. The three additional loci with a heterogeneity LOD (HLOD) > 3 were detected on 3q26, 9q21 and 18p11. Collectively, these studies provided evidence to reject the original hypothesis that a major locus for LOAD existed in the Amish genome.

As a complement to the dementia studies, successful aging (SA) has been investigated to identify genetic variation that may protect against neurodegenerative diseases. This phenotype is characterized by preservation of cognitive ability, physical function and social engagement. A genome-wide linkage screen of 5,944 SNPs was conducted in 214 Amish individuals (48 SA and 166 non-SA) (Edwards et al., 2011). Three loci reached significant heterogeneity log odds (HLOD) scores for multipoint linkage, chromosome 6 from 52-65Mb had a max HLOD of 4.49, chromosome 7 from 49-75Mb had a max HLOD of 3.11 and chromosome 14 from 42-53Mb had a max HLOD of 3.28. In a follow-up study of 263 individuals (74 SA and 189 controls), over 600,000 SNPs were examined for both association and linkage (Edwards et al., 2013). The results from this study suggest a novel linked and associated region on 6q25-27 with a maximum HLOD of 3.2 and minimum association p-value of 2.36×10^{-5}.

Genetic variation in the mitochondrial genome has been associated with longevity in multiple populations. Therefore, mitochondrial haplogroups were investigated in the Amish (Courtenay et al., 2012). Amish SA cases were more likely to carry Haplogroup X and less likely to carry Haplogroup J compared to controls. The association with Haplogroup X was novel as no significant associations had previously been reported for age-related diseases. Additionally, this haplogroup accounts for 7% of the Amish individuals in the study, but occurs in less than 5% of all European populations. The association with Haplogroup J replicated previous reports of a population-specific positive association with longevity.

Since these isolated populations differ from the general population, the specific variants may not be present in the same frequency or have the same effect. However, it is expected that the same genes and pathways implicated by the variants will also be associated and confer risk in the general population. By studying the genetics of these isolated populations, the limiting heterogeneity often occurring in complex population studies can be overcome, and variants or genes that help explain the missing heritability of LOAD can be identified.

Gaps in Knowledge Addressed

The previous studies detailed above support the conclusion that genetic heterogeneity does exist in this isolated population, and that genetic variation that contributes to disease risk continues to be undiscovered. To understand how the known risk loci identified in the general population contribute to disease risk in this population, total

genetic burden was calculated for Amish cases and controls and compared to unrelated individuals. If the known genetic risk burden is lower in the Amish, this would be more evidence that additional loci remain to be discovered in this population.

This current study also built on the previous dementia work conducted in the Amish by using whole-exome sequencing of a selected subset of the overall study population as a screening tool to identify variants harbored in the regions of the genome that are most likely to contribute risk. By then genotyping the most significant and interesting candidate variants from this screen in the full dataset, there was more power to detect an association between the variant and phenotype of interest (Figure I-4). If associations were detected in the full dataset, the variants or genes could be studied in the general population to learn more the underlying disease process in a more heterogeneous population. However, if this exonic variation did not confer susceptibility to disease risk, additional studies would need to be performed in this isolated population to continue searching for risk loci.

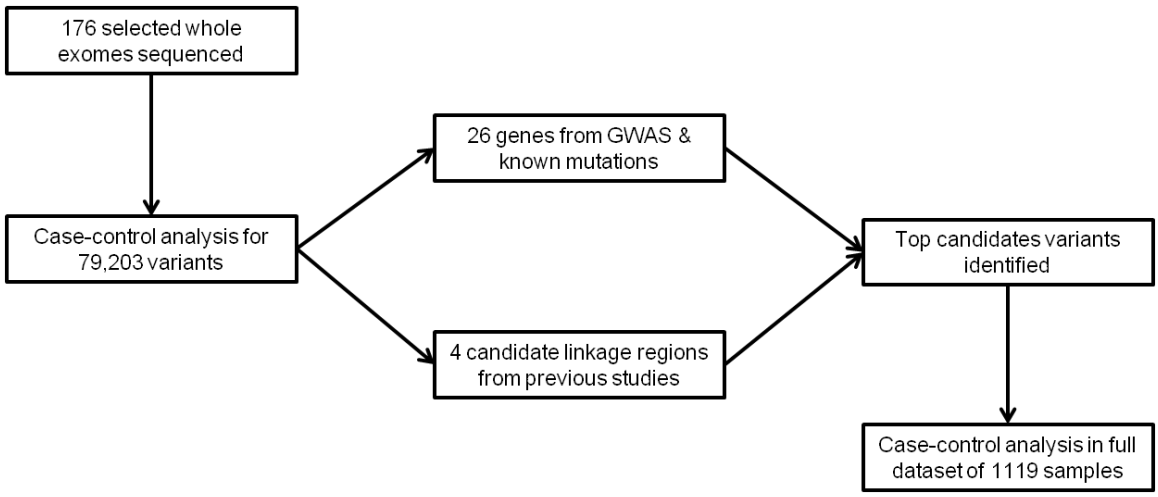

Figure I-4. Flow Diagram of this Study. Individuals were selected from the full Amish dataset for whole-exome sequencing. The variants identified from these data were used to screen two classes of variants, genes that are very near or contain GWAS hits and four candidate linkage regions implicated by previous studies. The top variants from these two classes were then genotyped in the full dataset and case-control association was performed.

CHAPTER II
ANALYSIS OF GENETIC RISK SCORE

Introduction

The first successful GWASes were published in 2005 and as of the end of 2013, 1,779 publications have reported associations for over 12,000 SNPs (Hindorff). Of these, about 2,000 associations are considered robust (Manolio, 2013). Recent studies suggest that for an associated GWAS marker to be suitable for translation to clinical care, and specifically for risk prediction use, the disease should be highly heritable, the marker should explain a large proportion of the expected heritability, the marker should be available for targeted genotyping in a high-risk group, the marker should increase predictive value, and that the disease management should have available preventative strategies (Manolio, 2013). When considering late-onset Alzheimer disease (LOAD) and these criteria, it is a highly heritable disease but the known risk loci only explain about 50% of the expected heritability estimates. High-risk individuals, those with a family history of disease and those with traumatic brain injury or relevant cardiovascular risk factors, could easily be genotyped for a targeted panel consisting of GWAS hits and *APOE* (see Chapter I). The predictive nature of the GWAS hits has not yet been sufficiently studied but most of the risk loci confer modest risk. Treatment options for LOAD are currently largely targeted at symptoms but current studies are investigating treatment in asymptomatic individuals and subjects with mild impairment. Some of these characteristics suggest that LOAD would be a suitable disease for which clinical risk prediction would be suitable, but several key pieces are still missing, most notably missing heritability, predictive markers, and preventative disease management.

Genetic risk scores, which sum the individual effects of multiple risk loci, have been studied for many complex diseases. For example, the odds ratio for type 2 diabetes associated with each additional genetic risk point, or each additional risk allele, was 1.19 and 1.16 for men and women, respectively, when adjusted for age and body mass index (BMI) (Cornelis et al., 2009). When the extremes of the distributions were investigated, individuals with BMI greater than 30, a main risk factor for the disease, and a risk score in the highest quartile had an odds ratio of 14.06 when compared to individuals with BMI less than 25 (not overweight or obese) and a risk score in the lowest quartile.

Age-related macular degeneration (AMD) is a well-studied phenotype and the known genetic risk loci explain 15-65% of the expected heritability, dependent on disease prevalence (Fritsche et al., 2013). With a large portion of the expected genetic risk identified, the use of genetic risk scores in this phenotype may provide useful categorization of individuals into low and high risk groups. Thirteen reported AMD risk loci were combined into a genetic risk score for 986 cases and 796 controls (Grassmann, Fritsche, Keilhauer, Heid, & Weber, 2012). These cases had a significantly higher mean score compared to controls ($p < 0.01$). Furthermore, the relative risk of AMD per risk unit was 2.72. The area-under-the-curve (AUC) for the receiver-operating characteristic curve (ROC) for the risk score was 0.82, similar to previous studies of risk score with fewer AMD loci (Gibson, Cree, Collins, Lotery, & Ennis, 2010; Seddon et al., 2009). Using 19 known loci to calculate risk score, a subsequent study distinguished between cases and controls with an AUC = 0.74 (Fritsche et al., 2013).

Another study investigated the predictability of genetic risk score on the progression from mild cognitive impairment (MCI) to AD (Rodriguez-Rodriguez et al., 2013). Eight LOAD risk SNPs were genotyped in 118 converters and 170 non-converters. Overall, the genetic risk score was not associated with the risk of converting from MCI to AD. The upper two tertiles did progress two-fold more rapidly than individuals in the lower tertile ($p = 0.047$ for second tertile and $p = 0.031$ for top tertile). While this study did find differences between risk scores for the phenotype extremes, there is not sufficient evidence to suggest genetic risk scores can be used to predict an individual's likelihood and risk of developing or progressing to LOAD. Many recent studies have identified numerous LOAD risk loci in the general Caucasian population. The most recent meta-analysis identified 21 replicated or novel markers associated with LOAD, many more than the eight investigated in the progression study (Lambert et al., 2013). Additional studies targeting identification of novel LOAD risk loci and investigating the predictive utility of the currently known loci may provide the necessary evidence and support for the use of genetic risk scores in LOAD if suitable disease treatments become available.

In addition to predicting high or low risk groups, genetic risk scores may identify different underlying genetic architecture between populations. As described in the Introduction, isolated founder populations go through a severe population bottleneck when the subpopulation is established by the small group of individuals. The genetic variation carried by these founders undergoes random drift within this subpopulation that alters allele frequencies and patterns of linkage disequilibrium (LD). Additionally, individuals from these populations tend to marry within their culture, thus limiting the amount of genetic variation introduced from the general population. The markers interrogated by genome-wide association studies (GWAS) tend to be tagging markers, or markers that

are in LD with the true risk variant. If the LOAD risk loci identified by the recent GWASes are tagging the risk variant and are not the risk variant themselves, and if the Amish have unique patterns of LD in these implicated regions because of the founder effect, it is unlikely that these known risk alleles will confer the same risk in this subpopulation.

Therefore, if the known risk alleles do not contribute the same risk in the Amish, it is hypothesized that the Amish cases will have a significantly lower burden of risk alleles when compared to the LOAD cases from a dataset of unrelated individuals. However, if all or a subset of the known genetic risk alleles do contribute to disease risk in the Amish, the Amish cases should tend to have a significantly higher genetic risk score than the Amish cognitively normal controls. To test these two related hypotheses, the total genetic risk score for the most recently identified risk loci was calculated for all individuals and compared across affection statuses and populations to characterize how the known LOAD risk loci contribute to risk burden in the Amish.

Methods

Study populations

The full dataset for which samples have been collected is comprised of individuals from the Amish communities in Adams, Elkhart and LaGrange Counties in Indiana and Holmes County in Ohio. This same study population is also the parent dataset for the additional experiments detailed in future chapters. Individuals were ascertained through public directories, public notices and referrals from previously enrolled participants. Over 30% of the Amish populations over the age of 65 have been contacted, and 87% of these individuals have consented to participate in the study. The Modified-Mini Mental

Status (3MS) exam was used to screen individuals during the initial interviews (Teng & Chui, 1987). Information from these baseline screens and additional cognitive testing were used to generate a consensus diagnosis according to the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Association (ADRDA) criteria (G. McKhann et al., 1984). Methods for ascertainment were reviewed and approved by the individual Institutional Review Boards of the respective institutions. Sample collection, DNA extraction, cognitive testing and affection statuses derived from the consensus diagnoses followed procedures detailed in previous studies conducted in these populations (Cummings et al., 2012). DNA from blood samples were allocated by the DNA banks at the Hussman Institute of Human Genomics at the University of Miami and the Center for Human Genetics Research at Vanderbilt University.

The Modified-Mini Mental Status (3MS) exam was used to screen individuals during the interviews. This exam is used as a screening test for dementia. The 3MS has added test items that sample a broader range of functions and difficulty levels while allowing a wider range of scores than the Mini Mental Status exam (MMSE). Additionally, this test has a higher reliability and validity of the scores achieved. The 3MS tests the individual's ability to orientate temporally and spatially, to recall personal information, to perform both rapid and long-term recall of 3 items, to recite in the forwards and backwards direction, to name highly recognizable objects, to generate a list of objects that meet a given criteria, to repeat a spoken sentence, to distinguish which members of a group are similar, to read and obey a written command, to write a spoken sentence, to copy a drawing, and to follow a three-stage command. These activities are scored on a scale from 0-100 and then adjusted for the education level of the individual being examined (Teng & Chui,

1987). Individuals scoring higher than an 87 were classified as “normal by screen.” Individuals with scores less than or equal to an 87 were examined with further cognitive tests using the neuropsychological battery developed by the Consortium to establish a Registry for Alzheimer’s Disease (CERAD) and the geriatric depression scale (GDS) (Morris et al., 1989).

A yearly case conference is held to review all information pertaining to the screening and further cognitive tests, and a consensus diagnosis is made that follows the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer’s Disease and Related Disorders Association (ADRDA) criteria. Definite Alzheimer’s disease can only be diagnosed if the patient has histopathological evidence of AD upon autopsy; therefore, no diagnoses of this class are made because autopsies are not permissible within the Amish culture. Probable AD is evidenced by clinical and neuropsychological examination that establishes dementia. Progressive cognitive impairments must be present in at least two areas of cognition and must be in the absence of other dementia diseases. Possible AD is diagnosed when dementia is present with an unknown etiology and no co-morbid diseases are believed to be the origin. Lastly, cognitive impairment no dementia (CIND) or mild cognitive impairment (MCI) can be diagnosed when dementia presents with focal signs, sudden onset, seizures or gait disturbances (G. McKhann et al., 1984).

For the dementia studies, individuals classified as “normal by screen” or “unaffected by exam” were categorized as healthy cognitive controls. Individuals classified “affected by history” or “affected by exam” with either possible or probable AD were categorized as

affected cases during the case conferences. Individuals who reported a history of or were diagnosed with CIND, MCI, stroke, non-Alzheimer dementia, neuropsychiatric disorder, impairment secondary to vascular injury, major depression, Lewey body dementia, trauma, progressive supranuclear palsy (PSP) or other disorders were classified as having an unknown affection status for the dementia study as the true AD affection status of these individuals is unclear.

For the 198 individuals missing the consensus diagnosis, the adjusted 3MS score was used to determine affection status with the same cut-off of 87. Individuals who achieved scores of greater than 87 were categorized as normal unaffected controls. Individuals with scores less than or equal to 87 (n = 9) were categorized as having unknown affection status as the etiology of the low score could not be determined (Khachaturian, Gallo, & Breitner, 2000). One individual scored a 22 on the unmodified Mini-Mental Status Exam (MMSE) and was categorized an unknown. Individuals ascertained for Parkinson's disease or autism studies were categorized as unknown controls if they lacked the proper cognitive test results. Additionally, individuals without screening, clinical, or cognitive data (n = 45) were classified as unknown controls as their true affection statuses could not be determined. Thirty-five of these individuals only had data on *APOE* status and lacked all clinical information.

In addition to comparing Amish LOAD cases and cognitively normal controls, cases and cognitively normal controls ascertained from a general clinical population were studied (Table II-1) (Naj et al., 2011). A collaborative study between researchers at the University of Miami and Vanderbilt University has ascertained Caucasian individuals

affected with LOAD unique from the Amish populations studied. These individuals have been diagnosed with probable or definite AD according to NINCDS-ADRDA criteria with an age of onset greater than 60. To make these diagnoses, documentation or a clinical history of significant cognitive impairment was present. Age- and gender-matched cognitively healthy controls were ascertained from the same regions and had a documented 3MS or MMSE score in the normal range. As the Amish are founded from European immigrants, this European-American dataset of unrelated individuals is of similar ancestry. As far as it was able to be determined, the samples in this dataset were not related to each other based upon the amount of sharing between individuals across a set of previously genotyped genome-wide data.

Table II-1. Demographics of Genetic Risk Score Samples.

Cohort	Affection status	Female	Total	Average age of exam/onset (standard deviation)
Amish	**LOAD case**	**63%**	**126**	**78 (7.75)**
	Cognitively normal control	**58%**	**503**	**79 (6.72)**
Unrelated	**LOAD case**	**63%**	**473**	**74 (8)**
	Cognitively normal control	**60%**	**498**	**74 (8)**

Risk loci used to estimate total burden

Total genetic risk score was calculated for each individual in the two study populations. Twenty-one single-nucleotide polymorphisms (SNPs) that reached genome-wide significant level in the most recent LOAD GWAS, referred to as GWAS hits in subsequent sections and chapters, were genotyped in the full Amish dataset (Lambert et al., 2013). This study used a genome-wide significant level of $p < 5 \times 10^{-8}$, which corrects the type I error for approximately 1,000,000 independent tests. Previous genotypic data for *APOE* were used to include this major genetic risk variant in the analysis (Cummings et al., 2012).

The genotyping and quality control measures for these data are described in more detail in Chapter IV. In summary, these GWAS hits were genotyped with additional variants in three Sequenom MassARRAY pools for both populations. Two GWAS hits failed to genotype via this method. Two additional GWAS hits were removed from analysis due to low calling efficiency (less than 95%). Samples from both populations were removed if they were genotyped in duplication, had low genotyping efficiency (less than 95%). Additionally for the Amish, samples that could not be related to other individuals were removed. Seventeen of the 21 GWAS hits passed this QC are were used to estimate total genetic risk (Table II-2).

Table II-2. Details of Risk Loci from Meta-Analysis Used to Calculate Total Genetic Risk Score. Alleles (major/minor), MAF, and overall OR Adapted from Lambert, et al, 2013. Chr = chromosome. Pos = position in bp. MAF = minor allele frequency. OR = odds ratio. Allele frequency was calculated using the 921 Amish samples and the 971 samples from the unrelated dataset that passed QC in the follow-up genotyping phase (Chapter IV). Each individual marker beta (converted from the published OR) was divided by the sum of all marker betas to calculate marker weights.

Marker	Chr	Position	Gene	Alleles	MAF	Overall OR	Amish MAF	Unrelated MAF	Weights
rs6656401	1	207692049	*CR1*	G/A	0.20	1.18	0.24	0.18	0.052
rs6733839	2	127892810	*BIN1*	C/T	0.41	1.22	0.45	0.40	0.062
rs35349669	2	234068476	*INPP5D*	C/T	0.49	1.08	0.45	0.50	0.024
rs190982	5	88223420	*MEF2C*	A/G	0.41	0.93	-	-	-
rs9271192	6	32578530	*HLA-DRB5/HLA-DRB1*	A/C	0.28	1.11	0.18	0.28	0.033
rs10948363	6	47487762	*CD2AP*	A/G	0.27	1.10	-	-	-
rs2718058	7	37841534	*NME8*	A/G	0.37	0.93	0.29	0.35	0.023
rs1476679	7	100004446	*ZCWPW1*	T/C	0.29	0.91	0.28	0.29	0.030
rs11771145	7	143110762	*EPHA1*	G/A	0.34	0.9	0.27	0.32	0.033
rs28834970	8	27195121	*PTK2B*	T/C	0.37	1.10	0.32	0.35	0.030
rs9331896	8	27467686	*CLU*	T/C	0.38	0.86	0.36	0.41	0.047
rs10838725	11	47557871	*CELF1*	T/C	0.32	1.08	0.35	0.31	0.024
rs983392	11	59923508	*MS4A6A*	A/G	0.40	0.90	-	-	-
rs10792832	11	85867875	*PICALM*	G/A	0.36	0.87	0.45	0.35	0.044
rs11218343	11	121435587	*SORL1*	T/C	0.04	0.77	0.05	0.04	0.082
rs17125944	14	53400629	*FERMT2*	T/C	0.09	1.14	0.05	0.11	0.041
rs10498633	14	92926952	*SLC24A4/RIN3*	G/T	0.22	0.91	0.20	0.22	0.030
rs8093731	18	29088958	*DSG2*	C/T	0.02	0.73	0.01	0.01	0.099
rs4147929	19	1063443	*ABCA7*	G/A	0.19	1.15	-	-	-
rs3865444	19	51727962	*CD33*	C/A	0.31	0.94	0.29	0.30	0.019
rs7274581	20	55018260	*CASS4*	T/C	0.08	0.88	0.10	0.08	0.040
***APOE* E4**	19	19q13.2	*APOE*	-	-	2.5	0.14	0.26	0.287

Estimation of total genetic risk score

The weighted genetic risk score was calculated by multiplying the number of risk alleles at each marker by the weight for that marker, and then summing across all markers (Equation 1). To determine the weight for each marker, the published odds ratio (OR) for the minor allele was converted to an OR for the risk allele. This was then converted to a beta effect by taking the natural logarithm of the risk OR. Each individual marker beta was divided by the sum of all marker betas (Equation 2, Table II-2). For *APOE*, the ε4 allele was coded as the risk allele and an OR of 2.5 was used in the weighting scheme (Lambert et al., 2013; Naj et al., 2011; Reitz, Brayne, & Mayeux, 2011; Strittmatter et al., 1993). No individual from either population was missing genotypes for more than three GWAS hits. If an individual was missing a genotype for a marker, the average allele frequency for the respective parent population was used to determine the average number of risk alleles carried. For the Amish population, an allele frequency that had been corrected for the relatedness of the individuals was used. MQLS (detailed in Chapter III) uses kinship coefficients to estimate and correct for the relatedness of the Amish samples when calculating allele frequencies and testing for association.

Equation 1. Total Genetic Risk Score Estimation. GRS_i = genetic risk score for the i-th individual. w_j = weighted effect size for the j-th GWAS SNP. x_{ij} = risk allele count for the i-th individual of the j-th GWAS SNP. n = total number of GWAS SNPs.

Equation 2. Weighted Effect Size Calculation. w_j = weighted effect size for the j-th GWAS SNP. $beta_j$ = published effect size of the j-th GWAS SNP. n = total number of GWAS SNPs.

