Exploring potential functional variants in the Alzheimer's disease associated genes, *CD2AP*, *EPHA1* and *CD33.*

Anne Braae, BSc. MSc.

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

July 2016

Abstract

Little is known about the molecular biology of late onset Alzheimer's disease (LOAD), the most common dementia in the elderly. Genetic loci associated with LOAD have been identified through genome-wide association studies (GWAS). However, the functional variants responsible for the observed GWAS association at each of the loci remain unknown. The aim of this project was to identify and assess potential functional rare variants at three associated loci, *CD2AP*, *EPHA1* and *CD33*. Target enriched, pooled sequencing of 96 post-mortem confirmed LOAD patient samples was used to identify 1273 variants within the three GWAS loci. Variants were prioritised using a combination of *in silico* functional annotation and putative disease association. Disease association was assessed through comparison to an independent, imputed LOAD GWAS dataset (2067 cases, 7376 controls). 18 coding and untranslated region variants and 9 noncoding variants were prioritised for further investigation. Potential splicing variants in *CD2AP* (6:47544253A>G) and *EPHA1* (rs6967117) were assessed using minigene assays, although neither were found to influence splicing products *in vivo*. Five untranslated variants from the three genes and a frameshift variant in *CD33* (rs201074739) were assessed for potential *cis*-regulatory consequences using allelic expression imbalance in brain tissues and B-lymphoblast cell lines. Only the frameshift variant displayed significant allelic expression imbalance and was found to be targeted for nonsense-mediated decay. None of the prioritised variants investigated were both functional and significantly associated with LOAD. However, pooled next generation sequencing using target enrichment successfully identified potential functional rare variants in *CD2AP*, *EPHA1* and *CD33*. Rare variants do have a role to play in late onset Alzheimer's disease. With the development of additional functional databases and improvements imputing rare variants from GWAS datasets, the combined prioritisation strategy used in this thesis will be useful for similar studies investigating causal GWAS variants.

Publications

Journal Articles

Anne Braae, Naomi Clement, James Turton, Jenny Lord, Tamar Guetta-Baranes … et al. Prioritizing splicing variants uncovered by next-generation sequencing the Alzheimer's disease candidate genes, *CLU*, *PICALM*, *CR1*, *ABCA7*, *BIN1*, *MS4A*, *CD2AP*, *EPHA1* and *CD33*. In preparation.

Imelda S. Barber, **Anne Braae**, Naomi Clement, Tulsi Patel, Tamar Guetta-Baranes et al. Mutation Analysis of Sporadic Early-Onset Alzheimer's Disease (sEOAD) Using the NeuroX Array. Experimental Neurology. Submitted September 2015.

Imelda S. Barber, Jennyfer M. García-Cárdenas, Chidchanok Sakdapanichkul, Christopher Deacon, Gabriela Zapata Erazo … **Anne Braae** … et al. 2015. Screening Exon 16 and 17 of the Amyloid Precursor Protein Gene in Sporadic Early-Onset Alzheimer's Disease (sEOAD). Neurobiology of Aging. Accepted pending minor revisions September 2015.

Rita Guerreiro, Valentina Escott-Price, Lee Darwent, Laura Parkkinen, Olaf Ansorge … **Anne Braae** … et al. 2015. Genome-wide analysis of genetic correlation in Dementia with Lewy Bodies, Parkinson's and Alzheimer's diseases. Neurobiology of Aging. In press, doi: 10.1016/j.neurobiolaging.2015.10.028.

Anne Braae, Christopher Medway, Minerva Carrasquillo, Steven Younkin, Patrick G Kehoe, Kevin Morgan and Alzheimer's Research UK (ARUK) Consortium. 2015. Blood type gene locus has no influence on ACE association with Alzheimer's disease. Neurobiology of Aging. 36(4): 1767.e1-2.

Jose Bras, Rita Guerreiro, Lee Darwent, Laura Parkkinen, Olaf Ansorge … **Anne Braae** … et al. 2014. Genetic analysis implicates APOE and lysosomal dysfunction in the etiology of Dementia with Lewy Bodies. Human Molecular Genetics. 23(23): 6139-614.

Carlos Cruchaga, Celeste M. Karch, Sheng Chih Jin, Bruno A. Benitez, Yefei Cai … **Anne Braae** …et al. 2014. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature 505 (7484): 550-554.

Bruno A. Benitez, Sheng Chih Jin, Rita Guerreiro, Rob Graham, Jenny Lord … **Anne Braae** … et al. 2014. Missense variant in TREML2 protects against Alzheimer's disease. Neurobiology of Aging. 35(6): 1510.e19-26.

Book Chapter

Christopher Medway, **Anne Braae**, Kevin Morgan. 2013. Erythropoietin-Producing Human Hepatocellular Carcinoma (EphA1). In: K. Morgan and M.M. Carrasquillo Ed. Genetic Variants in Alzheimer's Disease. ISBN 978-1-4614-7308-4.

Acknowledgements

There are many people who helped and supported me during my PhD. First, I would like to thank my supervisors Kevin Morgan and Noor Kalsheker for their support, guidance and encouragement. I am particularly grateful to Kevin for providing me with so many interesting opportunities during my project.

I am also indebted to the Alzheimer's Research UK (ARUK) for the initial funding of my PhD and for their support of the brain and DNA bank hosted at the Human Genetics laboratory at the University of Nottingham. The Big Lottery Fund also provided financial support for this work. Thank you to Rita Guerreiro and John Hardy at UCL for access to the Exome sequencing project data. I am also grateful for access to the ARUK/Mayo GWAS dataset, the Wellcome Trust Case Control Consortium dataset and the 1000 Genomes reference panel which allowed the imputation and subsequent association testing of rare variants.

Many of the laboratory techniques would have been impossible without advice from Tamar Guetta-Baranes. I am extremely grateful for the enthusiastic discussions provided by Sally Chappell. I also thank her for her insightful comments on Chapters 5 and 6 of this thesis. I thank Christopher Medway for his guidance on logistic regression. Thanks also to the others in the lab, Imelda Barber, Naomi Clement, Kristelle Brown, Keeley Brookes and Tulsi Patel for creating a fun, supportive and enjoyable working environment. To the PhD students who came before me, Jenny Lord and James C. Turton. I thank you for your initial work on the resequencing project. James, a special thank you for always taking the time to discuss my bioinformatics queries. Thanks to Lucy Millar (BMed.Sci. student), Lena Karrar (MSc. student) and Natalie Barker (MSc. student) for their dissertation work which is featured in Chapter 4, 5 and 6.

To my family and friends, a huge thank you for listening and for your advice and support these last few years. Patrick, I certainly wouldn't be where I am today without you. Thank you for always believing in me.

Lastly, I would like to thank the individuals who donated their tissues for research as without them, none of the research in this PhD would have been possible.