Regression analysis determines if there is a relationship between two variables, in this case, total genetic risk score and affection status or population. To compare cases to cognitively normal controls, logistic regression was performed (R, version 3.0.2) to estimate the correlation between these two variables in both populations. Moreover, Amish LOAD cases were compared to unrelated LOAD cases and Amish cognitively normal controls to unrelated cognitively normal controls. As the total genetic risk scores calculated for each of the Amish individuals is likely to be correlated with the scores for related Amish individuals, generalized estimating equation (GEE) was used to estimate a generalized model that incorporates the correlation between outcomes (Halekoh, Hojsgaard, & Yan, 2006; Liang & Zeger, 1986). This method has been employed in the "geepack" package available in R. By converting the kinship coefficient matrix generated by KinInbcoef (see Chapter III) to a correlation matrix, the relatedness of individuals was included in the analysis.

Results

To determine if the known genetic etiology of LOAD in general European Caucasian descent populations also impacts LOAD in the Amish, the total genetic risk score using known LOAD risk alleles was calculated and compared across affection and population groups (Figure II-1). When comparing this genetic risk score, Amish cases harbored a significantly higher burden of the known risk alleles ($\mu = 0.94$ genetic risk score)

compared to Amish cognitively normal controls (μ = 0.86) (logistic regression, $p = 5.99 \times 10^{-6}$). As expected, the unrelated cases also had a significantly higher burden (μ = 1.05) when compared to the unrelated cognitively normal controls (μ = 0.85) ($p < 2 \times 10^{-16}$). When compared to unrelated cases, Amish cases had a significantly lower burden of known risk alleles ($p = 2.56 \times 10^{-6}$). Cognitively normal Amish controls were not different from the unrelated controls (p = 0.381). When APOE was evaluated independent of the GWAS hits, the Amish LOAD cases were more different (μ = 0.45 risk alleles) than the unrelated cases (μ = 0.82) ($p = 9.76 \times 10^{-8}$) (Figure II-2,Table II-2). When affection status and population groups were compared using GEE instead of regression, the trends were similar but the p values from the GEE comparisons of Amish cases to Amish controls and Amish cases to unrelated cases were much more significant (Table II-3).

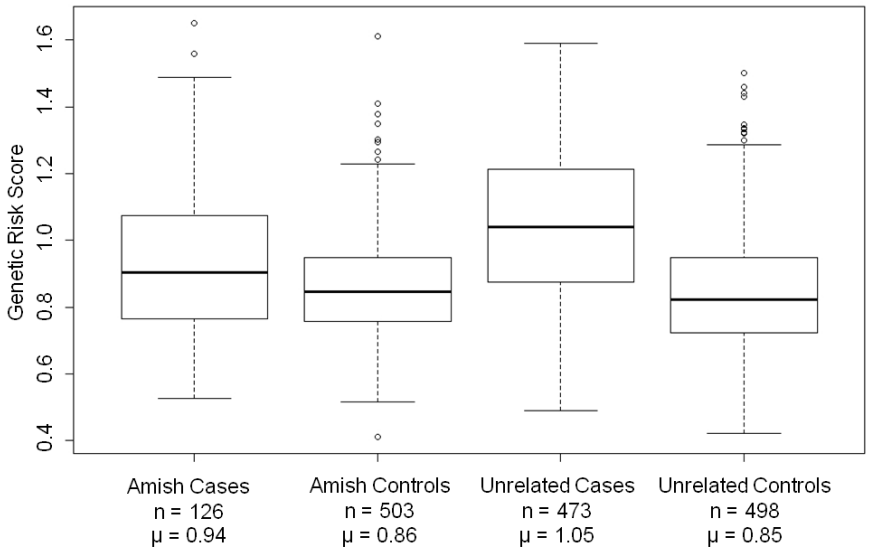

Figure II-1. Distributions of Total Genetic Risk Scores. Total genetic risk score averages and standard deviations were calculated for the 629 Amish LOAD cases and cognitively normal controls and the 971 LOAD cases and cognitively normal controls from the unrelated case-control dataset who passed QC for the follow-up genotyping phase. n = total number of individuals. μ = average total risk score for group.

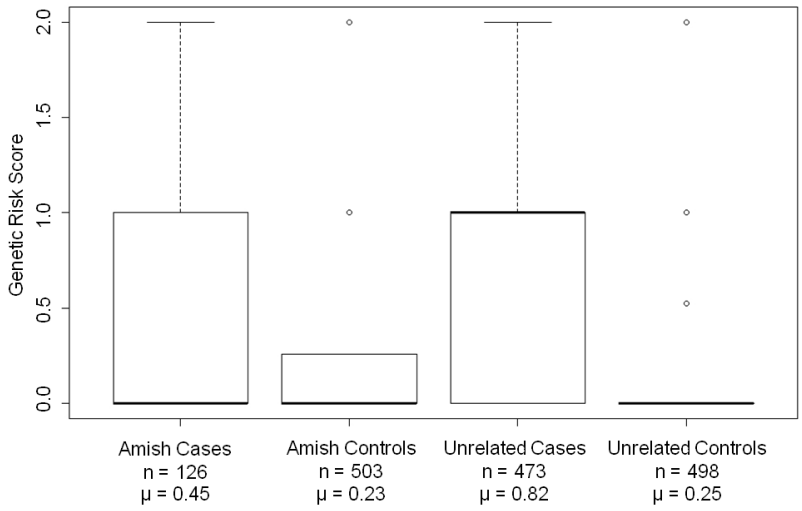

Figure II-2. Distributions of Genetic Risk Scores for *APOE* Only. Genetic risk score averages and standard deviations were calculated for the 629 Amish LOAD cases and cognitively normal controls and the 971 LOAD cases and cognitively normal controls from the unrelated case-control dataset who passed QC for the follow-up genotyping phase. n = total number of individuals. μ = average *APOE* risk score for group.

Table II-3. Comparison of statistical model p-values for genetic risk score analysis.

Group 1	Group 2	Regression p value	GEE p value
Amish cases	**Amish controls**	5.99×10^{-6}	$< 2 \times 10^{-16}$
Amish cases	**Unrelated cases**	2.56×10^{-6}	$< 2 \times 10^{-16}$
Amish controls	**Unrelated controls**	0.38	0.83

Discussion

The Amish cases tended to have a lower genetic risk score than the unrelated cases. This result suggests that the common variants implicated by GWAS explain a smaller proportion of genetic risk in the Amish than in the general population. This result is consistent with the lack of significant association observed for these risk loci in previous studies in the Amish. However, since Amish cases did tend to have a higher burden when compared to cognitively normal controls from the same population, these known risk loci do explain some of the expected genetic effects. The lack of correlation between

total risk and parent population for cognitively normal controls suggests that the Amish controls are genetically similar to controls from the general population. In the Amish from Elkhart, LaGrange and Holmes Counties, the *APOE* ε4 allele has a frequency of 0.18 in cases and 0.06 in Adams County, but in cases from the general Caucasian population this risk allele frequency is 0.42 (Corder et al., 1994; Cummings et al., 2012). This allele frequency disparity may in part explain the increase in difference in genetic burden between cases from the two datasets when only APOE was analyzed, but additional factors are likely to contribute as well.

The effect sizes published by the GWAS were estimated from a dataset of unrelated individuals distinct from the Amish populations. The unrelated individuals studied in this project were a subset of the meta-analysis in which the associations were detected. The largest difference across all comparisons was between the unrelated cases and cognitively normal controls. This result is consistent with the reported effect sizes and risk alleles and suggests total genetic risk was validly estimated.

The difference in genetic burdens may suggest different underlying genetic architecture between the two populations, resulting in different effect sizes or allele frequencies. Therefore, the effect sizes used to estimate total genetic risk may not reflect the true effect size of the risk locus in the Amish. Alternatively, the lower total risk harbored by the Amish may be explained by differences due to the risk allele. The risk alleles reported by the GWAS were also determined from a dataset distinct from the Amish populations. These markers are not likely to be the functional marker, but are likely to be in linkage disequilibrium with the true risk marker. It is possible that an alternate marker

would be a better surrogate in the Amish. As stated previously, there is a large difference in allele frequency for the *APOE* ε4 risk locus. Additionally, the overall minor allele frequencies between the two populations differ for multiple risk loci. *PICALM*, *SORL1*, *FERMT2*, *CR1*, *CASS4*, *DSG2*, and *HLA-DRB5*/*HLA-DRB1* all have at least a 20% frequency difference between the two populations. These large differences may be contributing to the significant differences in genetic burden and support the hypothesis that additional variation in or near these risk genes is contributing to disease susceptibility in the Amish because of differences in underlying genetic architecture. Exonic variation in these risk regions was investigated with additional genomic regions implicated by previous studies in the Amish, and is described in the following Chapters.

CHAPTER III
IDENTIFICATION OF VARIANTS FROM EXOME SEQUENCES

Introduction

The known genetic risk loci do not explain the total expected genetic risk in the general population, and the analysis of total risk burden suggests they explain even less in the Amish. Therefore, additional variants in known genes or previously unassociated genes may also confer susceptibility. Through the identification of additional risk variants or loci, more can be learned about the underlying biology and pathogenesis of AD that can inform future studies about diagnosis and treatment targets.

The common disease multiple rare variant hypothesis states that common diseases may be influenced by multiple causal, but very rare variants in one or more genes that have large effect sizes. This suggests additional rare variants with larger effects may explain unknown genetic risk for LOAD, as opposed to the common risk loci implicated by GWAS. Previous genetic analyses have been limited by molecular technology and statistical methods available for genotyping and analysis. New technologies for sequencing and statistical methods for variant analysis have allowed for the efficient and effective interrogation of rare variation. The clinical utility of genetic variation in disease prediction is complex and depends on many factors including the predictive value of the risk allele and the therapeutic implications for an asymptomatic at-risk individual (Manolio, 2013). If rare functional variants have a larger effect size and are more predictive of disease, there may be more clinical utility using these variants to identify at-risk individuals who might benefit from early treatment or increased screening. Whole-

genome sequencing determines the complete DNA sequence. Whole-exome sequencing is a targeted approach that only sequences the coding portion of the genome. By studying rare variants identified from whole-exome sequencing, portions of the human genome previously unstudied under the common disease common variant hypothesis can be interrogated.

While the costs for sequencing a whole-exome have decreased dramatically and will continue to do so, the sample sizes needed to have sufficient power to detect an association with a rare variant can still be cost prohibiting. For example, if a variant has a minor allele frequency of 1% and an odds ratio (OR) of 2, over 2,000 cases and 2,000 controls would be needed to have 80% power to detect an association if the type I error rate is 0.05. There are many ways to overcome this limitation, two of which have been employed in this study. First, as described in detail in Chapter I, isolated founder populations are advantageous for genetic studies for many reasons. The severe bottleneck that occurs in a founder population can alter allele frequencies. If a rare variant is carried by a founder and propagated through subsequent generations, this previously rare variant may be enriched in the isolated population. The power to detect an association increases because the allele frequency increases. Second, instead of sequencing all available samples and performing case-control association testing on all variants identified, a subset of individuals can be used. By selecting the individuals most likely to harbor genetic variation that is contributing to risk, and then screening this sequence data for candidate variation that can be genotyped through a more cost effective method in the full dataset, the limiting costs can be overcome.

Methods

Selection of subset of the Amish population for sequencing

From the larger dataset of 1,119 Amish individuals described in Chapter II, 176 individuals were selected for whole-exome sequencing for three different studies (LOAD, Parkinson's disease and age-related macular degeneration). One-hundred fifteen individuals were chosen for LOAD using the following prioritization (a) large sibships with both affected and unaffected individuals, (b) close relatives of sibships in (a), (c) *APOE* 2/3 and 3/3 affected individuals and their unaffected siblings, and (d) members of subpedigrees with the highest logarithm of odds (lod) scores from previous genetic linkage studies (Cummings et al., 2012). It was hypothesized that these cases and controls were the most likely subset in which to identify unidentified variants that confer risk to LOAD.

For the Parkinson's disease (PD) sequencing project, all 32 affected individuals were selected for sequencing. Up to two unaffected full siblings of those cases were chosen as controls, this consisted of 26 unaffected individuals and two individuals with unknown PD status. These individuals with an unclear affection status have been diagnosed with progressive supranuclear palsy (PSP), a neurodegenerative tauopathy. Symptoms include slowed movements and gait difficulty that can be confused for PD. Since PD is also a neurodegenerative disease and may co-occur with LOAD, all of the PD affected individuals are considered to have an unknown LOAD affection status. As described in future sections, the association software used in this study allows for the inclusion of unknown controls to increase sample size in a study of related individuals.

Four Amish individuals were chosen for the age-related macular degeneration (AMD) sequencing project. During the initial baseline screen for the dementia studies, this single nuclear family was identified to have three AMD affected siblings and eight unaffected individuals. The affected members lacked the known risk alleles in *CFH* and *ARMS2*. At the time, three affected and one unaffected members were selected as the other four members were unavailable for clinically evaluation for AMD. All four individuals were categorized as cognitively normal controls for the LOAD project. By including individuals from the additional studies, sample sizes were increased and more sequence data was available for the processing steps detailed below.

Whole-exome sequencing

Whole-exome sequencing was performed on DNA extracted from these selected individuals' blood. The DNA for this project was allocated and sequenced by the respective DNA banks and sequencing cores at both the Hussman Institute of Human Genomics at the University of Miami and the Center for Human Genetics Research at Vanderbilt University. The Agilent SureSelect Human All Exon 50 Mb capture kit was used to capture the exonic genomic DNA. This kit captures 50Mb targets consisting of all coding exons annotated by the GENCODE project and all exons annotated in the consensus coding sequence (CCDS), plus 10 base pairs of flanking sequence for each targeted region. Additionally, the kit captures small non-coding RNAs. The technology captures almost 80% of the exome sequenced at 20X coverage and 77% of the capture is on-target ± 200bp ("Datasheet 5990-6319EN," 2010).

This exonic library was then sequenced on the Illumina HiSeq 2000 with paired ends and read lengths of 75 base pairs. Next-generation sequencing (NGS) can be applied to *de novo* sequencing and re-sequencing. These technologies offer better detection of rare variants and more sequence information on a larger scale than previous sequencing methods. The Illumina sequencing-by-synthesis chemistry uses bridge amplification and reverse terminator chain sequencing. Bridge amplification generates clusters from a single fragment of DNA. After exome capture, adaptors are added to DNA fragments and then these fragments are immobilized on the surface. Amplification proceeds in cycles as nucleotides and polymerase are added to grow clusters. This process creates a cluster consisting of identical pieces of DNA. After bridge amplification, reverse terminator chain sequencing is performed. As a single base is incorporated, a fluorophore corresponding to the nucleotide is released by the polymerase. The polymerase can only incorporate a single base at a time allowing each base added to be detected. Released fluorophores are detected simultaneously by a four camera system. The overall miscall rate is typically reported around 1% and is largely due to desynchronization within clusters (Nielsen, Paul, Albrechtsen, & Song, 2011).

Processing of raw sequences and calling of variants

Sequence processing consisted of aligning reads, removing duplicates, realigning around local indels, recalibrating quality scores and calling of variants (Figure III-1). Using BWA (version 0.6.2), raw sequences reads were aligned to the UCSC hg19 human reference genome. Picard tools (version 1.74) was used in the process of marking duplicates. All steps performed in the Genome Analysis Tool Kit (GATK, version 2.1-10) following the best practices available at the time of processing, which consisted of local realignment around indels, base recalibration, variant calling (using the

UnifiedGenotyper), and variant recalibration. The reference bundle was downloaded from GATK and used for all processing steps.

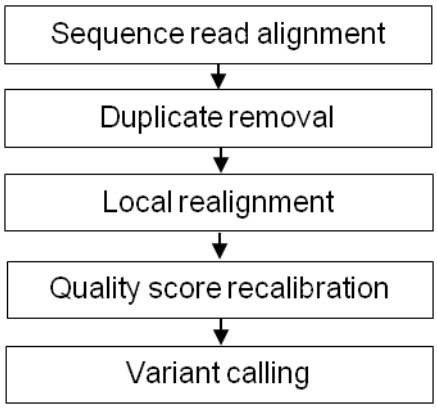

Figure III-1. Sequence Processing Pipeline. Read alignment performed by BWA, duplicate removal by Picard tools, and all other steps by GATK.

The first step in processing was to map reads to the genome. Aligning reads to a reference sequence reconstructs the individual's genome and then allows for mutations or indels to be identified. Due to the large amount of data and reads from a single run of the sequencer, algorithms for alignment must be able to efficiently and accurately map the data. Incorrect mapping may lead to errors in variant calls, so the aligner must handle both sequencing errors and real variation from the reference. BWA, which uses the Burrows-Wheeler transform (BWT) to compress data and align short sequences rapidly, was used to align the raw sequence reads. In BWT, substrings or characters that occur often are repeated in a row and can be compressed by a number of data compression methods. These types of aligners are fast, memory-efficient and are useful for mapping repetitive reads (Nielsen et al., 2011). BWA performs gapped global alignment for paired-end reads for short reads (up to 200bp) (Li & Durbin, 2009). When

aligning, the user can specify such values as maximum differences allowed, maximum gap and seed size as well as disallow long deletions or indels.

Paired-end reads are advantageous for alignment because each end is mapped to the genome, and from these positions, the aligner estimates insert size distribution and pairs the mapped locations (Li & Durbin, 2009). This information can be combined with the known fragment length from sample preparation and capture to select the paired locations that correspond with expected insert size. Therefore, paired-end reads provide more information for mapping sequence to the reference genome and reduce the number of locations to which a single read may map. Additionally, paired-end reads can reduce the difficulty associated with alignment for regions of high diversity, such as the major histocompatibility complex (Li & Durbin, 2009). For each individual sequenced, the two sequence files, one for each paired-end, were aligned to the most recent reference genome, hg19, released by UCSC in February 2009. This reference genome is also known as Genome Reference Consortium GRCh37. Then the two aligned files were merged into a single file for each individual.

Molecules of DNA can be sequenced in duplication, but the nature of high-throughput next-generation sequencing (NGS) would interpret duplicates as increased coverage of a site and therefore may affect the probability that a variant is real. Most software tools developed for NGS allow for the marking of duplicate reads and subsequent processing steps will ignore these duplicates. Duplicates were marked in this dataset using Picard tools. Realignment accounts for errors in the original alignment process. Local realignment around indels corrects mismatches that may affect variant calling. If an indel

is present in a read, bases that follow it will be mismatched to the reference genome and might look like variants. The reads were realigned around known sites of indels as suggested by GATK's best practices as it is efficient and little coverage is needed (DePristo et al., 2011).

One way to account for the errors and biases associated with NGS is through a phred quality score recalibration for each base. This phred quality score reflects the probability that the called base is an error and is generated by the sequencer incorporating the ambiguity of the fluorescent signal, the quality of neighboring bases and the quality of the entire read. The scale is bounded with a highest possible value of Q40, which corresponds to a probability of 0.0001 that the base is incorrect. The proportion of bases with quality scores of at least Q30 ($p = 0.001$ that base has been called incorrectly) is estimated at 74-80% (Minoche, Dohm, & Himmelbauer, 2011). However, this score may not accurately reflect the true error rate for this base. GATK recalibrates the raw quality score by incorporating the position of the base in the read, dinucleotide content, and the read group. Non-polymorphic sites act as controls and variants are not expected to be present. For these non-polymorphic sites, the software estimates the residual differences between the mismatch rate based on the raw quality score and the real number of mismatches compared to the reference genome (Nielsen et al., 2011). Variant calling algorithms then use this recalibrated quality score in determining the probability of a variant and the genotype likelihoods.

At this point in processing, GATK's DiagnoseTargets was used to analyze the coverage distribution for all samples for the intervals specified by the capture technology. This tool

categorizes regions or intervals of the genome with bad coverage, mapping or read mating. Across all samples, 78% of the intervals passed the filters used by DiagnoseTargets (Table III-1). Intervals could be assigned multiple flags describing coverage, mapping and mating. Of the intervals that did not pass, 22% were sequenced with low or insufficient median depth across samples. Low coverage was seen in 11% of intervals. This meant that there was less than the minimum depth at the locus, after applying the filters specified in the previous processing steps. Gaps in coverage, or absolutely no coverage, were observed in 2.5% of the intervals. The remaining categories of excessive coverage, bad mate, and poor quality were given to less than 1% each of intervals. Excessive coverage is defined as more than the specified maximum read depth at the locus, indicating some sort of mapping problem. Reads that are not properly mated suggest mapping errors. Poor quality indicates poor mapping quality of the reads if a fraction of all read in the intervals had low quality.

Table III-1. Summary of Exome Interval Coverage.

Category	Percent of Intervals
Pass	78.0
Low median depth	22.0
Low coverage	10.8
Coverage gaps	2.5
Poor quality	< 1
Bad mate	< 1
Excessive coverage	< 1

After sequence reads were mapped to the reference genome, a variant caller was used to identify those bases that are statistically different from the reference. GATK had two available variant callers, HaplotypeCaller and UnifiedGenotyper. The newer HaplotypeCaller uses both local *de novo* assembly and an advanced hidden Markov model (HMM) likelihood function. However, this caller is new and hasn't yet been fully

tested and debugged. Therefore variants were called using the UnifiedGenotyper (DePristo et al., 2011). This algorithm uses a Bayesian genotype likelihood model to estimate allele frequencies and likely genotypes across many samples. The posterior probability estimates variant alleles segregating at the locus and also the genotype for each sample. The recalibrated quality scores are used to eliminate and filter most of the false positive variants (DePristo et al., 2011). Best practice guidelines recommended a minimum confidence score threshold of Q30 for the coverage expected for our data (DePristo et al., 2011). By changing this threshold value, the number of false positive variants in the dataset can be affected. After the QC detailed below, the average coverage was the dataset was 58.60 ± 13.53, within the expected range. The input for the UnifiedGenotyper is the read data after processing, and the output is a multi-sample unfiltered variant call format (vcf) file. Multi-sample variant calling was performed using the intervals specified by the exome capture targets. This was done in batches by splitting the interval file every 5000 lines to run multiple processes in parallel and to decrease time and computing space needed.

The variant quality score recalibrator (VQSR) removes false positive machine artifacts. This step estimates the probability that the called variant is a true variant. For the VQSR to achieve the best results, at least 30 samples are needed (DePristo et al., 2011). Given the sample size of 176 for this project, VQSR was not limited by the number of samples available. The model used by VSQR is determined based on real variants provided in the input files, such as those polymorphic sites in HapMap on SNP chip arrays. This adaptive model is then applied to all variants discovered to determine the probability of the site being a real variant. This probability is reported as the log odds ratio (VQSlod) of a true variant versus a false variant under the Gaussian mixture model

(DePristo et al., 2011). By applying the error model built to the raw vcf file, a new recalibrated vcf file was generated and was ready for analysis.

Data management was performed using vcftools (version 0.1.9) for additional QC steps. Samples with an average read depth less than 30 (n = 9) were removed from the dataset (Table III-2). These samples also corresponded with samples that were called with low efficiency. The full sequencing dataset had previously been genotyped on the Affymetrix 6.0 GeneChip Human Mapping 1 million array set. The genotypes determined from the sequence data were compared to these previous genotypes to check the concordance rate. Of the 170,849 sequencing variants, 8,268 overlapped with the 610,611 SNPs passing QC from the previous GWAS study. Samples that had a concordance rate less than 90% (n = 3) were removed as confidence would be low in the validity of the variant calling. The remaining discordant genotypes were largely due to heterozygotyes in the GWAS genotyping being called as homozygotes in the sequence data and vice versa. A small percentage was due to homozygotes for one allele being called as homozygotes for the alternate allele. Gender determined by the percentage of X heterozygosity was compared to the genders recorded in the clinical data. Two samples were discordant and the genotypic gender did not agree with the charts, therefore these samples were removed from subsequent analysis.

Table III-2. Summary of Sample Quality Control Measures.

Samples	Cases	Controls	Unknowns	Total
Sequenced	59	68	49	176
Depth < 30	-5	-1	-3	-9
Concordance < 90%	-1	-1	-1	-3
Gender error	0	-1	-1	-2
Analyzed	**53**	**65**	**44**	**162**

There were a total of 170,849 variants called across the 162 whole exomes that passed the above QC measures (Table III-3). Of these variants, 153,272 passed the processing filter of a minimum phred-scaled quality threshold of 10. To control for missing data, variants with a calling efficiency of less than 80% were removed from analysis. The available statistical software for association analysis in a complex population like the Amish (detailed below) is restricted to testing biallelic markers therefore any multiallelic markers cannot be analyzed and were removed. As the exome was targeted and enriched for at the beginning of the sequencing preparation, all off-target intergenic and intronic variants were removed from further analysis. After this QC, 162 individuals and 79,203 biallelic variants were analyzed (Table III-4). This QCed dataset was 99.1% concordant for 8,268 exonic variants overlapping with previous genotyping and individuals were sequenced at a depth of 58.60 ± 13.53 (Figure III-2, Table III-5).

Table III-3. Summary of Variant Quality Control Measures.

Variants	Count
Variants called	170,849
Pass processing filters	153,272
Efficiency ≥ 80%	141,534
Biallelic	141,044
Exonic	**79,203**

Table III-4. Demographics of Sequencing Samples. Age of exam and onset averages and standard deviations were calculated for the 162 samples that passed QC for whole-exome sequencing.

Affection status	Female	Total	Average age of exam/onset (standard deviation)
LOAD case	55%	53	78 (6.92)
Cognitively normal	62%	65	76 (7.21)
Unclear or unknown	41%	44	78 (7.60)

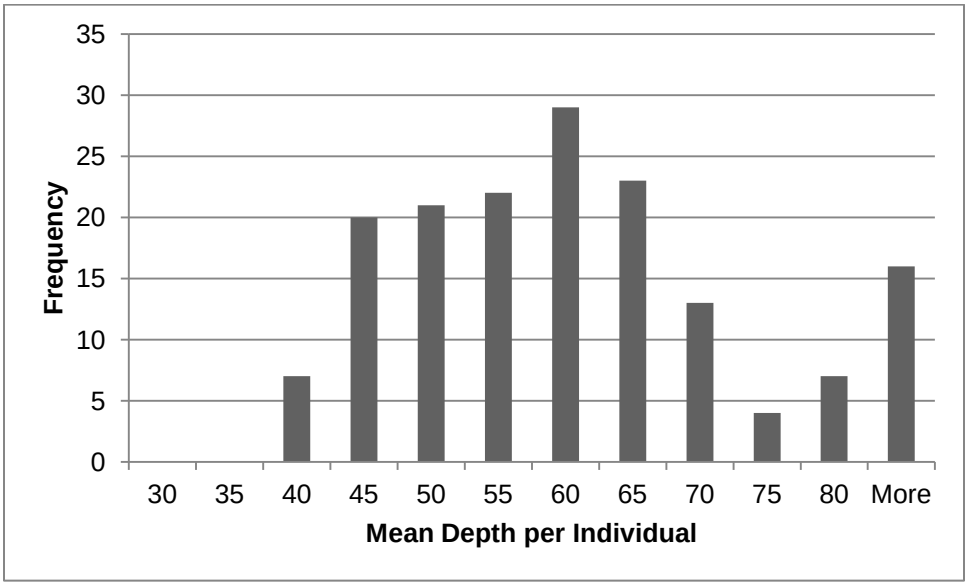

Figure III-2. Distribution of Mean Depth per Individual. Mean depth across 79,203 biallelic exonic sites calculated by vcftools for the 162 individuals passing the above QC.

Table III-5. Summary of Mean Depth per Site. Mean depth across 162 individuals calculated by vcftools for the 79.203 biallelic exonic sites passing the above QC. Percent Within = percent all of sites with at least that coverage.

Coverage	Percent Within
0X	100
10X	93.0
20X	82.0
30X	70.8
40X	59.6
50X	49.0
60X	39.0
70X	30.6
80X	24.0

Annotation of variants

Variants were annotated by SeattleSeq (online resource, version 134) and by ANNOVAR. These programs annotate each variant for a variety of important information, functions, and statistics. SeattleSeq was used to annotate the gene and function for all variants. Even though the capture intervals used during variant calling were specified by the exome capture technology, intronic and intergenic variants occur within these intervals and were therefore in the dataset. By annotating these variants with the gene and function from SeattleSeq, only exonic or variants in untranslated regions (those targeted by the molecular study design) were included in the association analysis and prioritization pool. Additionally, it was used to annotate PolyPhen prediction, Grantham scores, and conservation scores (phastCons and GERP) for each variant. PolyPhen scores predict whether an amino acid substitution is damaging, affects protein function, or if it is benign, lacking phenotypic effect (Ramensky, Bork, & Sunyaev, 2002). Grantham scores are based upon a formula to quantify the chemical dissimilarity of the amino acid substitution (Grantham, 1974). Lower scores correspond to more conservative substitutions and higher scores are more radical. PhastCons estimates the degree of evolutionary conservation among vertebrate genomes based upon phylogenetic tree (Siepel et al., 2005). Genomic evolutionary rate profiling (GERP) identifies constrained elements by quantifying rejected substitutions (Cooper et al., 2005). ANNOVAR was used to annotate a variant's inclusion in three catalogs of human variation: dbSNP build 137, ESP 6500 release, and 1000 Genomes April 2012 release (Wang, Li, & Hakonarson, 2010).

Analysis of single variant in cases versus controls

After variants had been called across all sequenced samples, this information was used to perform a case-control analysis of exonic variants. To focus further analysis on the variants most likely to contribute genetic risk to LOAD in the Amish, three classes of variants were screened. The classes of genes include 26 genes previously implicated in LOAD through GWAS, genes known to carry early-onset mutations, and genes located in four previously identified candidate linkage regions (Cummings et al., 2012; Harold et al., 2009; Hollingworth et al., 2011; Lambert et al., 2009; Lambert et al., 2013; Naj et al., 2011; Seshadri et al., 2010). These three classes of genes are the most likely to harbor variants that contribute to LOAD susceptibility in the Amish. In addition, and as a secondary screen, single variant case-control analysis was performed for all 79,203 sequencing variants to test for possible association with LOAD.

Known familial relationships were incorporated with the genetic information in the Modified Quasi-Likelihood Score (MQLS) test. This program accounts for the relatedness of individuals through kinship coefficients and corrects for pedigree structure while testing for association between a genetic marker and a binary trait (Thornton & McPeek, 2007). The KinInbcoef software is used to calculate kinship and inbreeding coefficients based upon the pedigree structure. The kinship coefficient is a measure of relatedness between two individuals and is the probability that two alleles sampled at random from each individual are identical and inherited from a common ancestor. The inbreeding coefficient is calculated for each individual by measuring the kinship coefficient between the individual's parents. The AGDB (see Chapter I) generated a pedigree consisting of 5,437 members that connects all analyzed samples across 13

generations. By specifying this pedigree structure to KinInbcoef, kinship coefficients were calculated for every possible pair in the full dataset.

The inputs for MQLS are the calculated coefficients, a marker data file such as a modified vcf file, and an estimated prevalence of the binary trait in the general population. To run multiple processes in parallel, the full set of variants was divided into subsets of 2000 variants each. The output is a p-value for each marker corresponding to the MQLS statistic and its place among the chi-squared null distribution. To correct for multiple testing, a Bonferroni correction was applied. While such a correction is considered overly conservative, this method will reduce Type I error and its conservative nature is not extreme (Lander & Botstein, 1989; Sidak, 1968). This determined a threshold of significance based upon the number of variants tested.

Prioritization of identified variants

To overcome low power due to small sample size in the initial screening population, exonic variants from the three classes of genes screened were prioritized for follow-up analysis in the full dataset. The additional information from this analysis may tease out which variants are contributing to the previously significant results or to identify new risk variants in known loci. Two criteria were used for prioritization. First, any variant with a nominally significant p-value (< 0.01) was chosen. This cut-off was chosen to reduce type II error, while accepting a high type I error rate. Second, variants were chosen if it was apparently novel by not being present in three catalogs of human variation (dbSNP build 137, ESP 6500 release, and 1000 Genomes April 2012 release). All variants prioritized for further analysis had a VQSLOD score greater than 2 signifying that the

chances the variant is a true variant is 100:1. As the larger Amish population has been studied previously, all variants that overlapped with the previous genome-wide analyses were omitted from further analysis to avoid redundant testing.

For the initial screen, each GWAS hit had been assigned to the gene in the LD block or to the nearest gene by the publishing researchers. However, a more thorough analysis of these published associations resulted in an expanded gene list based upon LD patterns determined by the CEU genotypes available in HapMap via downloads from Haploview (version 4.2). These newly implicated genes were then screened for additional variants, but were not included in the analyses of the full Amish dataset due to cost constraints.

In addition to screening the three classes of variants described previously, variants that were unique to cases or unique to controls were investigated. Because these variants only occurred in the exomes of cases or controls, the power to detect an association may have been too low in this subset analysis. By prioritizing these variants for genotyping in the full dataset, novel associations with previously unknown genes or pathways may be identified. To increase the likelihood that these uniquely observed variants are contributing to disease risk and are not derived from a *de novo* mutation and propagated in a subpedigree, the variant must be carried by at least 10 “unrelated” individuals. “Unrelated” in this context is defined as individuals who are related to other carriers with a kinship value less than that for first cousins.

Results

Screen of candidate genes for association with LOAD

The exomes sequenced harbored 155 exonic variants in the known AD genes screened (Table III-6). The most significant p-value among these genes was 0.0098 for position 10,054,789 on chromosome 19 in *ABCA7*. Within the candidate linkage regions, 557 exonic variants were identified and the most significant p-value was 0.00017 on chromosome 3 (Table III-7).

Table III-6. Summary of variants identified that are within or very near known AD genes. Counts are displayed for the number of variants present in the human variation catalogs of dbSNP build 137 (dbSNP), ESP 6500 release (ESP), and 1000 Genomes April 2012 release (1000G). The number of novel variants identified in each implicated gene is also shown. (*) Closest gene to GWAS hit.

Gene	Location	Variants	dbSNP	ESP	1000G	Novel
ABCA7*	19p13.3	20	19	18	18	1
APOE	19q13.2	0	0	0	0	0
APP	21q21.3	1	1	1	0	0
BIN1*	2q14	3	3	3	3	0
CASS4*	20q13.31	9	9	8	9	0
CD2AP*	6p12	1	1	1	1	0
CD33*	19q13.3	1	1	1	1	0
CELF1*	11p11	1	0	0	0	1
CLU*	8p21-p12	2	2	2	2	0
CR1*	1q32	9	9	9	9	0
DSG2*	18q12.1	8	7	7	5	1
EPHA1*	7q34	3	3	3	3	0
FERMT2*	14q22.1	5	4	4	4	1
HLA-DRB5/DRB1*	6p21.3	0	0	0	0	0
INPP5D*	2q37.1	2	2	2	2	0
MEF2C*	5q14	1	1	1	1	0
MS4A*	11q12.2	45	42	41	35	1
NME8*	7p14.1	0	0	0	0	0
PICALM*	11q14	2	2	2	2	0
PSEN1	14q24.3	1	1	1	1	0
PSEN2	1q31-q42	2	1	1	1	1
PTK2B*	8p21.1	12	10	11	10	1
SLC24A4/RIN3*	14q32.12	5	5	5	4	0
SORL1	11q23.2-q24.2	15	14	14	14	1
TREM2	6p21.1	1	1	1	1	0
ZCWPW1*	7q22.1	6	5	4	4	1

Table III-7. Summary of variants identified within implicated linkage regions. Counts are displayed for the number of variants present in the human variation catalogs of dbSNP build 137 (dbSNP), ESP 6500 release (ESP), and 1000 Genomes April 2012 release (1000G). The number of novel variants identified in each implicated linkage region is also shown. Chr = chromosome. Mbp = megabase pair.

Peak	Variants	dbSNP	ESP	1000G	Novel
Chr 2: 62-102 Mbp	282	277	273	276	4
Chr 3: 161-175 Mbp	54	54	54	54	0
Chr 9: 99-114 Mbp	158	157	156	156	1
Chr 18: 7-15 Mbp	63	62	62	62	0

Twenty-one additional genes were implicated by the published GWAS hits. Variants in three of these genes are not in LD with the published SNP, but are closely located to the marker. Variants in 18 of these genes were in high LD with the published SNP, but not reported in the literature. In total, 44 exonic variants were identified in these 21 genes (Table III-8). Four of these variants have previously been studied in this population, one variant in *AGBL2*, two variants in *GPR111*, and one variant in *GPR115*. None of these variants were associated with LOAD in the Amish with a MQLS p-value less than 0.01.

Table III-8. Summary of Additional Genes Implicated by GWAS hits. LD = linkage disequilibrium. High = variants in this gene are in high LD with published GWAS hit. Low = variants in this gene are in low LD with published GWAS hit. novel = not present in dbSNP 137, ESP 6500 release, or 1000 Genomes April 2012 release.

Gene	GWAS hit	LD Type	# variants	# novel
AGBL2	rs10838725	high	4	0
BCDIN3	rs1476679	high	0	0
C1QTNF4	rs10838725	high	0	0
CUGBP1	rs10838725	high	0	0
FNBP4	rs10838725	high	2	1
GATS	rs1476679	high	3	0
GPR111	rs9349407, rs10948363	high	5	0
GPR115	rs9349407, rs10948363	high	6	1
GPR141	rs2718058	low	2	0
KBTBD4	rs10838725	high	0	0
MGC57359	rs1476679	high	0	0
MTCH2	rs10838725	high	1	0
NDUFS3	rs10838725	high	1	1
NUP160	rs10838725	high	5	1
PILRA	rs1476679	high	3	0
PILRB	rs1476679	high	0	0
PLEKHC1	rs17125944	high	0	0
TRIM35	rs28834970	high	2	0
TSC22D4	rs1476679	high	4	1
TXNDC3	rs2718058	low	0	0
ZYX	rs11767557, rs11771145	low	6	1

Analysis of all variants for association with LOAD

Single variant case-control analysis was performed for all 79,203 sequencing variants to test for association with LOAD (Figure III-3). The most significant p-values were 1.25 x 10^{-6} for position 102,762,544 on chromosome 10 and position 91,503,598 on chromosome 15. Thirteen additional exonic variants had p-values less than 1 x 10^{-4} (Table III-9). None of these reach classical levels of genome-wide significance.

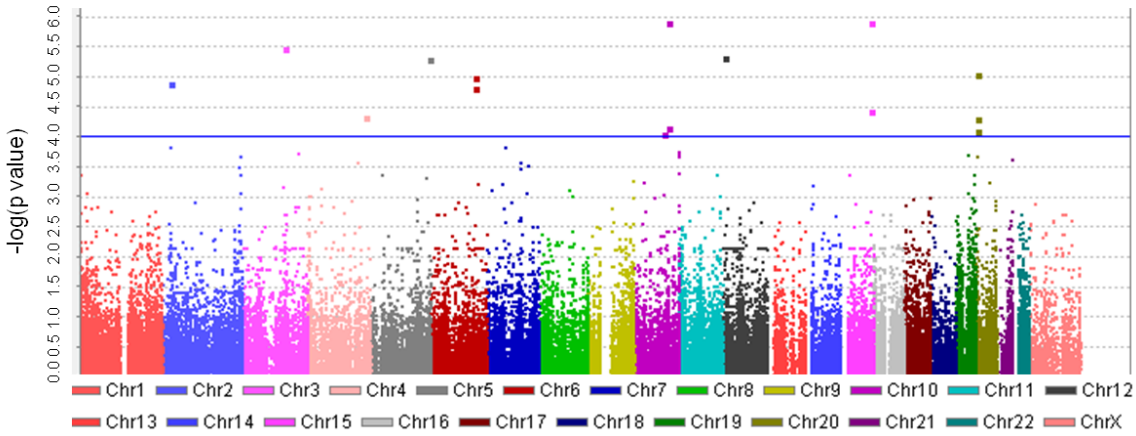

Figure III-3. Manhattan plot for MQLS p-values for 79,203 sequencing variants. -log(p value) = negative log base 10 of the p-value. Chr = chromosome. (—) p-value less than 1 x 10^{-4}. Analyzed variants are plotted on the x-axis by chromosomal position (each color represents a different chromosome). The y-axis is the negative logarithm of the MQLS association p-values. Each variant is plotted as a colored box.

Table III-9. MQLS-corrected allele frequencies and case-control association p-values for the top sequencing variants in the sequencing dataset. Chr = chromosome. MAF = minor allele frequency. Nucleotide position is based upon the UCSC hg19 human reference genome. Gene and Function annotated by SeattleSeq134.

Marker	Chr	Position	Case MAF	Control MAF	Overall MAF	p value	Gene	Function
rs41291476	10	102762544	0.0104	0.0019	0.0013	1.25E-06	*LZTS2*	synonymous
rs147224053	15	91503598	0.0104	0.0019	0.0013	1.25E-06	*RCCD1*	synonymous
rs4548	3	128525253	0.0987	0.0259	0.0525	3.31E-06	*RAB7A*	synonymous
rs11380	12	6601475	0.0156	0.0078	0.0072	4.68E-06	*MRPL51*	missense
rs201285308	5	176008380	0.0156	0.0078	0.0072	5.00E-06	*CDHR2*	missense
rs41279402	20	3785672	0.0414	0.0036	0.0108	9.28E-06	*CDC25B*	UTR-3
6_137234733	6	137234733	0.0403	0.0171	0.0199	1.01E-05	*PEX7*	UTR-3
rs11676272	2	25141538	0.5938	0.4151	0.4619	1.28E-05	*ADCY3*	missense
rs144407106	6	136710582	0.0511	0.0203	0.0277	1.53E-05	*MAP7*	synonymous
rs149872991	15	91496233	0.0278	0.005	0.0085	3.74E-05	*UNC45A*	missense
rs147643564	4	175158508	0.0651	0.0195	0.0318	4.67E-05	*FBXO8*	UTR-3
rs146399677	20	3785297	0.041	0.0076	0.0129	4.89E-05	*CDC25B*	synonymous
rs56400929	10	105762909	0.0104	0	0.001	7.00E-05	*SLK*	missense
rs150358287	20	3687141	0.0894	0.0374	0.0455	7.94E-05	*SIGLEC1*	stop-gained
rs34270879	10	90673047	0	0.0434	0.0352	9.17E-05	*STAMBPL1*	missense

Of the identified variants with a VQSLOD greater than 2, only three uniquely occurred in cases and six in controls. Only one variant unique to controls occurred in more than 10 unrelated individuals, as defined by the kinship coefficients of the individuals carrying the variant.

Of the biallelic exonic variants analyzed, the Amish exomes harbored 5,387 apparently novel variants that had a VQSlod score greater than 2. A minor allele frequency of 0.00309 corresponded to a single allele present in the 162 exomes. Based upon the MQLS-adjusted MAF, 858 of these novel variants were observed a single time or less after adjustment for the relatedness of individuals. Only 13 of these novel variants had a MAF greater than 0.05, the general threshold to be considered a “common” allele.

Prioritization of identified variants for further evaluation

A total of 56 variants (25 in AD genes, 30 in linkage regions, and 1 unique to cognitively normal controls) were identified from the sequencing data and met our criteria for prioritization for genotyping in the full Amish data (Table III-10-Table III-12).

Table III-10. Details of 25 top variants identified from 26 known AD genes for follow-up genotyping. Chr = chromosome. MAF = minor allele frequency. p value = MQLS association test p value. Function, Gene, PolyPhen, GERP, phastCons, Grantham scores are all annotated from SeattleSeq. # databases = number of human variation catalogs the variant is present in (dbSNP 137, ESP 6500 release, 1000 Genomes April 2012 release).