Table of Contents

List of abbreviations

3D	three dimensions
3C	chromatin conformation capture
Aβ	amyloid beta
ABCA7	ATP-binding cassette subfamily A member 7
ACH	amyloid cascade hypothesis
AD	Alzheimer's disease
ADGC	Alzheimer's Disease Genetics Consortium
AD-IG	Alzheimer's Disease Integrated Genome Research
ADRDA	Alzheimer's Disease and Related Disorders Association
AEI	allelic expression imbalance
ALS	amyotrophic lateral sclerosis
Alt	alternative allele
AML	acute myeloid leukemia
APOE	apolipoprotein E
APP	amyloid precursor protein
AR	ampicillin resistance gene
ARUK	Alzheimer's Research UK
AS	alternative splicing
BAM file	binary sequence alignment map file
BBB	blood brain barrier
BDGP	Berkley Drosophila Genome Project
BIN1	bridging integrator protein-1
Bp	base pair
CAA	cerebral amyloid angiopathy
CADD	Combined Annotation Dependent Depletion
CAL	candidate alignment locations
CD2AP	CD2-associated protein
CD33	sialic acid binding immunoglobulin-like lectin-3
CDCV	common disease/common variant
cDNA	complementary DNA
CDRV	common disease rare variant
CFS	cerebrospinal fluid
CHARGE	Cohort for Heart and Aging Research in Genomic Epidemiology
ChIP	chromatin immunoprecipitation
CLU	clusterin
COS-7	CV-1 in Origin carrying SV40

CR1	complement receptor type 1
CREB	cAMP response element-binding protein
CRISP	comprehensive read analysis for identification of single nucleotide polymorphisms from pooled sequencing
DMEM	Dulbecco's Modified Eagle Medium
DNA	deoxyribonucleic acid
EADI	European Alzheimer Disease Initiative
EBV	Epstein-Barr virus
ECACC	European Collection of Cell Cultures
ECRs	evolutionary conserved regions
EMEM	Eagle's Minimal Essential Medium
ENCODE	Encyclopedia of DNA Elements
EOFAD	early onset familial Alzheimer's disease
EPHA1	erythropoietin-producing human hepatocellular carcinoma 1
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
EtBr	ethidium bromide
eQTL	expression quantitative trait location
FBS	fetal bovine serum
FTD	frontotemporal dementia
GATK	genome analysis toolkit
gDNA	genomic DNA
GERAD	Genetics and Environmental Risk in Alzheimer's Disease consortium
GPI	glycophosphatidylinositol
GWAS	genome wide association study
HDL	high-density lipoprotein
HMPAO	hexamethylpropyleneamine oxime
hnRNP	heterogeneous nuclear ribonucleoprotein
HSF	human splicing finder
ID	identification
IGAP	International Genomics of Alzheimer's Project
iPSC	induced pluripotent stem cells
ISE	intronic splicing enhancer
ISS	intronic splicing silencer
ITIM	immunoreceptor tyrosine-based inhibitory motif
KASP	Competitive allele specific PCR

kDa	kilodalton
LBD	Lewy body disease
LCL	B-lymphoblastoid cell line
LD	linkage disequilibrium
lncRNA	long noncoding RNAs
LOAD	late onset Alzheimer's disease
MAF	minor allele frequency
MAPT	microtubule associated protein tau
MCI	mild cognitive impairment
MCMC	Markov Chain Monte Carlo
MCS	multiple cloning site
miRNA	micro RNA
MMSE	mini mental state examination
mRNA	messenger RNA
MS4A	membrane-spanning 4 domain family, subfamily A
NBS	National Blood Service
NCBI	National Center for Biotechnology Information
NEAA	non-essential amino acids
NEB	New England Biolabs inc.
NFT	neurofibrillary tau tangles
NGS	next generation sequencing
NHA	normal human astrocytes
NHGRI	National Human Genome Research Institute
NIA	National Institute on Aging
NICE	National Institute for Health and Care Excellence
NINCDS	National Institute of Neurological and Communicative Disorders and Stroke
NMD	nonsense-mediated decay
NMDA	N-methyl D-aspartate
NTC	no template control
OR	odds ratio
ori	origin of replication
PCR	polymerase chain reaction
PET	positron emission tomography
PICALM	phosphatidylinositol binding clathrin assembly protein
PLD3	phospholipase D3
PMI	post-mortem interval

pre-mRNA	precursor messenger RNA
PSEN1	presenilin 1
PSEN2	presenilin 2
P-tau	phosphorylated tau
PTV	protein-truncating variants
QC	quality control
RE	restriction enzymes
Ref	reference allele
RIN	RNA integrity number
RNA	ribonucleic acid
RNP	ribonucleoprotein
rRNA	ribosomal RNA
RPMI	Roswell Park Memorial Institute
rsID	reference SNP cluster identification
RS LTR	Rous sarcoma virus long terminal repeat
SAM	sterile alpha motif
SAM file	sequence alignment map file
SMA	spinal muscular atrophy
SNP	single nucleotide polymorphism
snRNP	small nuclear ribonucleoprotein
SORL1	sortilin-related receptor precursor 1
SPECT	single-photon emission computed tomography
SR	serine arginine
SRSF	serine arginine splicing factor
SV	SV40 origin of replication
SV40	simian vacuolating virus 40
Ta	annealing temperature
TAD	topologically associated domains
TAE	tris-acetate-EDTA
TE	target enrichment
TF	transcription factor
TFBS	transcription factor binding site
TRANSFAC	TRANScription FACtor database
TREM2	triggering receptor expressed on myeloid cells
TSS	transcription start site
TsTv	transition transversion ratio
T-tau	total tau

UCL	University College London
UTR	untranslated region
UV	ultraviolet
VCF	variant call format
VD	vascular disease
VEP	variant effect predictor
WES	whole exome sequencing
WGS	whole genome sequencing
ZYX	zyxin

1. Introduction

1.1. Alzheimer's and Dementia

Alzheimer's disease is a steady decline in cognitive function, including loss of memory and intellect. In the UK alone, dementia currently affects about 850 000 people which is predicted to increase to over two million people by 2051 (www.alzheimers.org.uk/dementia2014 accessed Oct 2014). Alzheimer's disease (AD), the most common cause of progressive dementia in the elderly, is responsible for 62% of these dementia cases (www.alzheimers.org.uk/statistics accessed May 2015). This is a growing concern for all.

1.1.1. What is Alzheimer's disease?

Alzheimer's disease (AD) was first documented in 1907 by the German clinical psychiatrist Alois Alzheimer during his work with Franz Nissl, staining the cerebral cortex of psychotic patients (Alzheimer 1907). AD is a neurodegenerative disorder, characterized by intracellular neurofibrillary tau tangles and extracellular amyloid beta (Aβ) protein plaques that ultimately result in neuronal death in specific regions of the brain.

1.1.2. How does it present?

Initial symptoms of this insidious disease include memory loss, loss of names, difficulty remembering specific words and forgetting recent events and appointments. Progression is slow, usually over about 10 years. Later symptoms include general confusion, language difficulties, apraxia and difficulty making plans and decisions. Symptoms of the final stages include wandering, disorientation, apathy, appetite changes, psychiatric and behavioural problems and incontinence. Importantly perhaps, not all forms of Alzheimer's disease will present with memory loss as the initial symptom (McKhann et al. 2011). It is now recognized that the disease exhibits a range of phenotypes (Alladi et al. 2007). The late onset sporadic form of Alzheimer's disease (LOAD) is the most common.

1.2. Neuropathology and Neurochemical changes

At the macroscopic level, Alzheimer patients show regional brain atrophy and subsequent enlargement of the ventricles, with many subcortical structures being lost. Areas particularly affected are the hippocampus and the temporal horn of the lateral ventricle (Perl 2010). Severe neuronal loss can occur in

regions such as the entorhinal cortex in the medial temporal lobe (Gómez-Isla et al. 1996). However some regions, like the superior temporal gyrus and the visual cortex, may escape complete degeneration (Halliday et al. 2003). This loss of brain tissue correlates with the loss of nerves and synapses associated with the condition. At the cellular level, there is a general loss of spine density and dendritic complexity (Arendt 2009).

In addition to widespread neuron loss, diseased brain tissue shows two key neuropathology changes at the microscopic level; the formation of amyloid plaques and the development of neurofibrillary tau tangles (NFT). Additional AD associated changes include cerebral amyloid angiopathy (CAA), region specific neuron death and glial activation. Other neuropathology changes that may be found include Lewy body disease (LBD), frontotemporal dementia (FTD) and vascular brain injury (Schellenberg and Montine 2012).

In the brain parenchyma, extracellular amyloid plaques result from the deposition of the amyloid-β (Aβ) peptide and the accumulation of neuronal processes around the peptides. In the eighties, amyloid plaques were shown to be associated with an inflammatory process. Plaques also contain complement factors due to activation of the complement cascade, and activated microglia (reviewed in Eikelenboom et al. 2006; Heneka et al. 2015).

Aβ is a 4.2 kDa peptide of 40-42 amino acids. The peptide is created from the cleavage of a larger single-pass transmembrane protein, the amyloid precursor protein (APP). This cleavage is catalyzed by β- and γ-secretases. While α-secretase also cleaves APP, the cleavage site is within the Aβ sequence and so does not generate Aβ (O'Brien and Wong 2011). The active catalytic subunit of the γ-secretase is coded by the *PSEN1* or *PSEN2* gene. Although APP is known to form Aβ, and is found in abundance in neurons, the endogenous function of this protein and the cleaved Aβ peptide is as yet unknown (O'Brien and Wong 2011). In transfected cell lines the soluble ectodomain, which is released by cleavage, regulates cell survival, growth and motility, and using mouse models, has been shown to be involved in neuronal migration during development (Young-Pearse et al. 2007).

NFTs are composed of pairs of abnormal hyper-phosphorylated tau proteins that occur within affected neurons. Found in the limbic regions and brainstem

in early disease, NFTs may later spread to other brain areas such as the cortex and subcortex (Hyman et al. 2012). Criteria proposed by Braak and Braak (1991) distinguish six stages of AD disease progression based on the amount and locations of the NFTs. The stages are; no NFTs, Braak stages I/II with NFTs in entorhinal cortex and nearby tissues, stages III/IV where NFTs are mainly found in the hippocampus and amygdala, with slight extension into the association cortex and lastly stages V/VI where NFTs are found throughout the neocortex. NFTs are not unique to Alzheimer's disease and occur in a range of other age-related neurodegenerative diseases collectively known as tauopathies.

It is still debated as to whether amyloid plaques or NFTs are the first to form. One hypothesis is that Aβ triggers tau hyper-phosphorylation, however it is also possible that both processes occur at the same time, simultaneously contributing to damage (Giacobini and Gold 2013).

1.3. Diagnosis

Alzheimer's disease can only be definitively diagnosed at post-mortem; however diagnostic criteria can be used to give a probable diagnosis. As of November 2013 the diagnostic rates on average in England are 48% (www.gov.uk/government/publications/dementia-care-and-support), so there is clearly a need for improvement. The National Institute for Health and Care Excellence (NICE) guidelines recommend using the diagnostic criteria outlined by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA) first published 1984 (McKhann et al. 1984). The research diagnostic criteria for late onset Alzheimer's disease (LOAD) from the International Working Group (IWG) for the diagnosis of Alzheimer's disease were revised twice, first proposing the inclusion of biomarkers in the diagnosis in 2007 (Dubois et al. 2007) and later in 2010 the inclusion of atypical LOAD presentation and the identification of asymptomatic patients who test positive for LOAD biomarkers (Dubois et al. 2010).

Subsequently, the NINCDS/ADRDA guidelines were revised in 2011 to bring the diagnostic criteria in line with research findings and to integrate biomarker evidence (including positron emission tomography (PET) imaging and cerebrospinal fluid (CSF) assays) into the diagnosis (Jack et al. 2011;

McKhann et al. 2011). LOAD is now thought of as a continuum of pathophysiological disease, broadly grouped into three phases. The preclinical or asymptomatic phase (Sperling et al. 2011), the prodromal symptomatic or mild cognitive impairment (MCI) phase (Albert et al. 2011) and the final, fully symptomatic phase recognized as Alzheimer's disease dementia. The revised criteria allow for earlier intervention, at the prodromal phase, with the diagnostic criteria for the preclinical phase recommended predominantly for research purposes into the secondary prevention of AD in the preclinical phase (Sperling et al. 2011; Morris et al. 2014).

These guidelines are again under revision following further research suggesting that the range of dementias included under the diagnostic criterion of AD is larger than initially thought and can include more pathological changes than the two hallmarks of amyloid plaques and neurofibrillary tau tangles (Reitz and Mayeux 2014). The most recent IWG recommendations, published in 2014 (Dubois et al. 2014), suggests that the diagnosis can be simplified to the presentation of Alzheimer's pathology as shown by an appropriate biomarker coupled with typical or atypical clinical AD phenotype.

The mini mental state examination (MMSE) may be initially used to assess the extent of memory loss or cognitive decline. The MMSE is a short test of cognitive function, often used as an initial screen for diagnosis. This is not definitive but may be useful as a tool to rule out the diagnosis of dementia and determine if further investigation is required (Wind et al. 1997; Mitchell 2009). Different dementias may present with different MMSE results further confounding diagnosis. The 2007 IGW revision recommended the use of a list based memory recall test to confirm typical presentation of LOAD (Dubois et al. 2007). Typical LOAD usually presents with an amnestic syndrome of the hippocampus.