Marker	Chr	Position	Case MAF	Control MAF	p value	Gene	Function	PolyPhen	GERP	phastCons	Grantham	# databases	MAF
1_227079478	1	227079478	0.008	0	0.046	*PSEN2*	synonymous	unknown	-8.95	0.319	NA	0	0.002
7_99998700	7	99998700	0.011	0	0.092	*ZCWPW1*	missense	unknown	-2.24	0	110	0	0.006
rs202188414	7	100001817	0	0.023	0.128	*ZCWPW1*	synonymous	unknown	-1.2	0	NA	1	0.017
8_27297871	8	27297871	0.011	0.019	0.575	*PTK2B*	missense	unknown	5.63	0.962	98	0	0.017
8_27300395	8	27300395	0.018	0.02	0.235	*PTK2B*	synonymous	unknown	3.14	1	NA	1	0.017
11_47505996	11	47505996	0.009	0.005	0.324	*CELF1*	synonymous	unknown	0.92	1	NA	0	0.005
11_59834482	11	59834482	0.031	0.018	0.537	*MS4A3*	missense	unknown	-1.13	0.002	64	0	0.025
rs138180929	11	59861473	0.006	0.007	0.417	*MS4A2*	stop-gained	unknown	4.23	0.997	NA	2	0.006
rs7929057	11	59980598	0.136	0.101	0.205	*MS4A4E*	utr-3	unknown	1.64	0.001	NA	2	0.105
rs147908272	11	60064732	0.012	0	0.332	*MS4A4A*	synonymous	unknown	-7.66	0	NA	2	0.007
11_60064763	11	60064763	0.003	0.005	0.393	*MS4A4A*	missense	benign	-0.63	0	21	1	0.002
rs148346043	11	60152688	0.028	0.007	0.247	*MS4A7*	missense	probably	2.76	0.996	60	2	0.013
rs144076317	11	60165392	0.009	0.015	0.661	*MS4A14*	missense	benign	-7.17	0	109	2	0.010
rs142892172	11	60183953	0.011	0	0.354	*MS4A14*	synonymous	unknown	-1.03	0	NA	2	0.002
11_60197218	11	60197218	0.038	0.016	0.039	*MS4A5*	missense	benign	0.97	0	98	1	0.022
rs200785869	11	60236016	0.017	0.025	0.958	*MS4A1*	utr-3	unknown	-1.57	0	NA	1	0.019
11_121454230	11	121454230	0	0.007	0.376	*SORL1*	missense	unknown	5.91	1	43	0	0.005
14_53348185	14	53348185	0.01	0.005	0.591	*FERMT2*	missense	unknown	5.92	1	56	0	0.007
17_7189779	17	7189779	0.008	0.004	0.686	*SLC2A4*	missense	unknown	5.11	1	64	0	0.004
rs62095193	18	29104689	0	0.005	0.834	*DSG2*	synonymous	unknown	-3.25	0.948	NA	2	0.002
18_29125783	18	29125783	0.012	0	0.720	*DSG2*	missense	unknown	5.99	0.998	56	0	0.005
rs147783767	19	1045209	0	0.005	0.703	*ABCA7*	missense	benign	1.15	0.21	26	2	0.003
19_1054789	19	1054789	0.025	0.012	0.010	*ABCA7*	missense	unknown	2.95	0.008	74	0	0.012
rs4811697	20	55033856	0.441	0.442	0.221	*CASS4*	utr-3	unknown	-6.28	0	NA	2	0.445
rs201970902	21	27484335	0.033	0.021	0.641	*APP*	synonymous	unknown	3.79	1	NA	2	0.021

Table III-11. Details of one variant identified that is unique to controls. Chr = chromosome. MAF = minor allele frequency. p value = MQLS association test p value. Function, Gene, PolyPhen, GERP, phastCons, Grantham scores are all annotated from SeattleSeq. # databases = number of human variation catalogs the variant is present in (dbSNP137, ESP 6500 release, 1000 Genomes April 2012 release).

Marker	Chr	Position	Case MAF	Control MAF	p value	Gene	Function	PolyPhen	GERP	phastCons	Grantham	# databases	MAF
rs16960199	19	54976265	0	0.072	0.045	*CDC42EP5*	utr-3	unknown	-3.91	0	NA	3	0.049

Table III-12. Details of 30 top variants identified from 4 implicated linkage regions for follow-up genotyping. Chr = chromosome. MAF = minor allele frequency. p value = MQLS association test p value. Function, Gene, PolyPhen, GERP, phastCons, Grantham scores are all annotated from SeattleSeq. #databases = number of human variation catalogs the variant is present in (dbSNP137, ESP 6500 release, 1000 Genomes April 2012 release).

Marker	Chr	Position	Case MAF	Control MAF	p value	Gene	Function	polyPhen	GERP	phastCons	Grantham	#databases	MAF
2_73519475	2	73519475	0.010	0.011	0.884	*EGR4*	missense	0.681	4.4	0.991	58	0	0.011
rs7598901	2	73675844	0.326	0.288	0.175	*ALMS1*	synonymous	unknown	-6.6	0	NA	3	0.308
rs1052161	2	73828538	0.373	0.340	0.216	*ALMS1*	missense	benign	-6.29	0.001	26	3	0.359
rs13538	2	73868328	0.206	0.222	0.643	*NAT8*	missense	benign	-7.73	0	155	3	0.220
rs2001490	2	73928098	0.352	0.338	0.417	*NAT8B*	missense	unknown	2.4	0.434	60	3	0.351
2_74709426	2	74709426	0.021	0.019	0.660	*CCDC142*	missense	0.99	2.77	0.029	60	0	0.023
rs2592551	2	85780131	0.384	0.421	0.924	*GGCX*	synonymous	unknown	4.65	1	NA	3	0.422
rs3731828	2	85806266	0.384	0.401	0.963	*VAMP8*	synonymous	unknown	2.93	1	NA	3	0.411
rs2276626	2	86259443	0.252	0.296	0.410	*POLR1A*	synonymous	unknown	-0.61	0.926	NA	3	0.274
rs8244	2	86371883	0.380	0.427	0.242	*IMMT*	synonymous	unknown	-10.5	0.002	NA	3	0.409
rs1050301	2	86400824	0.252	0.303	0.356	*IMMT*	missense	possibly	3.43	1	74	3	0.279
rs61748137	2	88383970	0.095	0.117	0.549	*SMYD1*	synonymous	unknown	-7.98	0.597	NA	3	0.112
rs11889464	2	95537501	0.08	0.055	0.992	*TEKT4*	synonymous	unknown	1.03	0.959	NA	3	0.078
2_95537526	2	95537526	0	0.007	0.666	*TEKT4*	missense	1	1.97	0.89	56	0	0.004
2_96781817	2	96781817	0.060	0.049	0.824	*ADRA2B*	synonymous	unknown	1.13	1	NA	0	0.057
rs1624844	2	97613616	0.338	0.266	0.019	*FAM178B*	synonymous	unknown	-4.79	0.849	NA	2	0.271
rs41280595	2	101580575	0.146	0.171	0.631	*NPAS2*	synonymous	unknown	0.91	0.649	NA	3	0.160
rs3772173	3	170078232	0.137	0.074	0.752	*SKIL*	missense	benign	4.42	1	64	3	0.113
rs5400	3	170732300	0.285	0.154	0.0002	*SLC2A2*	missense	benign	6.08	0.981	89	3	0.199
rs2787374	9	103054951	0.358	0.370	0.541	*INVS*	synonymous	unknown	4.57	0.998	NA	3	0.384
rs10761054	9	107379895	0.334	0.328	0.297	*OR13C9*	missense	possibly	-1.27	0.988	22	3	0.349
9_113018783	9	113018783	0	0.005	0.622	*TXN*	utr-5	unknown	1.37	0.005	NA	0	0.004
rs2281937	9	113169126	0.370	0.384	0.936	*SVEP1*	synonymous	unknown	-7.04	0.798	NA	3	0.371
rs35142681	9	113449489	0.080	0.032	0.009	*MUSK*	missense	benign	-7.08	0.006	81	3	0.043
rs73938538	18	7008583	0.126	0.090	0.006	*LAMA1*	synonymous	unknown	-7.48	0.664	NA	3	0.087
rs906807	18	9117867	0.198	0.243	0.903	*NDUFV2*	missense	benign	4.67	0.974	64	3	0.238
rs6505776	18	12984144	0.247	0.386	0.007	*SEH1L*	missense	benign	1	0.929	65	3	0.344
rs474337	18	13095609	0.247	0.377	0.009	*CEP192*	missense	benign	3.43	0.039	98	3	0.338
rs1786263	18	13116432	0.247	0.377	0.009	*CEP192*	missense	benign	4.76	1	102	3	0.338
rs12457503	18	14752957	0.350	0.470	0.004	*ANKRD30B*	synonymous	unknown	-3.21	0	NA	3	0.441

Discussion

In total, over 79,000 exonic variants were identified from the whole-exome sequence data. The Amish population harbored 605 previously uninterrogated exonic variants in three classes of the genes that are the most likely to contribute to risk of developing LOAD. These variants were identified by sequencing a selected subset of individuals who were the most probable to harbor identifiable risk loci. Given the small dataset, it is not surprising that no variant reached classical genome-wide significance levels. This lack of significance could be due to many reasons.

First, the available power to detect an association in this subset was likely very limited. Power is dependent on a number of variables, including sample size, allele frequency and effect size. The small sample size (162 exomes passing QC measures) is likely to be too small even to detect an association for a common allele with a moderate effect size. For example, if 162 unrelated cases and an equal number of controls were sequenced or genotyped for a variant with a minor allele frequency of 5% and an OR of 2, the power to detect an association is only 34.7% if the type I error rate is 0.05. This estimate assumes individuals are unrelated and therefore is an overestimate of the power in this population of related individuals. Sequencing methods allow for the detection of all variants present in a genome or exome, including those with small allele frequencies. The very low frequencies of some of the identified variants (less than 1%) also contributed to limited power.

Second, this screen only looked at exonic variants, or mutations within the coding region of the genome. The full exome screen, as a secondary analysis to the candidate regions,

only comprises 1% of the complete human genome. While functional variants are likely to contribute to disease risk and cause pathology, recent studies suggest the definition of “functional” needs to change and be expanded. The ENCODE project identified regions of the genome, outside of the exome, that function in transcriptional regulation, transcription factor binding sites, chromatin patterning, transcriptional promotion, epigenetic regulation of RNA processing, non-coding RNA, DNA methylation, transcription enhancement, and DNA structural interactions (Dunham et al., 2012). This screen looked at single base substitutions but sequence data has the ability to detect insertions and deletions. These classes of variation have been associated with disease risk for a variety of phenotypes and were not interrogated by this study. Therefore, it is likely that additional variants and mutations in regions outside of known AD genes, the implicated linkage peaks, and the exome, will confer susceptibility to LOAD and should also be interrogated in future studies.

As previously stated, the top candidate variants were selected from this initial screen for follow up genotyping in the larger, more complete Amish dataset. The variants selected from the known AD genes and the linkage regions were the most significantly associated with LOAD or were the most novel, as defined by their presence in catalogs of human variation. By genotyping these variants in over 1,100 samples, the power limitations of this screening population may be overcome and associations may be detected.

CHAPTER IV
VERIFICATION OF SELECTED VARIANTS AND EVALUATION OF THE SAMPLED AMISH POPULATION

Introduction

The sequencing experiment detailed in Chapter III has a number of associated problems and limitations that can be overcome by focused analysis in a larger study population. Variants identified from sequencing data can result from a true mutation in the DNA sequence, sample misidentification, sample contamination, an error made by the sequencing machine, poor base calling, misalignment, low depth or coverage, low quality scores, or other errors. Additionally, to detect an association with a variant of low effect size or of low frequency, large sample sizes are needed to have sufficient power.

In this follow-up phase of the overall study, a second genotyping technology was used on the full Amish dataset, including the subset of the samples used in the sequencing phase. This allows for confirmatory genotyping and concordance checking. If in the full dataset, a sequencing variant is monomorphic, (i.e. only one allele is present) the variant will fail to validate and is not a true variant. If large discrepancies are found for genotypes when comparing the sequence data to the follow-up genotyping, this would signal low confidence in the sequence data. Additionally, more accurate estimates of allele frequencies are possible as there is a larger sample of the full Amish population used. By genotyping a sequencing variant in the full Amish dataset, the sample sizes are increased and there is more power to detect an association.

The same factors that make the Amish advantageous for genetic studies (see Chapter I) also limit the ability to generalize findings for specific variants. The variants identified in the Amish dataset are likely to be present in different frequency in the general population and therefore may contribute a different level of risk to developing LOAD in this larger population. While the variants may not generalize, the genes and underlying disease processes implicated by them are likely to generalize in this dataset. Therefore, it is important to test whether an association detected in an isolated population, such as the Amish, generalizes and is also associated in a dataset derived from the general population to determine if the specific variant is important for disease risk or if additional variants in the gene or pathway may be identified. For the above reasons, it is necessary to verify the top 56 candidate variants identified from the sequence data and evaluate association with LOAD in the full Amish dataset of over 1,100 individuals. It is hypothesized that these top candidate variants will be associated with LOAD risk in the full Amish dataset.

Methods

Full Amish study population

The full dataset for which samples have been collected has been detailed in Chapter II. This dataset is comprised of individuals from the Amish communities in Adams, Elkhart and LaGrange Counties in Indiana and Holmes County in Ohio. From public directories and referrals from previously enrolled participants, individuals over the age of 80 were identified and all individuals over the age of 85 (as of the year 2006) have been ascertained. Over 30% of the Amish populations over the age of 65 have been contacted and 87% of these individuals have consented to participate in the study. The

Modified-Mini Mental Status (3MS) exam was used to screen individuals during the initial interviews (Teng & Chui, 1987). Information from these baseline screens and additional cognitive testing were used to generate a consensus diagnosis according to the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS) and the Alzheimer's Disease and Related Disorders Association (ADRDA) criteria (G. McKhann et al., 1984). Methods for ascertainment were reviewed and approved by the individual Institutional Review Boards of the respective institutions. Sample collection, DNA extraction, cognitive testing and affection statuses derived from the consensus diagnoses followed procedures detailed in previous studies conducted in these populations (Cummings et al., 2012).

Genotyping and verification of selected variants in the full set of Amish samples

Fifty-four of the prioritized variants described in Chapter III were genotyped in the full dataset using three Sequenom MassARRAY pools and two were genotyped via TaqMan assays. Thirty GWAS hits, including the 21 detailed in Chapter II used for the genetic risk score analysis, were genotyped in these pools in addition to the sequence variants.

Sequenom is a single base primer extension assay that genotypes variants in pools. An oligonucleotide primer anneals immediately upstream to the variant site being tested and then is extended using mass-modified dideoxynucleotide terminators. The products of this extension are detected via matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry. The distinct mass of each product confers the allele specificity of the assay. Cluster plots are generated for genotype calling when the intensity of the high mass product is plotted versus the intensity of the low mass product.

The 84 sequencing variants and GWAS hits, plus a variant identified from an age-related macular degeneration (AMD) study in this same population, were designed into three pools; the first had 31 variants and SNPs, the second 28, and the third 26 (Table IV-1).

TaqMan is a 5' nuclease assay used primarily for biallelic SNPs. The assay is a polymerase chain reaction (PCR) that contains two probes attached to a fluorescent reporter. Each probe binds to a single possible allele and is displaced by Taq DNA polymerase. Then, the polymerase cleaves the reporter which emits a wavelength that can be detected. The genotype calling software plots the fluorescence of one reporter versus the other. This plot generates clusters that correspond to the three possible genotypes of a biallelic marker, similar to the cluster plots from Sequenom. This assay is very accurate (~ 99.7%) and can be easy to run using the pre-designed on demand assays or may be harder if novel variants require assays to be designed (Shi, Myrand, Bleavins, & de la Iglesia, 1999). Fortunately, the two variants to be genotyped via this method had pre-designed assays available (Table IV-2). As the number of markers genotyped increases, this technology becomes expensive and time consuming. A single SNP is genotyped at a time on a 384-well plate and requires 5 ng of genomic DNA for each sample. In today's costs, genotyping four SNPs in 1536 samples via TaqMan is estimated to cost approximately $2,400, while genotyping 50 SNPs via Sequenom is estimated to cost approximately $12,000.

Table IV-1. Primer sequences for three sequenom pools. W1 = pool 1, W2 = pool 2, W3= pool 3.