The most specific biomarkers, when taken in combination with Alzheimer's pathology from post-mortem studies, are cerebral spinal fluid (CSF) biomarkers and amyloid photon emission tomography (PET). CSF biomarkers include $A\beta_{1\text{-}42}$ which shows an inverse relationship to the amount of amyloid deposited in the brain and tau in the forms of T-tau (total tau), that shows the state of neuronal degeneration and P-tau (phosphorylated tau), directly reflecting the pathology of the tau tangles (Blennow et al. 2010). The different

ratios of these three biomarkers are also useful for predicting progression of LOAD (Dubois et al. 2014).

LOAD progression can also be predicted with imaging. Using different amyloid-specific tracers for PET scanning reveals the amount of fibrillar amyloid deposition *in vivo* (Herholz and Ebmeier 2011). The most commonly used is ^{11}C-PiB (Pittsburgh compound B). Others include florbetapir (AV-45), flutemetamol (18F-PiB derivative), florbetaben (AV-1) and AZD4694. Levels of fibrillar amyloid as predicted by PET correlate strongly with post-mortem plaque load and provide good predictions on MCI progression to AD dementia, making amyloid-PET a useful surrogate marker of amyloid pathology in the brain (Clark 2011). Imaging can also be useful to distinguish the different dementias. For example, hexamethylpropyleneamine oxime (HMPAO) single-photon emission computed tomography (SPECT) can be used to distinguish AD from vascular or frontotemporal dementia although not if the patient has Down's syndrome (McNeill et al. 2007).

Boundaries between different dementia diagnoses are often blurred, and mixed diagnoses are common, particularly for AD with vascular disease (VD) and AD with Lewy body disease (LBD) as confirmed by clinical-pathological studies. Indeed it is rare to come across "pure" VD (Neuropathology Group. Medical Research Council Cognitive Function and Aging Study. 2001; Jellinger 2006). Further complicating post-mortem pathological diagnoses are Alzheimer-like brain changes, or Alzheimer pathology, in individuals who did not exhibit dementia symptoms during life. This is most likely due to the progressive nature of the disease, where physical manifestations could begin 20-30 years before behavioural and psychological symptoms present (Holtzman et al. 2011).

1.4. Management and social consequences

There is, as yet, no cure for AD. Therefore, treatment and management is supportive, with the aim of maintaining dignity and providing care so the individual can remain in the home environment for as long as possible. Two types of pharmacological interventions are recommended that are based on the observed decrease in acetylcholine and glutamate neurotransmitter levels which are associated with disease progression along with loss of neurons and cortical atrophy.

Stopping acetylcholine from being broken down and increasing the amount of acetylcholine in the brain can alleviate early symptoms (Tabet 2006). The National Institute of Clinical Excellence recommends the use of acetylcholinesterase inhibitors, donepezil, rivastigmine and galantamine, for mild to moderate cases of Alzheimer's disease while for moderate to severe Alzheimer's disease, the N-methyl D-aspartate (NMDA) receptor antagonist, memantine might be prescribed. Memantine is less specific, blocking glutamate transmission through the NMDA receptor channels throughout the brain (Johnson and Kotermanski 2006).

With the prevalence of LOAD predicted to more than triple in the next 40 years (Wortmann 2012), LOAD is set to become a major public health concern with considerable social and economic impact. Worldwide, the largest increase in dementia cases is predicted to occur in low to middle income countries (Wortmann 2012). The numbers of dementia cases are reaching a plateau in Western Europe, despite the increase in the aging population (Wu et al. 2015). This is presumably due to the population experiencing better health during early and mid-life periods. Dementia costs the world 604 billion US dollars per year (Wortmann 2012). The total cost of dementia to the UK economy today is estimated to be £26.3 billion, £11.6 billion of which (44%) is covered by unpaid carers (Prince et al. 2014).

1.4.1. Drug development

A large amount of resources and effort has gone into drug development for Alzheimer's disease reflecting the immense clinical need and large potential world market. As already mentioned, the first drugs developed for Alzheimer's disease focused on addressing the cholinergic deficit noted in the nucleus basalis of Meynert (Bowen et al. 1976; Davies and Maloney 1976). However since the production of the cholinesterase inhibitors and the NMDA antagonist, memantine, all further drug development primarily focused on disrupting the amyloid cascade hypothesis (ACH). The ACH was proposed by John Hardy and David Allsop in 1991 (Hardy and Allsop 1991) and suggests that it is the overproduction of Aβ which is ultimately responsible for the disorder (see section 1.5.2.1). Unfortunately, despite valiant efforts trialing drugs inhibiting Aβ aggregation (such as tramiprosate and scyllo-inositol), Aβ antibodies (bapineuzumab and solanezumab) and γ-secretase inhibitors (semagacestat and avagacestat) no drug target has yet been validated which

has been developed based on the ACH (Schneider et al. 2014). It is possible that this is because the study participants need to be recruited at the earlier MCI or even preclinical stages of disease progression. For example, data presented at the Alzheimer's Association International Conference in July 2015 indicates that solanezumab could slow the progression of the disease in mild AD individuals (Liu-Seifert et al. 2015). The results of the next phase of this study will be crucial for the validation of this Aβ antibody as a possible treatment for AD.

1.5. Aetiology

The specific cause(s) of AD are as yet unknown, and the main risk factor for the disease is increased age, as after 65 years of age prevalence doubles every 5 years. There are two main forms of the disease, the common late onset form of Alzheimer's disease (LOAD) and the rarer, familial early onset Alzheimer's disease (EOFAD). LOAD is a complex disorder, with both environmental and genetic risk factors.

1.5.1. Environment

The main environmental risk factor for LOAD remains age. Other risk factors that cannot be modified include ethnicity (Kalaria et al. 2008) and early developmental conditions (Borenstein et al. 2005). There is also some evidence that head injury during early life can increase risk of developing Alzheimer's disease (Fleminger et al. 2003). Socio-economic factors also play a part, as environmental enrichment, cognitive reserve and a diet high in anti-oxidants can be protective, but may not be accessible to all. Increased education as a marker for cognitive reserve has been found to increase the threshold for the clinical manifestation of brain atrophy among LOAD patients (Liu et al. 2012). The cognitive reserve theory posits that individuals with higher experience resources such as knowledge or education show higher levels of cognitive function as they age (see review Meng and D'Arcy 2012). Environmental risk factors that can be addressed include obesity and related conditions such as diabetes and hyperlipidemia and cardiovascular diseases including hypertension (Ertekin-Taner 2007).

1.5.2. Genetics

Early work on the genetics of AD focused on Down's syndrome patients, who often develop a disease similar to AD after 40 years of age, and families with early disease onset AD (<65 years) who also show increased risk of disease

development (Harris 1982). Early onset familial AD (EOFAD) has been documented as early as the 1930s and is rare with an autosomal dominant inheritance, however it was only after advances in molecular techniques that the gene locus was found on chromosome 21, and finally the gene itself was identified as *APP* (amyloid precursor protein) (Kang et al.; Goate et al. 1991). Mutations causing EOFAD are mainly clustered within a 54 amino acid segment, around or in the sequence of *APP* that codes for the Aβ peptide (see http://www.molgen.ua.ac.be/ADmutations). The different *APP* mutations discovered have helped reveal the molecular pathology of the disease.

For example, the Swedish mutation is a double substitution in the sequence before the Aβ peptide sequence that results in about three times more Aβ being produced from the cleavage of APP, possibly as a result of increasing β-secretase cleavage efficiency (Citron et al. 1992). This, together with the knowledge that trisomy 21 (Wisniewski et al. 1985), or even small duplications of chromosome 21 that duplicate the *APP* locus can cause AD changes (Rovelet-Lecrux et al. 2006), suggested that increased Aβ peptide alone is enough to cause AD. Other mutations showed that it was the ratio of Aβ42 to Aβ40 peptides that was important which led to the discovery that Aβ42 is more amyloidogenic and prone to aggregation (Jarrett et al. 1993).

The other two genes responsible for EOFAD are the presenilin genes, *PSEN1* and *PSEN2*, although these are even rarer than the *APP* mutations. PSEN1 and PSEN2 proteins form part of the γ-secretase complex and disease-causing mutations change the cleavage site of the complex, producing more Aβ42 after cleavage (Wakabayashi and De Strooper 2008). To date, the Alzheimer's disease and Frontotemporal Dementia Mutation Database (http://www.molgen.ua.ac.be/admutations/, accessed November 2015) lists 26 pathogenic mutations in *APP*, 151 in *PSEN1* and 18 in *PSEN2*.

1.5.3. Amyloid cascade hypothesis

For a long time, the central disease hypothesis for Alzheimer's disease was the amyloid cascade hypothesis (ACH) (Selkoe 1991; Hardy and Higgins 1992) which suggests that dysfunction in APP cleavage results in the overproduction of Aβ and it is the Aβ which is primarily responsible for the disease.

APP is sequentially cleaved by two different pathways – the amyloidogenic and the nonamyloidogenic pathways with the amyloidogenic pathway resulting in the production of Aβ (O'Brien and Wong 2011) (figure 1.1). The ACH theorizes that through the course of AD, the amyloidogenic pathway becomes the main cleavage pathway for the protein, resulting in the accumulation and aggregation of Aβ. This hypothesis was proposed following the genetic findings from familial early onset AD which implicated the processing of APP and the generation of Aβ in AD.

Downstream effects of the ACH include tau accumulation which together with Aβ accumulation lead to synaptic failure, neuronal loss and cognitive decline. An alternative model suggests that amyloid is the catalyst initiating tau neurofibrillary pathology in the temporal lobe which leads to tau accumulation in the cortex, synaptic failure and neuronal loss, but also directly contributing to cognitive decline (Sorg and Grothe 2015).

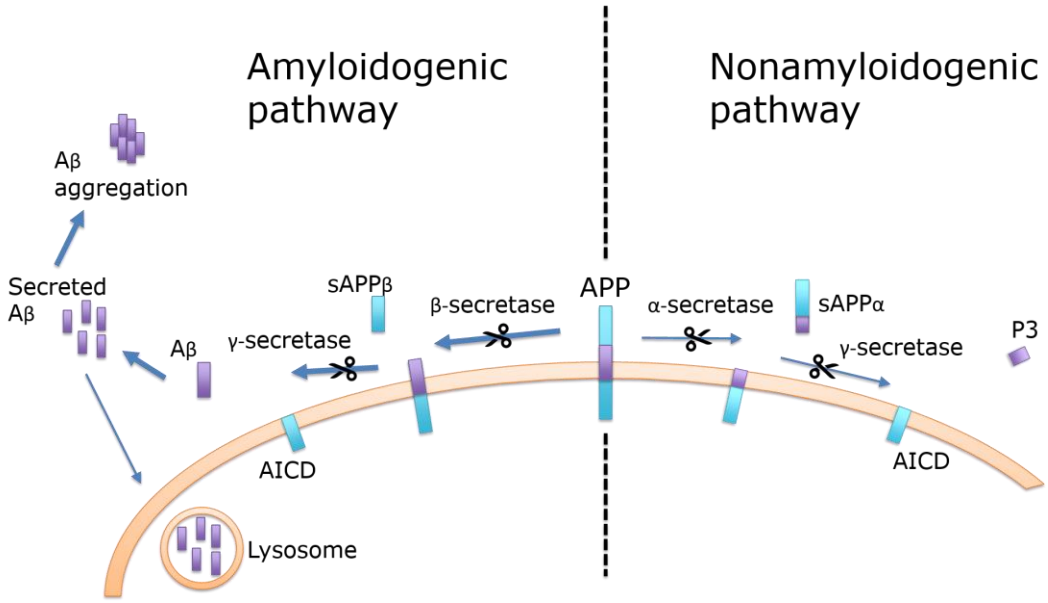

Figure 1.1. The amyloidogenic and nonamyloidogenic pathways for cleaving APP. The cell membrane is shown in orange, the Aβ section of APP is shown in purple and enzymatic cleavage is indicated with the scissors icon. γ-secretase is involved in both pathways, while β-secretase initially cleaves APP in the amyloidogenic pathway and α-secretase initially cleaves APP in the nonamyloidogenic pathway. Both pathways generate the amyloid precursor protein intracellular domain (AICD). Bold blue arrows indicate the pathway leading to Aβ aggregation and plaque formation. Figure adapted from (Yu et al. 2014).

There is a strong documented correlation between dementia and the accumulation of abnormal protein deposits including APP, however, the ACH appears to oversimplify the disease. AD has a complex biology, biochemistry and pattern of neurodegeneration which is not easily explained by the ACH.

Therefore, this hypothesis has been questioned. In part due to Alzheimer pathology in non-demented individuals and in part due to the short comings of therapeutic interventions targeting Aβ (Hardy and Selkoe 2002; Hardy 2009) (see also section 1.4.1). There is now sufficient evidence to suggest that this hypothesis could possibly be disregarded entirely (Herrup 2015).

1.6. Genetics of late onset Alzheimer's disease

The genetics behind the common form of AD, late onset Alzheimer's disease (LOAD) are rather more complicated than the genetics of the early familial onset form of the disease. Given that this thesis focuses on genes associated with the late onset form of the disease, any mention of AD from this point refers to the late onset form of the disease.