Pool	Variant	Secondary Primer	Primary Primer	Pool	Variant	Secondary Primer	Primary Primer
W1	rs2787374	ACGTTGGATGGATTCTGCCCAGCATGGAC	ACGTTGGATGAAGGCGCACTCAAGAGCTCA	W2	rs7274581	ACGTTGGATGCTCAGCCTCCCAAAGTGGAA	ACGTTGGATGAGCTTGTCAGACCGGGTAAG
W1	rs16960199	ACGTTGGATGTCGAGCTGAACGACGTCATC	ACGTTGGATGACACACCTTGGCCGTTTATG	W2	19_1054789	ACGTTGGATGCTGACCCTACATCTCCCCT	ACGTTGGATGTCCTCCAGTCCTGCCTCCT
W1	rs561655	ACGTTGGATGTCAAAATTGTATGCTGCCCC	ACGTTGGATGGTATGAAGTTAACCTGGGAG	W2	rs906807	ACGTTGGATGAGGCTATGATTTACTAAAC	ACGTTGGATGAAGCTCCTCCAGCTCCATTT
W1	rs28834970	ACGTTGGATGGCTAAGTGAAAGCAGCCGTC	ACGTTGGATGCTGGTCATTCCATATAAGTG	W2	17_63221207	ACGTTGGATGGTCTCCTTCTAGCCCCTTC	ACGTTGGATGGGTGAGGCGAAAGGCTTCC
W1	rs6656401	ACGTTGGATGGCTTGTAGATGCATCATTTCC	ACGTTGGATGGACAGAAGAGCAAAGGACAC	W2	8_27300395	ACGTTGGATGCTCCTTATCTGACGTGACTC	ACGTTGGATGTCCATGGCAATGTCCTTCTC
W1	8_27297871	ACGTTGGATGGGTGCTGGAGAAAGGAGAC	ACGTTGGATGCGGGTCATGAGGGTATAAAG	W2	18_29125783	ACGTTGGATGGAAACTGAATCGCTGAATGC	ACGTTGGATGGGTCATCTAGCTCTCCTTCA
W1	11_60197218	ACGTTGGATGGGGCTTTGAGTTGAAAAGG	ACGTTGGATGCCGGTGTTTCTGGTATTTCC	W2	rs147908272	ACGTTGGATGTGACCCCCAAATTGTGTACC	ACGTTGGATGAGCCTTAGCATGGGAATAAC
W1	1_227079478	ACGTTGGATGCACCCAGAAGAAGACTCCTA	ACGTTGGATGAGTCAAGGGAGGCTCAAAGA	W2	rs2588969	ACGTTGGATGTCATACTCTAGCAGAAGAGG	ACGTTGGATGATGCCCTTTGCTCTTCAGAC
W1	rs3865444	ACGTTGGATGACAACTGTTTACACCAGGGC	ACGTTGGATGAATCCTATATCCTGCTGGAC	W2	rs4938933	ACGTTGGATGGCAGGACTGGAAAAATCTGA	ACGTTGGATGGGCCAGTACCATTTTTGGAG
W1	rs6733839	ACGTTGGATGTTCCCGTTCCATCCTGTTTC	ACGTTGGATGGGAAAGAAATCTCTGTTCTGC	W2	rs10792832	ACGTTGGATGATCTTGAGGCCACTTAAAGG	ACGTTGGATGGAGATGAAGGCCATCCTTTC
W1	rs13538	ACGTTGGATGGCTCTGCCTGTTGATGATCC	ACGTTGGATGTTTTGCTATCCCCTGACGAC	W2	rs5400	ACGTTGGATGTAATCACCATGCTCTGGTCC	ACGTTGGATGTTCCAAGTGTGTCCCCAAGC
W1	rs35349669	ACGTTGGATGAAGGGACAAGCGCTTCTGGT	ACGTTGGATGTGAAAGTAGGAGCGGAGACT	W2	rs11218343	ACGTTGGATGTTACAGATGTGAGCCACTGC	ACGTTGGATGCACTCAATGTTCCAAGATCC
W1	rs2001490	ACGTTGGATGAGAGCTCCTACTGTGCCCA	ACGTTGGATGCCAAATCCTACCTGAGTGAG	W2	rs6701713	ACGTTGGATGGGAGGTGTTACAGCACACTA	ACGTTGGATGGGATTGACAGAGCTGTTAAG
W1	rs9331896	ACGTTGGATGAGAGGGATAAGAGCTCCGGT	ACGTTGGATGCATTTCATTCAGCTCTTCCC	W2	2_74709426	ACGTTGGATGAGAGCGCGAAGCTGCATCTC	ACGTTGGATGAGACTCTCGAGCCGCTGCTG
W1	rs11767557	ACGTTGGATGATGATGTCTTAGGGCATCTC	ACGTTGGATGCTTGTTGCCTCCATCAACAG	W2	11_59834482	ACGTTGGATGGACAACTCCTTAATGACTGG	ACGTTGGATGTGAACATTGCCAGTGCTAC
W1	rs62095193	ACGTTGGATGCATGTTTTGCAGCTTGAAGG	ACGTTGGATGTTTCATCTGCATCGAACAC	W2	rs6505776	ACGTTGGATGTTTCCCATGAAGCAGCCTAC	ACGTTGGATGCAGGGAACCTCAAATCCTTC
W1	rs10498633	ACGTTGGATGGGCACTTAGCAGACAAGATG	ACGTTGGATGTCCTACCTGATCCACAAAGC	W3	rs73938538	ACGTTGGATGTCTTCTGGGTGCAGCTTTTC	ACGTTGGATGTGATGGCTGCACAGAATCTC
W1	2_96781817	ACGTTGGATGACCAGGACCCCTACTCCGT	ACGTTGGATGCCAGGATGACCAGAGCGTT	W3	rs1624844	ACGTTGGATGATGAGATCCCACAGAAGACC	ACGTTGGATGTTTTCCACCTAGCTGCTGAC
W1	rs201970902	ACGTTGGATGCTTGGCAATACTGCAGGATG	ACGTTGGATGTGAATGTCCAGAATGGGAAG	W3	rs9349407	ACGTTGGATGTGAGTCAGTGAGTGGTGAGC	ACGTTGGATGGTTAGCTTTAGTGTATGGTG
W1	rs7598901	ACGTTGGATGCACTCACATAGAGAGAAGCC	ACGTTGGATGGGAGTGGCTGAAACTTTAG	W3	rs2281937	ACGTTGGATGGGACTATGGCTTCATGAAGG	ACGTTGGATGAGGTGAGTTTTGGAGCACCG
W1	rs8093731	ACGTTGGATGTAAGGCGGGACTCAGTAATC	ACGTTGGATGGGGATGTTAACAGTGGTTTTC	W3	rs3752246	ACGTTGGATGAACCACCCCTTGAACCTCAC	ACGTTGGATGTGCAGGGTAGGACATGCAG
W1	rs3731828	ACGTTGGATGTATGACCCAGAATGTGGAGC	ACGTTGGATGTCACTGTGGCTTCCAGATCC	W3	rs1052161	ACGTTGGATGTAAGTCAACGTCACTGCACC	ACGTTGGATGAGAGAGGCTGGCAGAGACC
W1	rs200785869	ACGTTGGATGATGGTCACTCCATGCAAAGG	ACGTTGGATGAGCTTCCAAGAGACATGCTG	W3	rs1532278	ACGTTGGATGCAAGATCTCACTCCCTGATG	ACGTTGGATGCTGTGTCAGCTGATGCTGAG
W1	rs41280595	ACGTTGGATGCCTCCTGTAATGGTTTTGAC	ACGTTGGATGTGGCAATGAAGCAAACCTCC	W3	11_47505996	ACGTTGGATGGATTTCCAAGAAGTGCACTG	ACGTTGGATGATCCGGCATTCTTCAATCTG
W1	14_53348185	ACGTTGGATGGTACTGCATACTTACTACCTG	ACGTTGGATGGGGTCATTGTTTTACTATAC	W3	17_7189779	ACGTTGGATGGTGAAGATGAAGAAGCCCAG	ACGTTGGATGTATGGGGCCCTACGTCTTC
W1	rs474337	ACGTTGGATGTAGGACAGTCCACGGCTCTT	ACGTTGGATGTACCAGTCAAAGGTCCTCAG	W3	rs1786263	ACGTTGGATGAGTTTATGCCCCAGAGGATG	ACGTTGGATGTTAAGTGTCCGTGATTCCCC
W1	rs4147929	ACGTTGGATGCACCACACTATGTCCCATTC	ACGTTGGATGCACAGTGTGGCGGGGCCAGCA	W3	rs148346043	ACGTTGGATGAATCCAGCAATTTCCACCAC	ACGTTGGATGAGAGTCCCCATTCTACTCAC
W1	rs9271192	ACGTTGGATGGATCAGCAGGGTATCTAAAG	ACGTTGGATGCCCAAGGAGCTCTGATAAAG	W3	rs61748137	ACGTTGGATGACCTGATGTTCTCATTGGGC	ACGTTGGATGAGAAGGATGCTTGGCTGAAC
W1	rs1476679	ACGTTGGATGGTCACATGGTACTTAGACTG	ACGTTGGATGATTCTCCCGATCTGTTCTGG	W3	rs7561528	ACGTTGGATGTTCAGGAAAGAAGACTCTAC	ACGTTGGATGACCATTTAGGCCATAGTTTC
W1	rs2592551	ACGTTGGATGGGTTAAGGTAGCCCAGTTCG	ACGTTGGATGCTATTCCTGGGACATGATGG	W3	rs3772173	ACGTTGGATGCCCCCCAGCGAAAAAAATGA	ACGTTGGATGCCTTCTTAACTGTTGGCACC
W1	rs138180929	ACGTTGGATGTTCCCCAGCTCCACAGATTG	ACGTTGGATGTGATGATGCTGTTTCTCACC	W3	rs11771145	ACGTTGGATGCGGACACCAAAGAATGCATA	ACGTTGGATGAACACCACGGAGTGGATTTG
W2	rs4811697	ACGTTGGATGATGAGGACTGTCTACCTCCC	ACGTTGGATGTTTCCCATAGGGCAGAGTTG	W3	rs12457503	ACGTTGGATGCAACACGGCTCTCCATTATG	ACGTTGGATGCCTTGTTTTGCACCTCGATG
W2	9_113018783	ACGTTGGATGGCTGTAAGGACCGATGGAAA	ACGTTGGATGTAAAGGGAGAGAGCAAGCAG	W3	rs2718058	ACGTTGGATGGAGAACGAGCATTGGGTTTC	ACGTTGGATGACAACATAAATCAACACAG
W2	rs10761054	ACGTTGGATGGCAGTGGCATCAATGTGAAC	ACGTTGGATGAGTTGGCCTGTGCTGACATC	W3	rs8244	ACGTTGGATGCAGAAACACCTACTATCCCG	ACGTTGGATGCAGAACAGTTGGCTTTGATG
W2	2_95537526	ACGTTGGATGAGCGCCTGGGTCTCTGTGG	ACGTTGGATGAGAACTGCTATGCTCGCTAC	W3	rs1050301	ACGTTGGATGTTCCAAGATTCAGTCGGGTC	ACGTTGGATGCCCTTTTGTTTTTGGAGTTG
W2	rs10948363	ACGTTGGATGTAGTGTGTTAGGATTTGAG	ACGTTGGATGACACAACACTTTAGTTCCAC	W3	2_73519475	ACGTTGGATGACCCTTGAAGGCAGAGACAG	ACGTTGGATGTGCCTGTATGAGCCTCAGC
W2	rs202188414	ACGTTGGATGCCAGTCCTCAACCTCCATTT	ACGTTGGATGGTTTCTGGAGCCGATTCAAC	W3	11_121454230	ACGTTGGATGATCCTCATCGGAACCATCCT	ACGTTGGATGAGGCCTCCAACTTCCAGTG
W2	7_99998700	ACGTTGGATGAAGTCCTCGCCATCACTGTT	ACGTTGGATGCTGGACCTGGAGCAACTCAT	W3	rs142892172	ACGTTGGATGGAGGATTTCTGGCCTTTGG	ACGTTGGATGGCAAACCAAAGCCTTGCAAT
W2	rs10838725	ACGTTGGATGTAGCTCTTCTGGAGACTGAG	ACGTTGGATGTTGTCGCCCACGATGGAGTA	W3	rs2276626	ACGTTGGATGTTCTCCTTGCGTTTGGCATC	ACGTTGGATGGCACATTGTGGATGCTGAAG
W2	rs17125944	ACGTTGGATGTACTGGTGCATGATTTTGCC	ACGTTGGATGGTTTGTGGAACAAGCTGGTG	W3	11_60064763	ACGTTGGATGTGACCCCCAAATTGTGTACC	ACGTTGGATGCAATGATGTGTATGGCATCT
W2	rs190982	ACGTTGGATGTACTGAGTCTTATTTTCCCC	ACGTTGGATGGTGTAGTTTTCTATGTGCTC	W3	rs983392	ACGTTGGATGATGGAACATTTGTGAAGTG	ACGTTGGATGTTAGACAACTAAGCTTGTGG
W2	rs144076317	ACGTTGGATGTGAGGACAACAAGGGGAAGT	ACGTTGGATGATCATTGTGGGCTTTGGAAC	W3	rs35142681	ACGTTGGATGTGGAAGACAGTGATGATGGC	ACGTTGGATGGCTCCACAACTCTCCACAG
W2	rs147783767	ACGTTGGATGCTGACCTGTCCCTGATCTTA	ACGTTGGATGCATCAAAATCCGCATGGAC				

Table IV-2. TaqMan assay designs for two variants genotyped via this method. [C/T] = possible alleles for variant of interest within context sequence.

Variant	Strand	Context Sequence
rs11889464	Forward	GCTACCACCAGGCCTTCGCCGACCG[C/T]GACCAGTCGGAGCGGCAGCGGCACG
rs7929057	Reverse	ACTGGTGATATTCTTTCCTAGACTA[C/T]CTCCAACCTAGAAAGAAATGAAAAT

The genotypic information from the Sequenom pools and the two TaqMan assays were combined and the same QC was performed on the complete dataset (Table IV-3). Seven variants (two GWAS hits and five sequencing variants) failed genotyping via the Sequenom pools (Table IV-4). Of the remaining 78 variants and SNPs, two sequencing variants were monomorphic in the larger dataset and thus failed to validate as a true variant. Additionally, two GWAS hits with low efficiency (genotypes called in less than 95% of samples) and one multiallelic sequencing variant were dropped from analyses. The available statistical software for association analysis in a complex population like the Amish (detailed in Chapter III) is restricted to testing biallelic markers therefore any multiallelic markers cannot be analyzed. This resulted in 48 sequencing variants and 25 total GWAS hits (17 were used in Chapter II) passing these QC measures.

Table IV-3. Summary of Variant QC. Sequence variant = variant identified from whole-exome sequence data. GWAS hit = SNP implicated by two recent meta-analyses (Lambert et al., 2013; Naj et al., 2011). AMD variant = variant identified from whole-exome sequence data from subset of individuals chosen for the AMD project. Complete dataset = all variants and markers genotyped in three Sequenom pools and two TaqMan assays.

	Sequence Variant	GWAS hit	AMD variant	Complete Dataset
Selected from sequence data	56	30	1	87
Failed to genotype	5	3	0	8
Failed to validate, monomorphic	2	0	0	2
Dropped due to low marker efficiency	0	2	0	2
Dropped due to multiallelic variant	1	0	0	1
Available for analysis	48	25	1	74

Table IV-4. Five Sequencing Variants that Failed Genotyping in the Verification Phase. Chr = chromosome. MAF = minor allele frequency. MQLS-adjusted MAFs and association p-value values are from the sequence data (see Chapter III).

Marker	Chr	Position	Case MAF	Control MAF	p value	Gene
rs5400	3	170,732,300	0.285	0.154	0.00017	*SLC2A2*
rs35142681	9	113,449,489	0.080	0.032	0.00865	*MUSK*
rs906807	18	9,117,867	0.198	0.243	0.90281	*NDUFV2*
rs474337	18	13,095,609	0.247	0.377	0.00856	*CEP192*
rs12457503	18	14,752,957	0.350	0.470	0.00369	*ANKRD30B*

Five of the 48 sequence variants passing the above QC had an initial concordance rate with the sequence data less than 97.5%. One of these variants, rs144076317, was low in concordance and appeared monomorphic in the larger dataset. Examination of the cluster plots generated by the calling software determined that the genotypes needed to be manually called as the graph showed weak clustering. The automated genotyping software for the Sequenom technology is called Typer ("MassARRAY Typer 3.4 Software User's Guide for iPLEX and hME," 2006). With this software, TyperAnalyzer is used to view and analyze the data collected from the spectra readings. It uses threshold values, 85-100%, that correspond to a conservative to moderate call to categorize the strength of the genotype in each well. The spectrum for each well shows the location of each allele peak. The height of each peak corresponds to the amount detected of each analyte or allele reporter (Figure IV-1). The software converts the spectra into a cluster plot by plotting the height of the peak for allele 1 versus to height of the peak for the alternate allele. The color and shape of the point reflects the genotype call for a particular sample. The cluster plot for a reliable assay will have data points that fall along the axis (homozygotes) and the diagonal (heterozygotes). For unreliable assays, data points will fall between clusters and may be labeled as “no calls” or alternatively, may be called as the wrong genotype if not enough data points exist for a given genotype for a clear cluster to be formed. This latter scenario was the case for rs144076317 as there were too few individuals with the alternate allele for cluster formation. This manually calling increased the concordance to 100% and called four individuals as heterozygous for the minor allele, including three individuals who were a part of the sequence dataset (Figure IV-2).

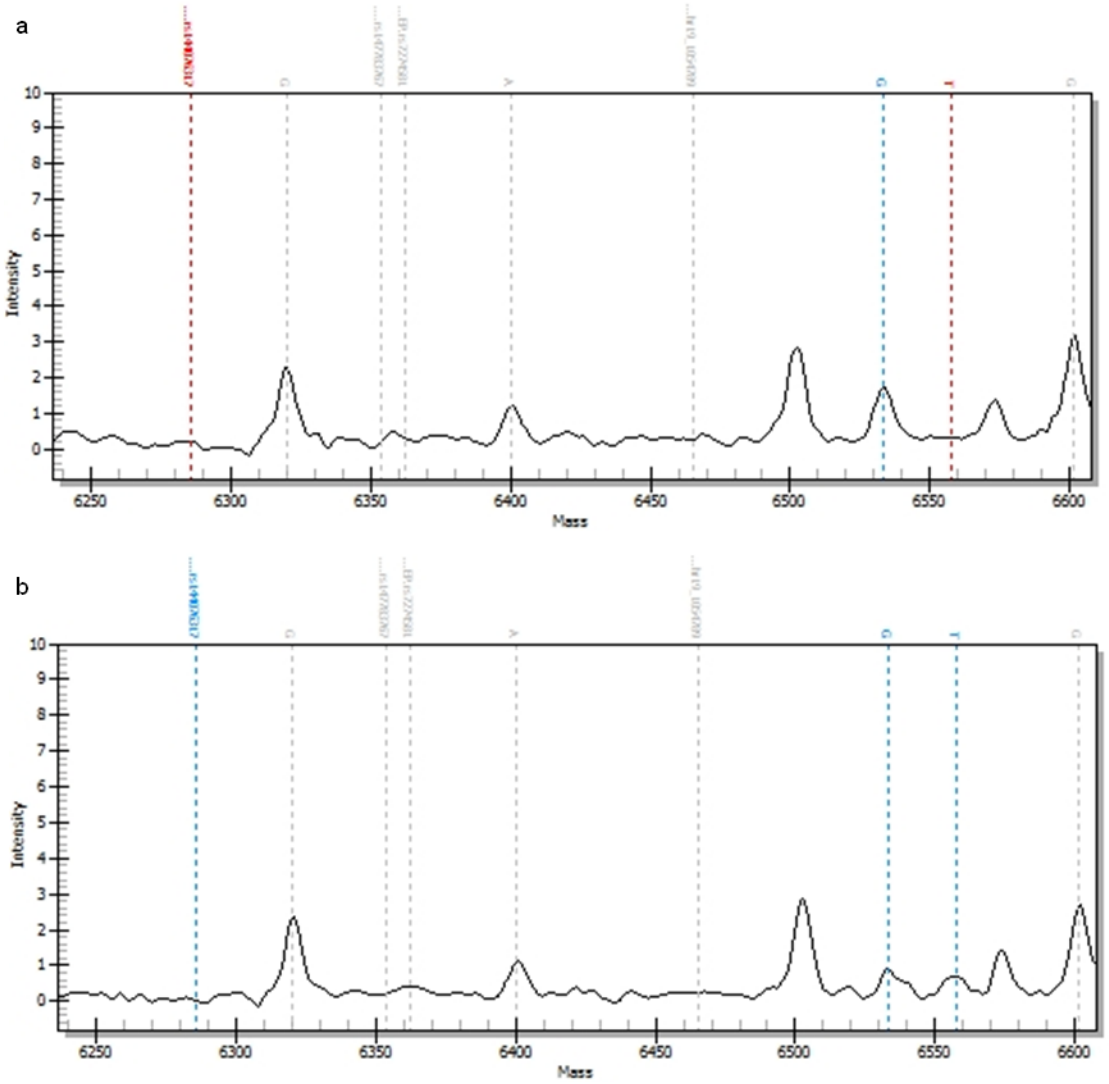

Figure IV-1. Spectrum peaks for rs144076317 from Sequenom MassARRAY Typer Software. (a) Spectrum from sample with a reliable peak for G/G genotype. (b) Spectrum from sample with reliable peak for G/T genotype.

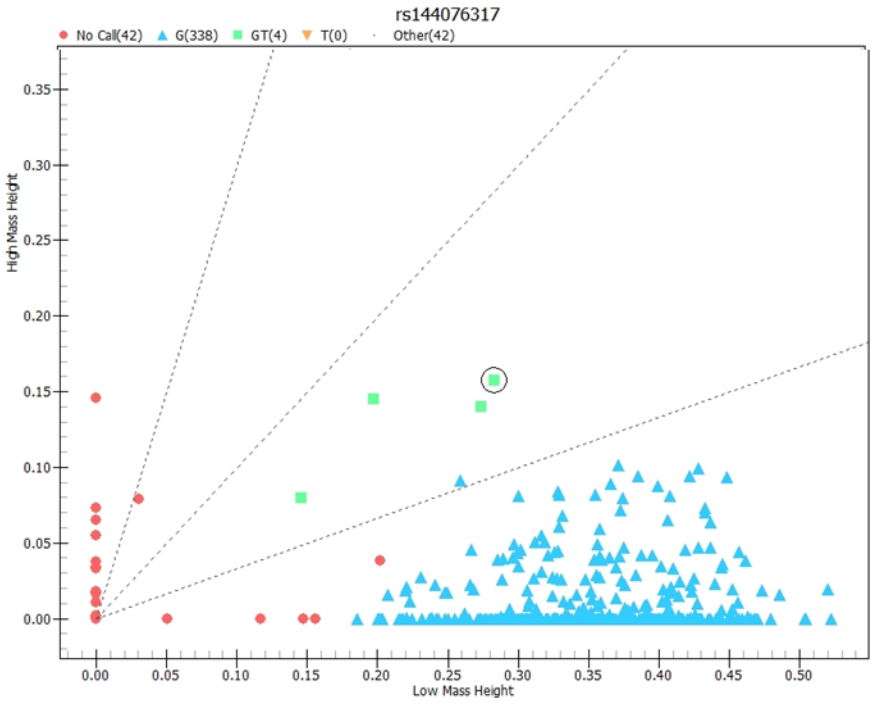

Figure IV-2. Log height plot for rs144076317 generated by Sequenom MassARRAY Typer Software. Log height plots each assay results by probe of extension rates using the reported log (height) of allele peak signals. (▲) or G = homozygotes for major allele G. (■) or GT = heterozygotes who were manually called G/T. (▼) or T = homozygotes for minor allele T. (●) = samples where no call was made for the genotype.

The majority of the discordant genotypes for the remaining four variants consisted of homozygotes for either allele from the sequence data being called as heterozygotes in the follow-up genotyping (Table IV-5). On average 83% of the discordant genotypes were of this manner for these variants. The remaining discordant genotypes consisted of sequence heterozygotes being called as homozygotes for either allele in the follow-up phase. Three of these variants were sequenced at a low depth, less than 10x on average across all individuals, but one was sequenced at an average depth of 100x. Low sequence depth could explain the discordance if the read counts were too small to contain one with the minor allele. For a heterozygous individual, the probability of a read containing the minor allele should be 0.5. Therefore, the probability of all reads containing the same allele is 0.0625 if the sequence depth is only 4. While small, this is

not a trivial probability. These variants with low concordance were flagged but not removed from further analysis. It was determined that if these variants were significantly associated with LOAD, then confirmatory genotyping would be necessary to be confident in the association results.

Table IV-5. Details of the four sequencing variants that were initially less than 97.5% concordant with the sequence data. AA to Aa = number of genotypes that were homozygous in the sequence data that were called as heterozygous in the follow-up genotyping. Aa to AA = number of genotypes that were heterozygous in the sequence data that were called as homozygous in the follow-up genotyping MAF = MQLS-adjusted minor allele frequency.

Marker	Discordant Genotypes	AA to Aa		Aa to AA		Concordance Rate	Sequence Depth	MAF
		Major	Minor	Major	Minor			
rs4811697	23	10	8	3	2	0.86	5.32	0.36
rs2001490	15	0	14	0	0	0.91	100.23	0.29
rs16960199	5	5	0	0	0	0.97	4.32	0.06
rs3731828	5	3	0	2	0	0.97	9.4	0.34

Two variants failed to validate and were monomorphic in the larger dataset. The first variant is a novel missense mutation and is located within the linkage region on chromosome 2 at position 95,537,526. The single heterozygote in the sequence dataset was sequenced at this position with a depth of 13 and 10 of the reads contained the alternate/minor allele. The alternate/minor allele was only observed in a single read in three of the remaining 166 individuals who were sequenced with an average depth of greater than 30. The MQLS p-value for this variant was 0.67 in the sequence data. The Sequenom genotyping software called this individual as homozygous for the referent allele. As the true genotype for this individual at the position cannot be resolved, this variant from removed from further analysis.

The second variant that failed to validate, rs62095193, is located on chromosome 18 at position 29,104,689 and is a synonymous mutation in *DSG2*, a LOAD risk gene (Lambert et al., 2013). A single heterozygote had the alternate/minor allele in both the sequence and the follow-up datasets. This individual was sequenced at this position with a depth of 32 and the alternate allele was present in 14 of the reads. Six of the 166 individuals with an average depth greater than 30 had a single read with the alternate allele. The MQLS p-value in the sequence dataset was 0.83. However, when MQLS adjusted the minor allele frequency for the relatedness of the individuals in the follow-up dataset, this variant was reported as monomorphic. This may be either a true non-validation of the variant or a genotyping error in the Sequenom pools. As only a single heterozygote was genotyped and as MQLS reported the variant as monomorphic, it was removed from further analysis in the context of no resolution on the nature of the variant.

A total of 1,143 samples were genotyped for the variants (Table IV-6). Of these, 24 samples were duplicates. Including duplicate samples allows for concordance checking across samples and plates to control for the quality of the genotyping assays. Two sets of duplicate samples were discordant at nine and 14 variants, respectively. With this level of discordant genotypes, both samples from each set were removed from analysis. For nine of the duplicate pairs that were concordant, the sample with the lower genotyping efficiency was removed. The remaining 11 duplicate pairs were concordant and genotyped at the same efficiency, so a random sample from each pair was chosen to be removed. Of the 1,119 remaining unique samples, 83 were dropped due to a genotyping efficiency below 95%. Two individuals were dropped from analysis for low concordance (68% and 77%, respectively) between follow-up genotyping and the sequencing data. To calculate kinship coefficients to adjust for relatedness, individuals

not currently in the AGDB and those who were not in the subsequent all-connecting pedigree (n = 113) were removed. This resulted in 921 samples passing all QC measures (Table IV-7).