1.6.1. APOE

The *APOE* (Apolipoprotein E) locus on chromosome 19 was first identified as having an association with AD in 1991 through a family study (Pericak-Vance et al. 1991). There are three *APOE* alleles in humans, ε2, ε3 and ε4 that all have different associations with AD (Strittmatter et al. 1993). The three alleles differ at two SNPs (single nucleotide polymorphisms) in exon 4 that result in different amino acids being incorporated into the protein at position 112 and 158 (Table 1.1). The differences in amino acids result in significant structural differences in the APOE protein produced from each isoform with conformational changes occurring in the lipid binding region of the protein (Yu et al. 2014).

Table 1.1. The different amino acid composition of the three *APOE* epsilon alleles. The amino acid is abbreviated using the three letter code.

***APOE* allele**	**Position 112**	**Position 158**
ε2	Cys	Cys
ε3	Cys	Arg
ε4	Arg	Arg

The ε2 allele is rare and protective against AD (Chartier-Harlin et al. 1994). This allele is associated with delayed age of onset (by up to 5 years) in individuals that do go on to develop AD. The common, neutral allele is ε3, while the ancestral allele and the one associated with a high risk of developing AD is ε4 (Corder et al. 1993). The effect of carrying one ε4 allele changes the disease risk curve to be 5 years earlier and with two ε4 alleles to 10 years earlier (Yu et al. 2014). Individuals homozygous for ε4 are more than eight times as likely to be affected by AD (Corder et al. 1993).

Produced by neuroglial cells, APOE is the main lipid transport protein in the central nervous system. It appears to be secreted as part of the high-density lipoprotein (HDL) like particles which also contain phospholipids and cholesterol. It is also found in the cerebrospinal fluid (CSF), in particles with a cholesterol ester core (LaDu et al. 1998). These lipoproteins with APOE are involved in cholesterol, lipid and reverse cholesterol transport although their function in cholesterol homeostasis is not yet defined (Holtzman et al. 2012).

The different isoforms appear to play a role in amyloid-β aggregation, with the ε4 allele associating with increased Aβ, Aβ oligomers and the accumulation of plaques in the brain (Christensen et al. 2010; Hashimoto et al. 2012; Koffie et al. 2012). However the exact role of *APOE* ε4 in the pathogenesis of Alzheimer's disease is still unknown (Kok et al. 2009). APOE protein is the most well characterized Aβ chaperone (Bu 2009). Precisely how each APOE isoform affects the clearance of Aβ is not known, however there are several theories including; Aβ clearance via microglia (Lee et al. 2012), astrocytes (Basak et al. 2012) and neurons (Li et al. 2012), the blood-brain barrier (Bachmeier et al. 2013), enzymatic degradation (Jiang et al. 2008), oligomer formation (Hashimoto et al. 2012) and Aβ clearance via drainage through interstitial fluid (Castellano et al. 2011) or perivasculature (Hawkes et al. 2012). APOE appears to influence the regional vulnerability of neurons (Xu et al. 1999). It is also thought that the different APOE isoforms can affect tau changes in AD brains *in vitro* (Yu et al. 2014). The APOE ε3 isoform inhibits tau hyperphosphorylation and destabilization of the neuronal cytoskeleton in AD (Strittmatter et al. 1994) while ε4 is neurotoxic due to the C-terminal truncation and it stimulates tau phosphorylation initiating neurofibrillary tangles (Harris et al. 2003).

APOE remains the strongest risk locus identified to date (Schellenberg and Montine 2012). Any genome wide association analyses need to control for the APOE region to allow other alleles with smaller odds ratios (ORs) to be detected.

1.6.2. Genome-wide association studies: uncovering AD genetics.

Genome wide association studies (GWAS) revolutionized the discovery of disease causing genes in common complex disorders. GWAS is an unbiased

approach for identifying common variants which associate with a particular disease or phenotype. Using linkage disequilibrium, marker or tag SNPs are identified which represent nearly all common variants in the genome. These tag SNPs are then compared between cases and controls to identify common variants which associate with disease. The theory behind GWAS being that genetic risk for common complex diseases will be attributed to allelic variants found in a high proportion of the population - the common disease/common variant (CDCV) hypothesis (Lander 1996; Reich and Lander 2001).

Initial GWAS investigating late onset AD were underpowered and, apart from the already known gene *APOE*, were unable to reproducibly find any additional candidate genes (Coon et al. 2007; Grupe et al. 2007; Reiman et al. 2007; Abraham et al. 2008; Bertram et al. 2008; Li et al. 2008; Beecham et al. 2009; Carrasquillo et al. 2009). However, these studies did reveal that *APOE* was clearly detected as the major risk allele and that there were no other risk alleles of similarly large effect size. Later GWAS with increased sample size and therefore increased power were able to identify, and successfully replicate, new genes associated with AD.

1.6.3. The first risk genes reproducibly identified through GWAS

In 2009 the first sufficiently powered GWAS were published and the first of the small effect risk loci were revealed; clusterin (*CLU*), phosphatidylinositol binding clathrin assembly protein (*PICALM*) and complement receptor type 1 (*CR1*) (Harold et al. 2009; Lambert et al. 2009). The Genetics and Environmental Risk in Alzheimer's Disease (GERAD) consortium (Harold et al. 2009) uncovered two novel SNPs, in *CLU* ($p = 1.4x10^{-9}$, OR= 0.84, CI = 0.79-0.89) and in *PICALM* ($p = 1.9x10^{-8}$, OR = 0.85, CI = 0.80=0.90) which used a two stage approach to reach the threshold for genome-wide significance.

This two stage GWAS approach was also used by the European Alzheimer Disease Initiative (EADI) (Lambert et al. 2009) using an independent set of samples. The EADI study also discovered a significant association in the combined stage one and two analysis between *CLU* (combined results, $p = 7.5x10^{-9}$, OR = 0.86, CI = 0.81-0.90) and *CR1* ($p = 3.5x10^{-9}$, OR = 1.21, CI = 1.14-1.29) and the study reported supporting evidence for the *PICALM* site identified by GERAD (Harold et al. 2009). These results have been consistently independently replicated (Jun et al. 2010; Wijsman et al. 2011).

The fourth small effect risk gene to be uncovered was bridging integrator protein-1 (*BIN1*) ($p = 1.59x10^{-11}$, OR = 1.15, CI = 1.11-1.20) (Seshadri et al. 2010). The study discovering this gene also used a staged approach to the analysis, with SNPs selected for stage 2 from SNPs in the first stage with p values less than 10^{-3} and SNPs being selected for the third final stage from those SNPs in stage 2 with p values less than 10^{-5} (Seshadri et al. 2010).

The next five genes were discovered in two studies in 2011. The three-stage design was again employed by Hollingworth et al. 2011 to analyse the findings of the CHARGE (Cohort for Heart and Aging Research in Genomic Epidemiology), GERAD (Genetic and Environmental Risk in Alzheimer's Disease consortium) and EADI (European Alzheimer Disease Initiative group) consortia. Stage 1 used 13685 controls and 6688 cases from the GERAD group to identify 61 SNPs with genome wide significance ($p <= 1x10^{-5}$) which were carried through to the second stage for testing with the GERAD2 dataset, excluding SNPs which had been previously tested in this dataset, and including an *in silico* replication dataset created from imputation of deCODE and AD-IG (German Alzheimer's Disease Integrated Genome Research Network GWAS dataset) totaling 4896 cases and 4903 controls. Novel SNPs which continued to show association at stage 2 were taken to stage 3 and tested for association in 8286 cases and 21258 controls, which included EADI2 and Mayo2 samples and *in silico* replication samples from CHARGE. The study found genome-wide significant results for *ABCA7* ($p = 5.0x10^{-21}$, OR = 1.23, CI = 1.17-1.28), the *MS4A* locus (*MS4A6A/MS4A4E)* ($p = 1.2x10^{-16}$, OR = 0.91, CI = 0.88-0.93) and found independent evidence for the loci identified in the ADGC study (Naj et al. 2011), showing genome wide significance at the meta-analysis (third stage) level for *CD2AP* ($p = 8.6x10^{-9}$, OR = 1.1, CI = 1.07-1.15) *CD33* ($p = 1.6x10^{-9}$, OR = 0.91, CI = 0.88-0.93) and *EPHA1* ($p=6.0x10^{-10}$, OR = 0.90, CI = 0.86-0.93) (Hollingworth et al. 2011).

A second study, published in the same journal issue as Hollingworth et al. 2011 confirmed these findings. The National Institute on Aging (NIA) AD Genetics Consortium (ADGC) also used a three stage design (Naj et al. 2011). Again, only SNPs with a p value less than a cut off of 10^{-5} were taken further to stage two, with only five loci with suggestive or genome wide significance being taken forwards to the third stage. This study confirmed the

association of *APOE, CR1, BIN1, PICALM* and *CLU* and identified four new loci which remained significant through all stages, namely *CD2AP* ($p = 8.6x10^{-9}$, OR = 1.11, CI 1.07-1.15)*, EPHA1* ($p = 6.0x10^{-10-}$, OR =0.90, CI 0.86-0.93)*, MS4A4A* ($p = 8.2 x10^{-12}$ OR = 0.91, CI 0.88-0.93) and *CD33* ($p = 1.6x10^{-9}$, OR = 0.91, CI 0.88-0.93) and a suggestive result for *ABCA7* ($p = 5.0x10^{-7}$, OR = 1.15, CI 1.09-1.21) (Naj et al. 2011).

The genes identified in 2011 implicated three new pathways in LOAD pathogenesis, namely immunity (innate and adaptive), cholesterol metabolism and cell membrane processes (including synaptic dysfunction) (Morgan 2011). Although a large proportion of genetic risk factors remained to be found (the so-called 'missing heritability'), since collectively these genes along with *APOE*, accounted for only an estimated 23% of the heritability of LOAD (So et al. 2011).

In 2013, Lambert et al. published a meta-analysis of over 74000 individuals and announced the discovery of 11 new susceptibility loci for LOAD in addition to the nine loci already identified in 2011. This analysis was the culmination of the International Genomics of Alzheimer's Project (IGAP). The meta-analysis was again performed in two stages, with the first stage using the combined imputed genotype data from 4 previously published GWAS datasets with 17008 LOAD cases and 37154 controls. For the replication dataset, SNPs which were moderately associated ($p <1x10^{-3}$) were taken forwards for genotyping in an additional 8572 LOAD cases and 11312 controls (Lambert et al. 2013). The new loci identified were *HLA-DRB5-HLA-DRB1* ($p = 2.9x10^{-12}$, OR = 1.11, CI 1.08-1.15), *PTK2B* ($p = 7.4x10^{-14}$, OR = 1.10, CI 1.08-1.13), *SORL1* ($p = 9.7x10^{-15}$, OR = 0.77, CI 0.72-0.82), *SLC24A4-RIN3* ($p = 5.5x10^{-9}$, OR = 0.91, CI 0.88-0.94), *INPP5D* ($p = 3.2x10^{-8}$, OR = 1.08, CI 1.05-1.11), *MEF2C* ($p = 3.2x10^{-8}$, OR = 0.93, CI 0.90-0.95), *NME8* ($p = 4.8x10^{-9}$, OR = 0.93, CI 0.90-0.95), *ZCWPW1* ($p = 5.6x10^{-10}$, OR = 0.91, CI 0.89-0.94), *CELF1* ($p = 1.1x10^{-8}$, OR = 1.08, CI 1.05-1.11), *FERMT2* ($p = 7.9x10^{-9}$, OR = 1.14, CI 1.09-1.19) and *CASS4* ($p = 2.5x10^{-8}$, OR = 0.88, CI 0.84-0.92).

In terms of the pathways in which the 11 new genes are found, they appear to group according to previously identified pathways such as immune response (*HLA-DRB5-HLA-DRB1*, *INPP5D*, *MEF2C*), APP processing (*SORL1*,

CASS4), Tau pathology (*CASS4*, *FERMT2*), cell migration (*PTK2B*) and lipid transport and endocytosis (*SORL1*). However, there were a few additional pathways suggested by the new genes. *MEF2C* is also involved in hippocampal synaptic function whilst *PTK2B* associates with long term potentiation (LTP) in the hippocampus CA1 region. This gene, *PTK2B*, is also interesting as it is found approximately 130kb from *CLU,* with the genes being separated by a recombination hot spot. *CELF1*, *NME8* and *CASS4* are also implicated in cytoskeletal function and axonal transport.

1.7. Searching for causal variants

Although the genes above have been identified in GWAS as being associated with AD, the functional variants responsible for the association have yet to be discovered. This is because the tag SNPs used to test for association in GWAS were selected because they represent most of the genetic variation found within a particular linkage disequilibrium (LD) block. The association of one tag SNP therefore represents the association of a particular LD block. The true disease-causing variant could be any of the variants within that LD block. A known caveat of GWAS is the subsequent difficulty identifying and definitively proving the disease-causing allele responsible for generating the disease association at a specific GWAS locus (MacArthur et al. 2014). It is also important to bear in mind that the GWAS locus is often referred to by the name of the closest coding gene. In many cases this gene may be located several thousand bases away from the GWAS associated tag SNP or 'hit' (Schaub et al. 2012). As the functional disease-causing variant is not known, studies following up on GWAS hits often use the associated tag SNPs (and the closest coding gene) to investigate AD pathology. Confirming the pathogenic variant would allow more thorough investigations of the molecular mechanisms involved in disease pathogenesis and ensure that the correct gene was being studied. This would ultimately further the goals of complex disease genetics which is to advance the knowledge of disease biology, and identify additional targets for diagnosis and treatment.