Table IV-6. Summary of Sample QC. AGDB = Anabaptist Genealogy Database.

	Cases	Controls	Unknowns	Complete Dataset
Genotyped samples	144	625	374	1143
Removed due to discordant duplicates	0	0	4	4
Removed due to lower efficiency duplicate	1	3	5	9
Removed random duplicate sample	1	6	4	11
Removed due to genotyping efficiency < 95%	15	40	28	83
Removed due to concordance < 80%	1	0	1	2
Removed because not in AGDB or pedigree	0	73	40	113
Available for analysis	126	503	292	921

Table IV-7. Demographics of Samples Used For Follow-up Genotyping. Age of exam and onset averages and standard deviations were calculated for the 921 samples which passed QC for follow-up genotyping.

Affection status	Female	Total	Average age of exam/onset (standard deviation)
LOAD case	**63%**	**126**	**78 (7.75)**
Cognitively normal	**58%**	**503**	**79 (6.72)**
Unclear or unknown	**49%**	**292**	**80 (6.82)**

Case-control dataset of unrelated individuals

A collaborative study between researchers at the University of Miami and Vanderbilt University has ascertained approximately 1000 European-American individuals unique from the Amish populations (Table IV-8). These individuals are the same as those described in Chapter II. As the Amish are founded from European immigrants, this European-American dataset of unrelated individuals is of similar ancestry. These individuals have been diagnosed with probable or definite AD according to NINCDS-ADRDA criteria with an age of onset greater than 60 (G. McKhann et al., 1984). To make these diagnoses, documentation or a clinical history of significant cognitive impairment

was present. Age- and gender-matched cognitively healthy controls were ascertained from the same regions and had a documented 3MS or MMSE score in the normal range (over 78) ("Canadian study of health and aging: study methods and prevalence of dementia," 1994; Tombaugh, 2005). The cases and controls are demographically similar with an average age of onset or age at exam of 74 and female percentage of 63% and 60%, respectively (Naj et al., 2011).

Table IV-8. Demographics of Samples from the Unrelated Datasett. Age of exam and onset averages and standard deviations were calculated for the 971 samples which passed QC for follow-up genotyping.

Affection Status	Female	Total	Average age of exam/onset (standard deviation)
LOAD case	63%	473	74 (8)
Cognitively normal control	60%	498	74 (8)

Case-control analysis

Case-control association in the Amish was performed using the Modified Quasi-Likelihood Score (MQLS) test, which corrects for the relatedness of individuals (Thornton & McPeek, 2007). This association software was described in detail in Chapter III. Type 1 error rates for the method are not inflated when used for the Amish (Cummings et al., 2013). A conservative Bonferroni correction for the number of tests performed (48 sequencing variants passing QC) was used to determine the threshold for the level of significance. Previous studies in this Amish population investigated this software's power to detect associations (Cummings et al., 2013). For dominant and additive models, there was greater than 90% power to detect an association at $p < 0.05$ when the simulated odds ratio (OR) was at least 2 and the minor allele frequency was held constant at 0.2. For genome-wide data, the Bonferroni-corrected p-value is traditionally 5×10^{-8}. If the OR is 5, there was < 90% power to detect an association for dominant and additive

models, but this power dropped significantly, less than 5%, if the OR was less than or equal to 2. In the analysis of the unrelated dataset, logistic regression was performed in PLINK (version 1.07) with *APOE* as a covariate.

Age of onset analysis

So far, all the analyses described have focused on variants that contribute risk to disease status. As an alternative hypothesis, these variants may be acting as a modifier of LOAD. One of the most commonly examined measures for LOAD is age at onset. Age of onset was recorded for 105 of the 127 cases that passed genotyping QC measures detailed above. Previous studies in these populations demonstrate the expected relationship between *APOE* genotype and age of onset (Cummings et al., 2012). An association score test (mmscore in GenABEL R package, version 1.7-6), that adjusts for the relatedness of the case samples using kinship coefficients, was used to determine if there was a relationship between age of onset and genotype in this Amish dataset. *APOE* genotype was used as a covariate in this analysis. To replicate any significant results, a similar analysis was performed in the outbred case-control dataset using the same information for 464 cases in a linear regression with APOE status as a covariate.

Results

Similarly to the analysis of the whole-exome sequence data described in Chapter III, each of the variants that passed QC during the follow-up phase was tested to see if an allele was associated with LOAD in this population. No variant passed the significance threshold when corrected for 48 tests ($p < 0.001$). The most significant result ($p = 0.0012$) was for rs73938538 (MAF 0.087), a synonymous variant in *LAMA1* within the

linkage peak on chromosome 18. No other variant was significant at a threshold of p < 0.05. Seven of the 48 markers had a p-value less than 0.1 (Table IV-9).

Table IV-9. MQLS-corrected allele frequencies and case-control association p-values for the variants in the full dataset. Chr = chromosome. MAF = minor allele frequency. (*) Variant in implicated linkage regions. (+) Variant in implicated AD gene. (#) Variant unique to controls.

Marker	Chr	Pos	Case MAF	Control MAF	p value	Gene
rs73938538*	18	7008583	0.8487	0.9234	0.0012	*LAMA1*
11_47505996^{+}	11	47505996	0.0069	0.0001	0.0543	*CELF1*
rs1786263*	18	13116432	0.3063	0.3384	0.0550	*CEP192*
rs6505776*	18	12984144	0.3162	0.3423	0.0758	*SEH1L*
rs8244*	2	86371883	0.4026	0.4523	0.0775	*IMMT*
rs3772173*	3	170078232	0.1287	0.1635	0.0788	*SKIL*
rs4811697^{+}	20	55033856	0.3946	0.37	0.0920	*CASS4*
2_74709426*	2	74709426	0.0169	0.0005	0.1167	*CCDC142*
rs1624844*	2	97613616	0.2656	0.2391	0.1223	*FAM178B*
2_73519475*	2	73519475	0	0.0117	0.1243	*EGR4*
2_96781817*	2	96781817	0.0533	0.0236	0.1272	*ADRA2B*
7_99998700^{+}	7	99998700	0.0131	0.0069	0.1811	*ZCWPW1*
18_29125783^{+}	18	29125783	0.0042	0.0002	0.1912	*DSG2*
rs2276626*	2	86259443	0.2768	0.3375	0.2015	*POLR1A*
rs142892172^{+}	11	60183953	0.0107	0.0048	0.2104	*MS4A14*
11_60064763^{+}	11	60064763	0.0053	-0.0005	0.2131	*MS4A4A*
11_60197218^{+}	11	60197218	0.0171	0.0103	0.2890	*MS4A5*
rs201970902^{+}	21	27484335	0.0287	0.0146	0.3253	*APP*
rs16960199$^{\#}$	19	54976265	0.037	0.0583	0.3613	*CDC42EP5*
rs138180929^{+}	11	59861473	0.0073	0.0201	0.3631	*MS4A2*
rs7929057^{+}	11	59980598	0.1332	0.0738	0.3681	*MS4A4E*
11_59834482^{+}	11	59834482	0.0308	0.0199	0.4125	*MS4A3*
rs13538*	2	73868328	0.1806	0.2483	0.4776	*NAT8*
rs61748137*	2	88383970	0.1409	0.1319	0.5190	*SMYD1*
rs41280595*	2	101580575	0.0924	0.1094	0.5396	*NPAS2*
11_121454230^{+}	11	121454230	0	0.0012	0.5526	*SORL1*
rs147783767^{+}	19	1045209	0	0.002	0.6346	*ABCA7*
rs10761054*	9	107379895	0.3296	0.2884	0.6480	*OR13C9*
8_27300395^{+}	8	27300395	0	0.0009	0.6483	*PTK2B*
rs3731828*	2	85806266	0.3303	0.3579	0.6711	*VAMP8*
rs2592551*	2	85780131	0.3161	0.3549	0.6728	*GGCX*
rs11889464*	2	95537501	0.0785	0.0528	0.6920	*TEKT4*
rs202188414^{+}	7	100001817	0.0183	0.0124	0.6977	*ZCWPW1*
rs200785869^{+}	11	60236016	0.0222	0.0124	0.7086	*MS4A1*
rs2001490*	2	73928098	0.7192	0.6956	0.7121	*NAT8B*
17_7189779^{+}	17	7189779	0.0081	0.0045	0.7210	*SLC2A4*
19_1054789^{+}	19	1054789	0.0034	0.0111	0.7230	*ABCA7*
rs2281937*	9	113169126	0.3966	0.3849	0.7321	*SVEP1*
rs144076317^{+}	11	60165392	0.0038	0.0019	0.7365	*MS4A14*
rs148346043^{+}	11	60152688	0.0092	0.0178	0.7424	*MS4A7*
1_227079478^{+}	1	227079478	0.0019	0.0071	0.7619	*PSEN2*
9_113018783*	9	113018783	0.0082	0.0024	0.7637	*TXN*
rs2787374*	9	103054951	0.3913	0.3852	0.8000	*INVS*
14_53348185^{+}	14	53348185	0	0	0.8205	*FERMT2*
8_27297871^{+}	8	27297871	0.0174	0.0126	0.9210	*PTK2B*
rs147908272^{+}	11	60064732	0.0066	0.0097	0.9265	*MS4A4A*
rs7598901*	2	73675844	0.3029	0.3525	0.9600	*ALMS1*
rs1052161*	2	73828538	0.6766	0.6277	0.9657	*ALMS1*
rs62095193^{+}	18	29104689	NA	NA	NA	*DSG2*

To determine if the rs73938538 association replicated in a dataset of unrelated cases and controls, the variant was genotyped in 473 LOAD affected individuals and 498 cognitively normal controls. When the variant was tested for association with LOAD, it failed to replicate (logistic regression with *APOE* as a covariate, p = 0.28). In this unrelated dataset, the minor allele frequency (MAF) in cases was 0.081 and 0.094 in controls, which is the opposite direction of effect of the minor allele in the Amish. There is at least 90% probability to detect an association, if present, when the effect size is at least 0.125 with a type I error probability of 0.05. The association program used in the Amish, MQLS, does not return an OR or effect size for the variant being tested so this power calculation is an estimate that may vary based on the true effect size.

Additionally, the score test for association in the Amish between age of onset and genotype for rs73938538 was not significant (corrected p-value = 0.60, df=1). As this result was not significant, it was not tested for in the unrelated dataset.

Discussion

A synonymous variant in *LAMA1*, rs73938538, is associated with LOAD in the Amish just below experiment-wide significance (Figure IV-3). *LAMA1* encodes the laminin alpha subunit. Laminins are a major functional component of the basement membranes of many tissues. This protein is involved in pathways for axon guidance, extracellular matrix interactions, and cell adhesion and migration. Laminin is found underlying the endothelium of blood vessel walls and different isoforms may contribute to vascular homeostasis (Yousif, Di Russo, & Sorokin, 2013). The alpha1 subunit of laminin is

expressed in the basal lamina of blood vessels in the central nervous system, mostly confined to capillary walls (Virtanen et al., 2000). There is strong evidence to suggest the etiology of LOAD may include cerebrovascular dysregulation and that the neuronal degeneration is secondary to this dysregulation (Bomboi et al., 2010; Cullen, Kocsi, & Stone, 2006). Amyloid is deposited in arteries leading to leakage and hemorrhage. Selective breakdown of the blood-brain barrier may compromise the effectiveness of amyloid removal. However, there are conflicting reports on the cause and effect of this degeneration and dysregulation and which occurs first. The association of the synonymous variant rs73938538 with LOAD in the Amish suggests that the cerebrovascular homeostasis and dysregulation may contribute to the underlying pathology and degeneration in this isolated population.

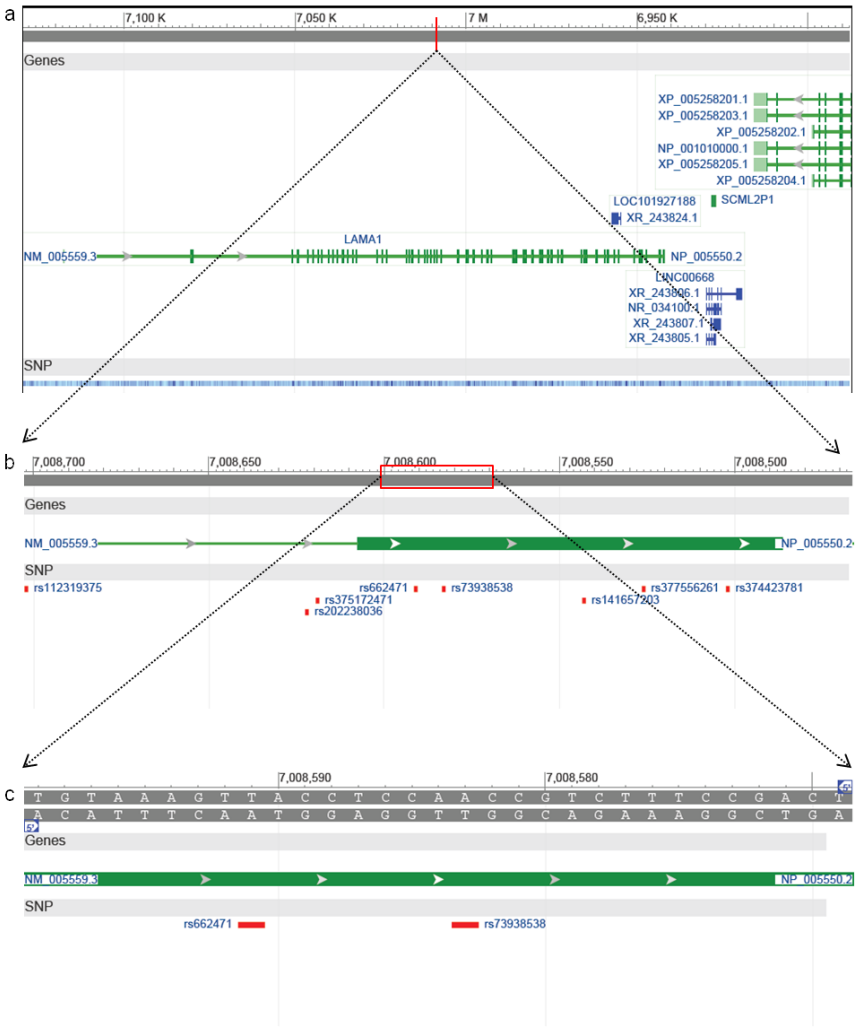

Figure IV-3. *LAMA1* Gene Position and Structure. Derived from the National Center for Biotechnology Information (NCBI) genomic region browser (http://www.ncbi.nlm.nih.gov/gene). (a) Full structure of *LAMA1* and surrounding genes. k = kilobase. m = megabase. (b) Exon level structure of *LAMA1* gene surrounding *rs73938538*. (c) Sequence level structure of *LAMA1* gene around *rs73938538*.

While synonymous, or “silent”, mutations do not change the amino acid encoded by the DNA sequence, these changes may cause human disease by altering protein expression, conformation and function. Messenger RNA (mRNA) nucleotides encode enhancers for splicing and mutations may change the efficiency if they occur at exon-intron boundaries (Pagani, Raponi, & Baralle, 2005). If the synonymous mutation occurs in the binding site of a microRNA, this change may alter the degradation patterns of the

transcript and therefore alter protein expression levels (Friedman, Farh, Burge, & Bartel, 2009). However, as this variant does not occur at such a boundary or binding site, it is unlikely to be contributing through these mechanisms.

A synonymous mutation can affect the speed of translation at many points. If the mutation encodes for a less abundant transfer RNA (tRNA), this can slow down the translation machinery (Zhang, Hubalewska, & Ignatova, 2009). Additionally, evidence suggests rare and common codons are distributed across mRNA to create control points for the speed of elongation. If a synonymous mutation occurs in one of these control points or encodes for a rarer tRNA, it can alter the speed of translation and therefore co-translational folding of specific protein secondary structure and cause misfolded proteins (Powers & Balch, 2008; Tsai et al., 2008; Tuller et al., 2010). This synonymous variant encodes for a more common valine codon (GTG) which has a frequency of 2.91 in highly expressed human genes and 2.78 in all human genes than the referent allele does (GTT) which has frequencies of 1.12 and 1.11, respectively (Karlin & Mrazek, 1996; Lavner & Kotlar, 2005). Therefore, the associated synonymous variant may increase disease risk through inefficient translation or abnormal co-translational folding of a protein important for the function of the basement membrane of the cerebrovasculature by substituting for a more common codon.

The lack of generalization in the unrelated dataset may be due to several reasons. First, the association detected in the Amish population may be a false positive and therefore not a true association. If this is true, the association should not be detected in any other study population or dataset. However, there is at least 90% probability to detect an association, if present, when the effect size is at least 0.125 with a type I error probability

of 0.05. Second, the association may be easier to detect in a dataset from a more homogeneous background. Third, the phenotype-genotype correlation may have arisen separately in the Amish after the founding of the population and could therefore be unique to this genetically isolated population. Fourth, the association with this variant and LOAD may be the result of an interaction with another genetic variation or a component of the environment that is unique to the Amish culture or way of life. If this interacting factor was untested or unaccounted for in this study, and therefore not reproduced in the unrelated dataset, the association may not be detected.

Five sequencing variants failed genotyping in the Sequenom pools. These variants can be genotyped via alternative methods (TaqMan genotyping, direct Sanger sequencing, etc.) to overcome this limitation. These variants may be associated with LOAD in the Amish, but were not interrogated because of the failure to genotype. If these variants are associated in the Amish and that association generalizes in the unrelated dataset, new knowledge about the disease process could be learned.

These results suggest that exonic variation in associated LOAD genes and regions implicated by previous linkage studies do not contribute risk to LOAD in the Amish (within the limits of the power to detect effects), beyond the possible association with *LAMA1*. Other areas of the genome, intronic regulatory elements, epigenetic modifications, or previously unassociated genes, may be harboring variation that confers susceptibility in the Amish but were not interrogated by this study. Additional studies examining the non-exonic variation of known AD genes and the candidate linkage

regions, as well as other portions of the genome, are likely to identify new variation that confers susceptibility to developing LOAD in the Amish.

CHAPTER V
CONCLUSION

Summary

Alzheimer disease is the most common cause of dementia and occurs in over 30% of the US population over the age of 85 years. The pathophysiology underlying the disease is not yet fully understood, and many theories have been proposed to explain observed findings from research (see Chapter I). However, this work has not yet translated into effective treatment options capable of significantly slowing disease course or targeting the underlying biological pathogenesis. Additional research has focused on identifying drug targets and predictive biomarkers by studying the many environmental, health and genetic risk factors associated with the late-onset form of this disease. However, the currently known genetic risk factors do not explain the expected risk based on heritability estimates.

The main goal of this project was to help overcome the genetic and environmental heterogeneity present in most genetic studies of LOAD by studying the isolated founder population of the Amish communities in Ohio and Indiana. To better understand how the risk loci previously identified in the general population contribute to disease risk in this special population, total genetic burden was calculated for Amish cases and controls and compared to unrelated individuals from the general population. This study also built on the previous dementia work conducted in the Amish by whole-exome sequencing a selected subset of the overall study population. These data were used as a screening tool to identify variants harbored in the genomic regions that are most likely to contribute to disease risk. By then genotyping the top candidate variants from this initial screen in

the full dataset, statistical power to detect an association between the variant and phenotype of interest was increased.

The Amish cases tended to have a lower genetic risk score than the unrelated cases from the general population. This result suggests that the common variants implicated by GWAS explain a smaller proportion of genetic risk in the Amish than in the general population. This is consistent with the lack of significant associations observed for these risk loci in previous studies in the Amish. However, since Amish cases did have a higher burden when compared to cognitively normal controls from the same population, it can be assumed these known risk loci do explain some of the expected genetic effects. The lack of correlation between total genetic risk and the parent population for cognitively normal controls suggests that the Amish controls are genetically similar to controls from the general population in the lack of previously reported risk loci. It is likely that additional variation outside of the currently reported risk loci confers susceptibility to LOAD in the Amish population. Exonic mutations are a likely source of this risk and are the easiest and most feasible to interrogate.

In total, over 79,000 exonic variants were identified from the whole-exome sequence data. The Amish population harbored 605 previously uninterrogated exonic variants in three classes of genes that are considered the most likely to contribute to risk of developing LOAD. These variants were identified by sequencing a selected subset of individuals who were the most probable to harbor identifiable risk loci. Given the small dataset, it is not surprising that no variant reached classical genome-wide significance levels. This lack of significance could be due to many reasons, including limited power or

the regions of the genome screened were too narrow. The top candidate variants therefore were selected from genes known to carry early-onset mutations, genes previously implicated through GWAS, and genes located in four implicated candidate linkage regions. These variants were then genotyped and analyzed in the larger, more complete Amish dataset.

Of these top candidates, a synonymous variant in *LAMA1*, rs73938538, is associated with LOAD in the Amish just below experiment-wide significance. Laminins are a major functional component of the basement membranes of many tissues. This gene encodes the laminin alpha subunit, which is expressed in the basal lamina of blood vessels in the central nervous system. The association of the synonymous variant rs73938538 with LOAD in the Amish suggests that cerebrovascular homeostasis and dysregulation may contribute to the underlying pathology and neurodegeneration in this isolated population. While the presence of this variant does not change the amino acid sequence of *LAMA1*, it does encode a more common codon than the referent allele does, and may increase disease risk through inefficient translation or abnormal co-translational folding of the resulting protein.

Overall, these results indicate that exonic variation in a majority of previously associated LOAD genes, and regions implicated by previous linkage studies, does not contribute to risk for LOAD in the Amish dataset studied. However, a potential relationship between a variant in the *LAMA1* gene and risk for LOAD was identified in this special population. The predicted function of this gene is also relevant to LOAD pathophysiology, suggesting it as a strong candidate gene for follow-up studies.

Future Directions

The studies described in this thesis work have generated many additional questions and potential future studies that should be conducted to address the new knowledge gaps. To further elucidate the functional relevance and consequences of the association with rs73938538 and LOAD risk, future studies should be conducted to confirm that *LAMA1* is expressed in the relevant cerebrovasculature endothelium. If this expression pattern reflects the pattern of neuronal loss and cortical atrophy, this evidence would reinforce the contribution to disease risk for this variant. Then similar differential expression studies could be performed to test the effects of the associated variant. Negative results from these functional and expression studies would suggest that the observed association was a false positive or a lack of positive results could suggest that a confounding factor may have been present in the association testing that was not available or measured in these follow-up studies.

Functional studies investigating the translational effects of rs73938538 should be conducted in cell types expressing the mutated gene and protein product. The most relevant cell type for this study would be endothelial cells from cerebrovasculature tissue, but an initial study could be performed using any cell type that expresses the gene product to determine if the variant does affect translation. If this variant does alter either speed of translation or co-translational folding, differences could be detected by comparing protein lysate from cells with the variant allele to those with the referent allele, at the same stage of growth, via Western blotting. Significant decreases in “mutant” protein expression would provide additional support that the association detected in the

Amish affects translation efficiency and is not a false positive. To assess the actual rate of translation, one approach would be to generate a construct containing the variant capped and polyadenylated *LAMA1* mRNA fused to a luciferase reporter and a construct containing the normal mRNA. Translation could then be monitored and compared by following the accumulation of luciferase activity over time (Zeenko et al., 2008). To determine if the variant alters co-translation folding, a method of SDS-PAGE analysis of nascent chains accumulating during *LAMA1* translation could be used to monitor translation kinetics (Komar, Lesnik, & Reiss, 1999).

To fully understand the differences in genetic architecture suggested by the risk score analysis, the reason for the large discrepancy between the regression p-value and the p-value from GEE should be further investigated. If a null correlation matrix is inputted to GEE, the results should be similar to those from the normal regression. However, if the p-values resulting from a null correlation structure are still dissimilar, then another method may be needed to determine if GEE is over-correcting for the relatedness or if the method is properly testing the stated hypothesis.

The same sequence screening and follow-up genotyping datasets could be analyzed in different ways to provide alternate results and interpretations. The studies detailed in the preceding chapters were conducted on a variant or gene level. By expanding the scope of the screening to included pathways implicated by the three classes of genes (those implicated by GWAS hits, known to carry early-onset mutations, and located in four previously identified candidate linkage regions), additional variation may be identified that confers disease susceptibility.

These studies focused on single base substitutions, but sequence data can be used to identify small insertions or deletions (indels). Other types of larger structural variation, larger indels or translocations, may have been captured if the borders incorporated or crossed the intervals targeted by the enrichment technology.

One-hundred thirteen individuals were removed from the analysis in the larger Amish dataset because they did not have an AGDB identification (ID) number, and therefore could not be related to other individuals, or because the AGDB was unable to connect them into the pedigree. This represents a significant number of samples, many of which are the most recently ascertained individuals. For individuals without an ID, the appropriate information should be gathered and the AGDB should be queried to determine if these individuals do have an ID and if they can be connected into a new, potentially larger pedigree. If genome-wide genotyping has been previously performed on an individual for whom the AGDB cannot find an ID, this genetic information could be used to estimate the relationship with other individuals in the dataset. By comparing the genetic relatedness and other demographic information (age, sex, county, etc.), an estimate of the individual's location within the pedigree could be made. Similar estimates could be made for individuals that the AGDB cannot connect into the pedigree, even with a known ID. These analyses would increase the sample size and power of the study without potentially having to perform additional genotyping.

Alternative methods of genotyping could be used to generate genotypes for the five candidate variants that failed in the follow-up phase detailed in Chapter IV. These

variants can be genotyped via alternative methods (TaqMan genotyping, direct Sanger sequencing, etc.) to overcome this limitation. These variants may be associated with LOAD in the Amish, but were not interrogated because of the failure to genotype. If these variants are associated in the Amish and that association generalizes in the unrelated dataset, new knowledge about the disease process could be learned.

Additional studies examining the non-exonic variation of known AD genes and the candidate linkage regions, as well as other areas of the genome (e.g. intronic regulatory elements, regions prone to epigenetic modifications, and previously unassociated genes), are likely to identify new variation that confers susceptibility to developing LOAD in the Amish. This study focused on three classes of genes and variants that occurred uniquely in cases or controls, significantly limiting the portion of the exome investigated. While the classes of genes studied were the most likely to harbor risk variants, it is still likely that other genes outside of these classes may contribute to LOAD. As common variation on genotyping chips has been investigated in the Amish, whole-genome sequencing would be an effective method to identify additional variation.

To efficiently conduct a whole-genome sequencing study, similar approaches using a screening subset of the large population may be necessary due to cost constraints. Selection of individuals for sequencing should be determined by the goals of the study (Cirulli & Goldstein, 2010). Within family-based studies, sequencing the most distantly related affected individuals lowers the number of variants that will be shared due to common ancestry and therefore increases the likelihood that a variant is shared because of common affection status. Although the Amish individuals sampled for this study are all

related to each other in a single pedigree, this work and previous studies have found that genetic heterogeneity does exist in this population (Cummings et al., 2012). To control for this heterogeneity, the single all-connecting pedigree could be broken into smaller subpedigrees from which the most distantly related individuals within these smaller families could be chosen for sequencing.

Another study design for sequencing a subset of individuals is selecting individuals at one or both ends of a disease or risk spectrum. Within the ascertained Amish dataset, the genetic risk score distributions detailed in Chapter II could be used to identify cognitively normal individuals with extremely high risk burdens and affected individuals with extremely low risk burdens. These controls may harbor protective variants that are modifying their high burden and these cases may harbor unidentified risk variants. By performing whole-genome sequencing in these well-defined subsets of the overall dataset, these types of functional non-exonic or modifying variants could be identified and then genotyped in the full dataset to test for association.

Whole-genome sequence data could also be used for imputation or creating a reference genome specific to the Amish (Holm et al., 2011; Le & Durbin, 2011). Imputation can fill in missing genotypes based on the correlation and predictability of already genotyped markers. Imputation can be a cheaper alternative to whole-genome sequencing every sample if enough genotypic and haplotype data exist to efficiently and effectively impute a large number of variants. If imputation was successful in the Amish, then follow-up direct genotyping of candidate variants may not be needed. Imputation accuracy is

affected by minor allele frequencies, population ancestry, genotyping platform, inclusion of trios, and reference panel size (Marchini & Howie, 2010).

Ascertainment of new individuals is ongoing in the context of the dementia study and another study focused on age-related macular degeneration. Continued focus on complete ascertainment of all individuals over the age of 60 would ensure that all members of the communities are included. Re-ascertainment should focus on individuals of “unknown” affection status and younger cognitively normal controls. Further neuropsychological study and re-evaluation of “unknown” individuals may help elucidate the primary underlying disease process and allow for better phenotyping of those individuals for future genetic analyses. By re-screening cognitively normal controls that are on the younger end of the age spectrum and individuals diagnosed with MCI, disease progression and genetic factors contributing to status conversion may be investigated in future studies. Continued re-evaluation of all subjects would reinforce the diagnoses and phenotypes assigned. While LOAD is a disease restricted to older individuals, ascertainment of the younger generations would incorporate the necessary genetic information needed for phasing and haplotype analysis that has been limited by little vertical genotyping in the current dataset.

REFERENCES

Abbott, R. D., White, L. R., Ross, G. W., Masaki, K. H., Curb, J. D., & Petrovitch, H. (2004). Walking and dementia in physically capable elderly men. *JAMA, 292*(12), 1447-1453. doi: 10.1001/jama.292.12.1447

Acevedo, A., & Loewenstein, D. A. (2007). Nonpharmacological cognitive interventions in aging and dementia. *J Geriatr Psychiatry Neurol, 20*(4), 239-249. doi: 10.1177/0891988707308808

Agarwala, R., Biesecker, L. G., & Schaffer, A. A. (2003). Anabaptist genealogy database. *Am J Med Genet C Semin Med Genet, 121C*(1), 32-37. doi: 10.1002/ajmg.c.20004

American Psychiatric Association., & American Psychiatric Association. DSM-5 Task Force. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5* (5th ed.). Washington, D.C.: American Psychiatric Association.

Arriagada, P. V., Growdon, J. H., Hedley-Whyte, E. T., & Hyman, B. T. (1992). Neurofibrillary tangles but not senile plaques parallel duration and severity of Alzheimer's disease. *Neurology, 42*(3 Pt 1), 631-639.

Ashley-Koch, A. E., Shao, Y., Rimmler, J. B., Gaskell, P. C., Welsh-Bohmer, K. A., Jackson, C. E., . . . Pericak-Vance, M. A. (2005). An autosomal genomic screen for dementia in an extended Amish family. *Neuroscience Letters, 379*(3), 199-204. doi: DOI 10.1016/j.neulet.2004.12.065

Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., . . . Vital Elderly Study, G. (2002). Effects of cognitive training interventions with older adults: a randomized controlled trial. *JAMA, 288*(18), 2271-2281.

Ballatore, C., Lee, V. M., & Trojanowski, J. Q. (2007). Tau-mediated neurodegeneration in Alzheimer's disease and related disorders. *Nat Rev Neurosci, 8*(9), 663-672. doi: 10.1038/nrn2194

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

Berkrot, B. (2012). Pfizer, J&J scrap Alzheimer's studies as drug fails. *Reuters*. Retrieved from reuters.com website:

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet, 39*(1), 17-23. doi: 10.1038/ng1934

Bomboi, G., Castello, L., Cosentino, F., Giubilei, F., Orzi, F., & Volpe, M. (2010). Alzheimer's disease and endothelial dysfunction. *Neurol Sci, 31*(1), 1-8. doi: 10.1007/s10072-009-0151-6

Brenner, D. E., Kukull, W. A., van Belle, G., Bowen, J. D., McCormick, W. C., Teri, L., & Larson, E. B. (1993). Relationship between cigarette smoking and Alzheimer's disease in a population-based case-control study. *Neurology, 43*(2), 293-300.

Brunnstrom, H. R., & Englund, E. M. (2009). Cause of death in patients with dementia disorders. *Eur J Neurol, 16*(4), 488-492. doi: 10.1111/j.1468-1331.2008.02503.x

Buee, L., Bussiere, T., Buee-Scherrer, V., Delacourte, A., & Hof, P. R. (2000). Tau protein isoforms, phosphorylation and role in neurodegenerative disorders. *Brain Res Brain Res Rev, 33*(1), 95-130.

Canadian study of health and aging: study methods and prevalence of dementia. (1994). *CMAJ, 150*(6), 899-913.

Carlson, M. C., Helms, M. J., Steffens, D. C., Burke, J. R., Potter, G. G., & Plassman, B. L. (2008). Midlife activity predicts risk of dementia in older male twin pairs. *Alzheimers Dement, 4*(5), 324-331. doi: 10.1016/j.jalz.2008.07.002

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet, 11*(6), 415-425. doi: 10.1038/nrg2779

ClinicalTrials.gov. Effect of LY2062430 on the Progression of Alzheimer's Disease (EXPEDITION2). from http://clinicaltrials.gov/ct2/show/study/NCT00904683?term=NCT00904683&rank=1

ClinicalTrials.gov. A Long-Term Safety And Tolerability Extension Study Of Bapineuzumab In Alzheimer Disease Patients. from http://clinicaltrials.gov/ct2/show/NCT00998764?term=NCT00998764&rank=1

Cooper, G. M., Stone, E. A., Asimenos, G., Program, N. C. S., Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res, 15*(7), 901-913. doi: 10.1101/gr.3577405

Coppola, G., Chinnathambi, S., Lee, J. J., Dombroski, B. A., Baker, M. C., Soto-Ortolaza, A. I., . . . Geschwind, D. H. (2012). Evidence for a role of the rare p.A152T variant in MAPT in increasing the risk for FTD-spectrum and Alzheimer's diseases. *Hum Mol Genet, 21*(15), 3500-3512. doi: 10.1093/hmg/dds161

Corder, E. H., Saunders, A. M., Risch, N. J., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Jr., . . . et al. (1994). Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nat Genet, 7*(2), 180-184. doi: 10.1038/ng0694-180

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., . . . Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science, 261*(5123), 921-923.

Cornelis, M. C., Qi, L., Zhang, C., Kraft, P., Manson, J., Cai, T., . . . Hu, F. B. (2009). Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. *Ann Intern Med, 150*(8), 541-550.

Corrada, M. M., Berlau, D. J., & Kawas, C. H. (2012). A population-based clinicopathological study in the oldest-old: the 90+ study. *Curr Alzheimer Res, 9*(6), 709-717.

Courtenay, M. D., Gilbert, J. R., Jiang, L., Cummings, A. C., Gallins, P. J., Caywood, L., . . . Scott, W. K. (2012). Mitochondrial Haplogroup X is associated with successful aging in the Amish. *Human Genetics, 131*(2), 201-208. doi: DOI 10.1007/s00439-011-1060-3

Cruchaga, C., Haller, G., Chakraverty, S., Mayo, K., Vallania, F. L., Mitra, R. D., . . . Consortium, N.-L. N. F. S. (2012). Rare variants in APP, PSEN1 and PSEN2 increase risk for AD in late-onset Alzheimer's disease families. *PLoS One, 7*(2), e31039. doi: 10.1371/journal.pone.0031039

Cullen, K. M., Kocsi, Z., & Stone, J. (2006). Microvascular pathology in the aging human brain: evidence that senile plaques are sites of microhaemorrhages. *Neurobiol Aging, 27*(12), 1786-1796.

Cummings, A. C., Jiang, L., Velez Edwards, D. R., McCauley, J. L., Laux, R., McFarland, L. L., . . . Haines, J. L. (2012). Genome-wide association and linkage study in the Amish detects a novel candidate late-onset Alzheimer disease gene. *Ann Hum Genet, 76*(5), 342-351. doi: 10.1111/j.1469-1809.2012.00721.x

Cummings, A. C., Torstenson, E., Davis, M. F., D'Aoust, L. N., Scott, W. K., Pericak-Vance, M. A., . . . Haines, J. L. (2013). Evaluating power and type 1 error in large pedigree analyses of binary traits. *PLoS One, 8*(5), e62615. doi: 10.1371/journal.pone.0062615

. Datasheet 5990-6319EN. (2010): Agilent Technologies.

Demuro, A., Mina, E., Kayed, R., Milton, S. C., Parker, I., & Glabe, C. G. (2005). Calcium dysregulation and membrane disruption as a ubiquitous neurotoxic mechanism of soluble amyloid oligomers. *J Biol Chem, 280*(17), 17294-17300. doi: 10.1074/jbc.M500997200

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet, 43*(5), 491-498. doi: 10.1038/ng.806

Doll, R., Peto, R., Boreham, J., & Sutherland, I. (2000). Smoking and dementia in male British doctors: prospective study. *BMJ, 320*(7242), 1097-1102.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C., Doyle, F., . . . Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi: Doi 10.1038/Nature11247

Edwards, D. R. V., Gilbert, J. R., Hicks, J. E., Myers, J. L., Jiang, L., Cummings, A. C., . . . Scott, W. K. (2013). Linkage and association of successful aging to the 6q25 region in large Amish kindreds. *Age, 35*(4), 1467-1477. doi: DOI 10.1007/s11357-012-9447-1

Edwards, D. R. V., Gilbert, J. R., Jiang, L., Gallins, P. J., Caywood, L., Creason, M., . . . Scott, W. K. (2011). Successful Aging Shows Linkage to Chromosomes 6, 7, and 14 in the Amish. *Ann Hum Genet, 75*, 516-528. doi: DOI 10.1111/j.1469-1809.2011.00658.x

Eli Lilly and Company Announces Top-Line Results on Solanezumab Phase 3 Clinical Trials in Patients with Alzheimer's Disease. (2012). [Press release]. Retrieved from https://investor.lilly.com/releasedetail.cfm?ReleaseID=702211

Evenhuis, H. M. (1990). The natural history of dementia in Down's syndrome. *Arch Neurol, 47*(3), 263-267.

Fleminger, S., Oliver, D. L., Lovestone, S., Rabe-Hesketh, S., & Giora, A. (2003). Head injury as a risk factor for Alzheimer's disease: the evidence 10 years on; a partial replication. *J Neurol Neurosurg Psychiatry, 74*(7), 857-862.

Fratiglioni, L., Paillard-Borg, S., & Winblad, B. (2004). An active and socially integrated lifestyle in late life might protect against dementia. *Lancet Neurol, 3*(6), 343-353. doi: 10.1016/S1474-4422(04)00767-7

Fratiglioni, L., & Wang, H. X. (2007). Brain reserve hypothesis in dementia. *J Alzheimers Dis, 12*(1), 11-22.

Friedman, R. C., Farh, K. K., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res, 19*(1), 92-105. doi: 10.1101/gr.082701.108

Fritsche, L. G., Chen, W., Schu, M., Yaspan, B. L., Yu, Y., Thorleifsson, G., . . . Consortium, A. M. D. G. (2013). Seven new loci associated with age-related macular degeneration. *Nat Genet, 45*(4), 433-439, 439e431-432. doi: 10.1038/ng.2578

Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., . . . Pedersen, N. L. (2006). Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry, 63*(2), 168-174. doi: 10.1001/archpsyc.63.2.168

Giannakopoulos, P., Herrmann, F. R., Bussiere, T., Bouras, C., Kovari, E., Perl, D. P., . . . Hof, P. R. (2003). Tangle and neuron numbers, but not amyloid load, predict cognitive status in Alzheimer's disease. *Neurology, 60*(9), 1495-1500.

Gibson, J., Cree, A., Collins, A., Lotery, A., & Ennis, S. (2010). Determination of a gene and environment risk model for age-related macular degeneration. *Br J Ophthalmol, 94*(10), 1382-1387. doi: 10.1136/bjo.2010.182568

Goate, A., Chartier-Harlin, M. C., Mullan, M., Brown, J., Crawford, F., Fidani, L., . . . et al. (1991). Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature, 349*(6311), 704-706. doi: 10.1038/349704a0

Gold, G., Kovari, E., Corte, G., Herrmann, F. R., Canuto, A., Bussiere, T., . . . Giannakopoulos, P. (2001). Clinical validity of A beta-protein deposition staging in brain aging and Alzheimer disease. *J Neuropathol Exp Neurol, 60*(10), 946-952.

Goldstein, F. C., Ashley, A. V., Gearing, M., Hanfelt, J., Penix, L., Freedman, L. J., & Levey, A. I. (2001). Apolipoprotein E and age at onset of Alzheimer's disease in African American patients. *Neurology, 57*(10), 1923-1925.

Graff-Radford, N. R., Green, R. C., Go, R. C., Hutton, M. L., Edeki, T., Bachman, D., . . . Farrer, L. A. (2002). Association between apolipoprotein E genotype and Alzheimer disease in African American subjects. *Arch Neurol, 59*(4), 594-600.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science, 185*(4154), 862-864.

Grassmann, F., Fritsche, L. G., Keilhauer, C. N., Heid, I. M., & Weber, B. H. (2012). Modelling the genetic risk in age-related macular degeneration. *PLoS One, 7*(5), e37979. doi: 10.1371/journal.pone.0037979

Guerreiro, R., Wojtas, A., Bras, J., Carrasquillo, M., Rogaeva, E., Majounie, E., . . . Alzheimer Genetic Analysis, G. (2013). TREM2 variants in Alzheimer's disease. *N Engl J Med, 368*(2), 117-127. doi: 10.1056/NEJMoa1211851

Guo, Z., Cupples, L. A., Kurz, A., Auerbach, S. H., Volicer, L., Chui, H., . . . Farrer, L. A. (2000). Head injury and the risk of AD in the MIRAGE study. *Neurology, 54*(6), 1316-1323.

Hahs, D. W., McCauley, J. L., Crunk, A. E., McFarland, L. L., Gaskell, P. C., Jiang, L., . . . Haines, J. L. (2006). A genome-wide linkage analysis of dementia in the Amish. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics, 141B*(2), 160-166. doi: Doi 10.1002/Ajmg.B.30257

Halekoh, U., Hojsgaard, S., & Yan, J. (2006). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software, 15*(2), 1-11.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., . . . Williams, J. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet, 41*(10), 1088-1093. doi: 10.1038/ng.440

Hebert, L. E., Scherr, P. A., Beckett, L. A., Funkenstein, H. H., Albert, M. S., Chown, M. J., & Evans, D. A. (1992). Relation of smoking and alcohol consumption to incident Alzheimer's disease. *Am J Epidemiol, 135*(4), 347-355.

Hebert, L. E., Weuve, J., Scherr, P. A., & Evans, D. A. (2013). Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology, 80*(19), 1778-1783. doi: 10.1212/WNL.0b013e31828726f5

Henderson, A. S., Easteal, S., Jorm, A. F., Mackinnon, A. J., Korten, A. E., Christensen, H., . . . Jacomb, P. A. (1995). Apolipoprotein E allele epsilon 4, dementia, and cognitive decline in a population sample. *Lancet, 346*(8987), 1387-1390.

Hindorff, L., MacArthur, J (European Bioinformatics Institute), Morales, J (European Bioinformatics Institute), Junkins, HA, Hall, PN, Klemm, AK, and Manolio, TA. A Catalog of Published Genome-Wide Association Studies. . Retrieved January 2, 2014, from www.genome.gov/gwastudies

Holden, K. F., Lindquist, K., Tylavsky, F. A., Rosano, C., Harris, T. B., Yaffe, K., & Health, A. B. C. s. (2009). Serum leptin level and cognition in the elderly: Findings from the Health ABC Study. *Neurobiol Aging, 30*(9), 1483-1489. doi: 10.1016/j.neurobiolaging.2007.11.024

Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J. C., Carrasquillo, M. M., . . . Williams, J. (2011). Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat Genet, 43*(5), 429-435. doi: 10.1038/ng.803

Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadottir, H. T., Zanon, C., . . . Stefansson, K. (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet, 43*(4), 316-320. doi: 10.1038/ng.781

Hsiung, G. Y., Sadovnick, A. D., & Feldman, H. (2004). Apolipoprotein E epsilon4 genotype as a risk factor for cognitive decline and dementia: data from the Canadian Study of Health and Aging. *CMAJ, 171*(8), 863-867. doi: 10.1503/cmaj.1031789

Jakob-Roetne, R., & Jacobsen, H. (2009). Alzheimer's disease: from pathology to therapeutic approaches. *Angew Chem Int Ed Engl, 48*(17), 3030-3059. doi: 10.1002/anie.200802808

Jonsson, T., Atwal, J. K., Steinberg, S., Snaedal, J., Jonsson, P. V., Bjornsson, S., . . . Stefansson, K. (2012). A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature, 488*(7409), 96-99. doi: 10.1038/nature11283

Jonsson, T., Stefansson, H., Steinberg, S., Jonsdottir, I., Jonsson, P. V., Snaedal, J., . . . Stefansson, K. (2013). Variant of TREM2 associated with the risk of Alzheimer's disease. *N Engl J Med, 368*(2), 107-116. doi: 10.1056/NEJMoa1211103

Karlin, S., & Mrazek, J. (1996). What drives codon choices in human genes? *J Mol Biol, 262*(4), 459-472. doi: 10.1006/jmbi.1996.0528

Khachaturian, A. S., Gallo, J. J., & Breitner, J. C. (2000). Performance characteristics of a two-stage dementia screen in a population sample. *J Clin Epidemiol, 53*(5), 531-540.

King, M. E., Kan, H. M., Baas, P. W., Erisir, A., Glabe, C. G., & Bloom, G. S. (2006). Tau-dependent microtubule disassembly initiated by prefibrillar beta-amyloid. *J Cell Biol, 175*(4), 541-546. doi: 10.1083/jcb.200605187

Kivipelto, M., Helkala, E. L., Hanninen, T., Laakso, M. P., Hallikainen, M., Alhainen, K., . . . Nissinen, A. (2001). Midlife vascular risk factors and late-life mild cognitive impairment: A population-based study. *Neurology, 56*(12), 1683-1689.

Komar, A. A., Lesnik, T., & Reiss, C. (1999). Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett, 462*(3), 387-391.

Korenberg, J. R., Chen, X. N., Schipper, R., Sun, Z., Gonsky, R., Gerwehr, S., . . . et al. (1994). Down syndrome phenotypes: the consequences of chromosomal imbalance. *Proc Natl Acad Sci U S A, 91*(11), 4997-5001.

Lambert, J. C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., . . . Amouyel, P. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet, 41*(10), 1094-1099. doi: 10.1038/ng.439

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., . . . Amouyel, P. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet*. doi: 10.1038/ng.2802

Lander, E. S., & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics, 121*(1), 185-199.

Launer, L. J., Andersen, K., Dewey, M. E., Letenneur, L., Ott, A., Amaducci, L. A., . . . Hofman, A. (1999). Rates and risk factors for dementia and Alzheimer's disease: results from EURODEM pooled analyses. EURODEM Incidence Research Group and Work Groups. European Studies of Dementia. *Neurology, 52*(1), 78-84.

Launer, L. J., Masaki, K., Petrovitch, H., Foley, D., & Havlik, R. J. (1995). The association between midlife blood pressure levels and late-life cognitive function. The Honolulu-Asia Aging Study. *JAMA, 274*(23), 1846-1851.

Lavner, Y., & Kotlar, D. (2005). Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene, 345*(1), 127-138. doi: 10.1016/j.gene.2004.11.035

Le, S. Q., & Durbin, R. (2011). SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res, 21*(6), 952-960. doi: 10.1101/gr.113084.110

Lee, J. H., Cheng, R., Schupf, N., Manly, J., Lantigua, R., Stern, Y., . . . Mayeux, R. (2007). The association between genetic variants in SORL1 and Alzheimer disease in an urban, multiethnic, community-based cohort. *Arch Neurol, 64*(4), 501-506. doi: 10.1001/archneur.64.4.501

Lee, S. H., Harold, D., Nyholt, D. R., Consortium, A. N., International Endogene, C., Genetic, . . . Visscher, P. M. (2013). Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum Mol Genet, 22*(4), 832-841. doi: 10.1093/hmg/dds491

Leibson, C. L., Rocca, W. A., Hanson, V. A., Cha, R., Kokmen, E., O'Brien, P. C., & Palumbo, P. J. (1997). Risk of dementia among persons with diabetes mellitus: a population-based cohort study. *Am J Epidemiol, 145*(4), 301-308.

Lesser, G., Kandiah, K., Libow, L. S., Likourezos, A., Breuer, B., Marin, D., . . . Neufeld, R. (2001). Elevated serum total and LDL cholesterol in very old patients with Alzheimer's disease. *Dement Geriatr Cogn Disord, 12*(2), 138-145. doi: 51248

Levy-Lahad, E., Wasco, W., Poorkaj, P., Romano, D. M., Oshima, J., Pettingell, W. H., . . . et al. (1995). Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science, 269*(5226), 973-977.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi: 10.1093/bioinformatics/btp324

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika, 73*(1), 13-22. doi: DOI 10.1093/biomet/73.1.13

Luchsinger, J. A., Tang, M. X., Stern, Y., Shea, S., & Mayeux, R. (2001). Diabetes mellitus and risk of Alzheimer's disease and dementia with stroke in a multiethnic cohort. *Am J Epidemiol, 154*(7), 635-641.

Mann, D. M. (1988). The pathological association between Down syndrome and Alzheimer disease. *Mech Ageing Dev, 43*(2), 99-136.

Manolio, T. A. (2013). Bringing genome-wide association findings into clinical use. *Nature Reviews Genetics, 14*(8), 549-558. doi: Doi 10.1038/Nrg3523

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet, 11*(7), 499-511. doi: 10.1038/nrg2796

. MassARRAY Typer 3.4 Software User's Guide for iPLEX and hME. (2006). In Sequenom (Ed.), (Vol. Doc 11546, R03 CO 060094).

McCauley, J. L., Hahs, D. W., Jiang, L., Scott, W. K., Welsh-Bohmer, K. A., Jackson, C. E., . . . Haines, J. L. (2006). Combinatorial Mismatch Scan (CMS) for loci associated with dementia in the Amish. *Bmc Medical Genetics, 7*. doi: Artn 19

Doi 10.1186/1471-2350-7-19

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology, 34*(7), 939-944.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., . . . Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers & Dementia, 7*(3), 263-269. doi: DOI 10.1016/j.jalz.2011.03.005

Merchant, C., Tang, M. X., Albert, S., Manly, J., Stern, Y., & Mayeux, R. (1999). The influence of smoking on the risk of Alzheimer's disease. *Neurology, 52*(7), 1408-1412.

Michikawa, M. (2003). Cholesterol paradox: is high total or low HDL cholesterol level a risk for Alzheimer's disease? *J Neurosci Res, 72*(2), 141-146. doi: 10.1002/jnr.10585

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol, 12*(11), R112. doi: 10.1186/gb-2011-12-11-r112

Morris, J. C., Heyman, A., Mohs, R. C., Hughes, J. P., van Belle, G., Fillenbaum, G., . . . Clark, C. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology, 39*(9), 1159-1165.

Mortimer, J. A., van Duijn, C. M., Chandra, V., Fratiglioni, L., Graves, A. B., Heyman, A., . . . et al. (1991). Head trauma as a risk factor for Alzheimer's disease: a collaborative re-analysis of case-control studies. EURODEM Risk Factors Research Group. *Int J Epidemiol, 20 Suppl 2*, S28-35.

Myers, R. H., Schaefer, E. J., Wilson, P. W., D'Agostino, R., Ordovas, J. M., Espino, A., . . . Wolf, P. A. (1996). Apolipoprotein E epsilon4 association with dementia in a population-based study: The Framingham study. *Neurology, 46*(3), 673-677.

Naj, A. C., Jun, G., Beecham, G. W., Wang, L. S., Vardarajan, B. N., Buros, J., . . . Schellenberg, G. D. (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet, 43*(5), 436-441. doi: 10.1038/ng.801

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet, 12*(6), 443-451. doi: 10.1038/nrg2986

Ott, A., Slooter, A. J., Hofman, A., van Harskamp, F., Witteman, J. C., Van Broeckhoven, C., . . . Breteler, M. M. (1998). Smoking and risk of dementia and Alzheimer's disease in a population-based cohort study: the Rotterdam Study. *Lancet, 351*(9119), 1840-1843.

Ott, A., Stolk, R. P., van Harskamp, F., Pols, H. A., Hofman, A., & Breteler, M. M. (1999). Diabetes mellitus and the risk of dementia: The Rotterdam Study. *Neurology, 53*(9), 1937-1942.

Oyama, F., Cairns, N. J., Shimada, H., Oyama, R., Titani, K., & Ihara, Y. (1994). Down's syndrome: up-regulation of beta-amyloid protein precursor and tau mRNAs and their defective coordination. *J Neurochem, 62*(3), 1062-1066.

Pagani, F., Raponi, M., & Baralle, F. E. (2005). Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc Natl Acad Sci U S A, 102*(18), 6368-6372. doi: 10.1073/pnas.0502288102

Pendlebury, S. T., & Rothwell, P. M. (2009). Prevalence, incidence, and factors associated with pre-stroke and post-stroke dementia: a systematic review and meta-analysis. *Lancet Neurol, 8*(11), 1006-1018. doi: 10.1016/S1474-4422(09)70236-4

PericakVance, M. A., Johnson, C. C., Rimmler, J. B., Saunders, A. M., Robinson, L. C., DHondt, E. G., . . . Haines, J. L. (1996). Alzheimer's disease and apolipoprotein E-4 allele in an Amish population. *Ann Neurol, 39*(6), 700-704. doi: DOI 10.1002/ana.410390605

Plassman, B. L., Havlik, R. J., Steffens, D. C., Helms, M. J., Newman, T. N., Drosdick, D., . . . Breitner, J. C. (2000). Documented head injury in early adulthood and risk of Alzheimer's disease and other dementias. *Neurology, 55*(8), 1158-1166.

Podewils, L. J., Guallar, E., Kuller, L. H., Fried, L. P., Lopez, O. L., Carlson, M., & Lyketsos, C. G. (2005). Physical activity, APOE genotype, and dementia risk: findings from the Cardiovascular Health Cognition Study. *Am J Epidemiol, 161*(7), 639-651. doi: 10.1093/aje/kwi092

Polvikoski, T., Sulkava, R., Haltia, M., Kainulainen, K., Vuorio, A., Verkkoniemi, A., . . . Kontula, K. (1995). Apolipoprotein E, dementia, and cortical deposition of beta-amyloid protein. *N Engl J Med, 333*(19), 1242-1247. doi: 10.1056/NEJM199511093331902

Powers, E. T., & Balch, W. E. (2008). Costly mistakes: translational infidelity and protein homeostasis. *Cell, 134*(2), 204-206. doi: 10.1016/j.cell.2008.07.005

Raffaitin, C., Gin, H., Empana, J. P., Helmer, C., Berr, C., Tzourio, C., . . . Barberger-Gateau, P. (2009). Metabolic syndrome and risk for incident Alzheimer's disease or vascular dementia: the Three-City Study. *Diabetes Care, 32*(1), 169-174. doi: 10.2337/dc08-0272

Ramensky, V., Bork, P., & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res, 30*(17), 3894-3900.

Reitz, C., Brayne, C., & Mayeux, R. (2011). Epidemiology of Alzheimer disease. *Nat Rev Neurol, 7*(3), 137-152. doi: 10.1038/nrneurol.2011.2

Reitz, C., Cheng, R., Rogaeva, E., Lee, J. H., Tokuhiro, S., Zou, F., . . . Environmental Risk in Alzheimer Disease, C. (2011). Meta-analysis of the association between variants in SORL1 and Alzheimer disease. *Arch Neurol, 68*(1), 99-106. doi: 10.1001/archneurol.2010.346

Reitz, C., Tokuhiro, S., Clark, L. N., Conrad, C., Vonsattel, J. P., Hazrati, L. N., . . . Mayeux, R. (2011). SORCS1 alters amyloid precursor protein processing and variants may increase Alzheimer's disease risk. *Ann Neurol, 69*(1), 47-64. doi: 10.1002/ana.22308

Reitz, C., Tosto, G., Vardarajan, B., Rogaeva, E., Ghani, M., Rogers, R. S., . . . Alzheimer's Disease Genetics, C. (2013). Independent and epistatic effects of variants in VPS10-d receptors on Alzheimer disease risk and processing of the amyloid precursor protein (APP). *Transl Psychiatry, 3*, e256. doi: 10.1038/tp.2013.13

Rodriguez-Rodriguez, E., Sanchez-Juan, P., Vazquez-Higuera, J. L., Mateo, I., Pozueta, A., Berciano, J., . . . Combarros, O. (2013). Genetic risk score predicting accelerated progression from mild cognitive impairment to Alzheimer's disease. *J Neural Transm, 120*(5), 807-812. doi: 10.1007/s00702-012-0920-x

Rogaev, E. I., Sherrington, R., Rogaeva, E. A., Levesque, G., Ikeda, M., Liang, Y., . . . et al. (1995). Familial Alzheimer's disease in kindreds with missense mutations in a

gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature, 376*(6543), 775-778. doi: 10.1038/376775a0

Rogaeva, E., Meng, Y., Lee, J. H., Gu, Y., Kawarai, T., Zou, F., . . . St George-Hyslop, P. (2007). The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. *Nat Genet, 39*(2), 168-177. doi: 10.1038/ng1943

Rovio, S., Kareholt, I., Helkala, E. L., Viitanen, M., Winblad, B., Tuomilehto, J., . . . Kivipelto, M. (2005). Leisure-time physical activity at midlife and the risk of dementia and Alzheimer's disease. *Lancet Neurol, 4*(11), 705-711. doi: 10.1016/S1474-4422(05)70198-8

Savelieff, M. G., Lee, S., Liu, Y., & Lim, M. H. (2013). Untangling amyloid-beta, tau, and metals in Alzheimer's disease. *ACS Chem Biol, 8*(5), 856-865. doi: 10.1021/cb400080f

Scarmeas, N., Levy, G., Tang, M. X., Manly, J., & Stern, Y. (2001). Influence of leisure activity on the incidence of Alzheimer's disease. *Neurology, 57*(12), 2236-2242.

Schofield, P. W., Tang, M., Marder, K., Bell, K., Dooneief, G., Chun, M., . . . Mayeux, R. (1997). Alzheimer's disease after remote head injury: an incidence study. *J Neurol Neurosurg Psychiatry, 62*(2), 119-124.

Seddon, J. M., Reynolds, R., Maller, J., Fagerness, J. A., Daly, M. J., & Rosner, B. (2009). Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. *Invest Ophthalmol Vis Sci, 50*(5), 2044-2053. doi: 10.1167/iovs.08-3064

Seshadri, S., Beiser, A., Kelly-Hayes, M., Kase, C. S., Au, R., Kannel, W. B., & Wolf, P. A. (2006). The lifetime risk of stroke: estimates from the Framingham Study. *Stroke, 37*(2), 345-350. doi: 10.1161/01.STR.0000199613.38911.b2

Seshadri, S., Fitzpatrick, A. L., Ikram, M. A., DeStefano, A. L., Gudnason, V., Boada, M., . . . Consortium, E. (2010). Genome-wide analysis of genetic loci associated with Alzheimer disease. *JAMA, 303*(18), 1832-1840. doi: 10.1001/jama.2010.574

Sherrington, R., Rogaev, E. I., Liang, Y., Rogaeva, E. A., Levesque, G., Ikeda, M., . . . St George-Hyslop, P. H. (1995). Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature, 375*(6534), 754-760. doi: 10.1038/375754a0

Shi, M. M., Myrand, S. P., Bleavins, M. R., & de la Iglesia, F. A. (1999). High throughput genotyping for the detection of a single nucleotide polymorphism in NAD(P)H quinone oxidoreductase (DT diaphorase) using TaqMan probes. *Mol Pathol, 52*(5), 295-299.

Sidak, Z. (1968). On Multivariate Normal Probabilities of Rectangles - Their Dependence on Correlations. *Annals of Mathematical Statistics, 39*(5), 1425-&.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., . . . Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res, 15*(8), 1034-1050. doi: 10.1101/gr.3715005

Skoog, I., Hesse, C., Aevarsson, O., Landahl, S., Wahlstrom, J., Fredman, P., & Blennow, K. (1998). A population study of apoE genotype at the age of 85: relation to dementia, cerebrovascular disease, and mortality. *J Neurol Neurosurg Psychiatry, 64*(1), 37-43.

Slooter, A. J., Cruts, M., Hofman, A., Koudstaal, P. J., van der Kuip, D., de Ridder, M. A., . . . van Duijn, C. M. (2004). The impact of APOE on myocardial infarction, stroke, and dementia: the Rotterdam Study. *Neurology, 62*(7), 1196-1198.

So, H. C., Gui, A. H., Cherny, S. S., & Sham, P. C. (2011). Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genetic Epidemiology, 35*(5), 310-317. doi: 10.1002/gepi.20579

Sokolov, Y., Kozak, J. A., Kayed, R., Chanturiya, A., Glabe, C., & Hall, J. E. (2006). Soluble amyloid oligomers increase bilayer conductance by altering dielectric structure. *J Gen Physiol, 128*(6), 637-647. doi: 10.1085/jgp.200609533

Solfrizzi, V., Scafato, E., Capurso, C., D'Introno, A., Colacicco, A. M., Frisardi, V., . . . Italian Longitudinal Study on Ageing Working, G. (2010). Metabolic syndrome and the risk of vascular dementia: the Italian Longitudinal Study on Ageing. *J Neurol Neurosurg Psychiatry, 81*(4), 433-440. doi: 10.1136/jnnp.2009.181743

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., & Roses, A. D. (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci U S A, 90*(5), 1977-1981.

Teng, E. L., & Chui, H. C. (1987). The Modified Mini-Mental State (3MS) examination. *J Clin Psychiatry, 48*(8), 314-318.

Thies, W., Bleiler, L., & Alzheimer's, A. (2013). 2013 Alzheimer's disease facts and figures. *Alzheimers Dement, 9*(2), 208-245. doi: 10.1016/j.jalz.2013.02.003

Thornton, T., & McPeek, M. S. (2007). Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am J Hum Genet, 81*(2), 321-337. doi: 10.1086/519497

Tombaugh, T. N. (2005). Test-retest reliable coefficients and 5-year change scores for the MMSE and 3MS. *Arch Clin Neuropsychol, 20*(4), 485-503. doi: 10.1016/j.acn.2004.11.004

Tsai, C. J., Sauna, Z. E., Kimchi-Sarfaty, C., Ambudkar, S. V., Gottesman, M. M., & Nussinov, R. (2008). Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *J Mol Biol, 383*(2), 281-291. doi: 10.1016/j.jmb.2008.08.012

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., . . . Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell, 141*(2), 344-354. doi: 10.1016/j.cell.2010.03.031

Unverzagt, F. W., Kasten, L., Johnson, K. E., Rebok, G. W., Marsiske, M., Koepke, K. M., . . . Tennstedt, S. L. (2007). Effect of memory impairment on training outcomes in ACTIVE. *J Int Neuropsychol Soc, 13*(6), 953-960. doi: 10.1017/S1355617707071512

van der Walt, J. M., Scott, W. K., Slifer, S., Gaskell, P. C., Martin, E. R., Welsh-Bohmer, K., . . . Pericak-Vance, M. A. (2005). Maternal lineages and Alzheimer disease risk in the Old Order Amish. *Human Genetics, 118*(1), 115-122. doi: DOI 10.1007/s00439-005-0032-x

Verghese, J., Lipton, R. B., Hall, C. B., Kuslansky, G., & Katz, M. J. (2003). Low blood pressure and the risk of dementia in very old individuals. *Neurology, 61*(12), 1667-1672.

Verghese, J., Lipton, R. B., Katz, M. J., Hall, C. B., Derby, C. A., Kuslansky, G., . . . Buschke, H. (2003). Leisure activities and the risk of dementia in the elderly. *N Engl J Med, 348*(25), 2508-2516. doi: 10.1056/NEJMoa022252

Virtanen, I., Gullberg, D., Rissanen, J., Kivilaakso, E., Kiviluoto, T., Laitinen, L. A., . . . Ekblom, P. (2000). Laminin alpha1-chain shows a restricted distribution in epithelial basement membranes of fetal and adult human tissues. *Exp Cell Res, 257*(2), 298-309. doi: 10.1006/excr.2000.4883

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res, 38*(16), e164. doi: 10.1093/nar/gkq603

Whitmer, R. A., Sidney, S., Selby, J., Johnston, S. C., & Yaffe, K. (2005). Midlife cardiovascular risk factors and risk of dementia in late life. *Neurology, 64*(2), 277-281. doi: 10.1212/01.WNL.0000149519.47454.F2

WHO. (2011, June 2011). Fact Sheet No 310.

Wieringa, G. E., Burlinson, S., Rafferty, J. A., Gowland, E., & Burns, A. (1997). Apolipoprotein E genotypes and serum lipid levels in Alzheimer's disease and multi-infarct dementia. *Int J Geriatr Psychiatry, 12*(3), 359-362.

Yousif, L. F., Di Russo, J., & Sorokin, L. (2013). Laminin isoforms in endothelial and perivascular basement membranes. *Cell Adh Migr, 7*(1), 101-110. doi: 10.4161/cam.22680

Zeenko, V. V., Wang, C., Majumder, M., Komar, A. A., Snider, M. D., Merrick, W. C., . . . Hatzoglou, M. (2008). An efficient in vitro translation system from mammalian cells lacking the translational inhibition caused by eIF2 phosphorylation. *RNA, 14*(3), 593-602. doi: 10.1261/rna.825008

Zhang, G., Hubalewska, M., & Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol, 16*(3), 274-280. doi: 10.1038/nsmb.1554