In order to fully investigate the GWAS loci, further sequencing of the area is proposed to discover rare causative mutations responsible for the GWAS association signal. This is the synthetic association hypothesis which suggests that common variants revealed through GWAS reflect associations of rarer mutations (Dickson et al. 2010; Wang et al. 2010a). According to the

common disease common variant (CDCV) hypothesis, which formed the basis for GWAS, if common diseases were caused only by common disease alleles with low penetrance (MAF > 5% and OR <1.2) the disease alleles should have been relatively straightforward to identify in a small number of cases and controls. In reality, tens of thousands of samples have been used to date for Alzheimer's disease GWAS (Lambert et al. 2013), and the population attributable risk is still only 60% (Medway and Morgan 2014). This suggests that rare variants with larger odds ratios may be responsible for the missing heritability.

The common disease rare variant hypothesis (CDRV) suggests that it is the synthetic associations of these rare variants with high penetrance which are ultimately responsible for causing disease and which are the real cause of the GWAS associated variants identified in the common complex diseases. The synthetic association hypothesis is not thought to reflect the true genetic architecture for the majority of the GWAS associated variants as expected outcomes for this model are not observed (Anderson et al. 2011; Wray et al. 2011). For example, low allele frequencies for the majority of GWAS hits and large odds ratios for rare variants resulting in increased detection through family studies have not been found (Wray et al. 2011). However, using next generation sequencing to resequence loci identified through common GWAS variants is still a useful method for identifying rare variants (Wray et al. 2011).

Rare variants at two GWAS loci (*CLU* and *ABCA7*) have been found to be associated with the disease independently of the GWAS tag variants (Bettens et al. 2012; Steinberg et al. 2015) and rare variants in the three familial genes, *APP*, *PSEN1* and *PSEN2* are associated with increased disease risk in late onset AD (Cruchaga et al. 2012). Rare coding variants in *TREM2* (Guerreiro et al. 2013),*TREML2* (Benitez et al. 2014) and *SORL1* (Pottier et al. 2012) were identified as increasing Alzheimer's disease risk before the GWAS loci rs9381040 ($p = 6.3 \times 10^{-7}$) (nearby *TREM2* and *TREML2*) and rs11218343 ($p = 9.7x10^{-15}$) (*SORL1*) were identified (Lambert et al. 2013). Therefore investigating rare variation in the other late onset GWAS loci could reveal additional variants affecting disease risk and provide further clarification on the molecular mechanisms leading to disease.

The aim of this PhD project is to use next generation sequencing to deep resequence three of the genetic loci associated with AD in an attempt to uncover rare functional variants. In particular, the last three genes that were uncovered in the combined meta-analyses of 2011 (Hollingworth et al. 2011; Naj et al. 2011), *CD2AP*, *EPHA1* and *CD33.*

1.8. CD2-associated protein *(CD2AP)*

First discovered in 1998 associating with CD2, CD2-associated protein (CD2AP) is thought to be an integral structural component of the immunological synapse (Dustin et al. 1998). An 80kDa cytoplasmic scaffold protein, CD2AP plays a role in actin remodeling, helping to secure proteins to the actin cytoskeleton and is also involved in endocytosis and membrane trafficking. In the immune synapse, the protein helps stabilize and strengthen initial contact (Dustin et al. 1998) and is also involved in stabilizing the synapse between natural killer cells and their target cells thereby playing a part in cytotoxic cell processes (Ma et al. 2010).

The gene on chromosome 6p12 is 149.5kbp long, with 18 exons and lacks a TATA promoter sequence (Lu et al. 2008; He et al. 2011). A proximal promoter was predicted 558bp upstream of the translation start site (TSS) and has been found to contain binding site motifs for several transcription factors (Lu et al. 2008). Transcription factors predicted to bind to this proximal promoter include CREB (cAMP response element-binding protein), with a binding site -553 to -546bp from the TSS and three SP1 binding sites, -490 to -469bp, -486 to -480bp (Lu et al. 2008) and -348 to -340bp (Xu et al. 2011) upstream of the TSS. Recently another binding site for the AP-1 like motif 85 to 78bp upstream of the TSS was also found (Xu et al. 2011). Additional proteins involved in the regulation of this gene include the Lim homeobox transcription factor 1, beta (LMX1B) which has also been found to bind upstream of the start site of the gene, with binding sites at -2855, -1817 and -1170 found to be functional (Miner et al. 2002).

There is only one validated transcript of 5425bp (NM_012120.2) that translates a 639 amino acid protein (NP_036252) which contains three N-terminal SH3 (Src homology-3) domains and an adjacent proline-rich domain (The UniProt Consortium 2012). SH3 domains are often involved in actin cytoskeleton organization and the proline-rich domain may assist SH3 binding.

The C-terminus of the protein contains a coiled-coil domain, which could potentially allow homodimerization or interaction with other membrane proteins (Tossidou et al. 2010). The protein domains of CD2AP show similarities to those of CIN85 (Cbl-interacting protein of 85kDa) a protein involved in receptor signaling and endocytosis (Take et al. 2000).

When it comes to involvement in human disease, CD2AP has largely been investigated for its role in kidney disease due to the expression of this protein in the specialized glomerular capillary cells, podocytes (Shih et al. 2001). Mature podocytes form specialized cell-cell adherens junctions at the basement membrane called the slit diaphragm which excludes proteins from filtrate in the kidneys. CD2AP is thought to be integral in maintaining the integrity of these adherens junctions. The role of CD2AP at the blood brain barrier (BBB) has also been investigated after the protein was found to be highly expressed in cerebrovascular endothelial cells (Lehtonen et al. 2008; Uhlen et al. 2010; Zhang et al. 2014). CD2AP has recently been shown to maintain the integrity of the BBB in mice (Cochran et al. 2015), presumably performing a similar function at the BBB as in the slit diaphragm.

The gene was first implicated in Alzheimer's disease following the large GWAS study by Naj et al 2011 and was subsequently also found by Hollingworth et al 2011. The associated SNP, rs9349407, had an odds ratio of 1.11 (CI = 1.07-1.15, $p=8.6x10^{-9}$, n=48589). A replication study in 2011 in Caucasians failed to confirm the association, although this may have been due to insufficient power (Carrasquillo et al. 2011). No association was also found in replication studies in Han Chinese (Tan et al. 2012) and Korean (Chung et al. 2012) cohorts. In the meta-analysis of 2013 (Lambert et al. 2013), the associated SNP, rs10948363, had an overall odds ratio of 1.10 (CI = 1.07-1.13, $p=5.2x10^{-11}$, n=74046). The two SNPs, rs9349407 and rs10948363, are in complete linkage disequilibrium (D' and $r^2 = 1$) in the 1000 Genomes phase 3 European cohort. In this cohort, rs9349407, a C to G change, is found with a minor allele frequency of 25%, while rs10948363, 34384bp away, is an A to G change with the same minor allele frequency (25%).

Located in the intronic regions of *CD2AP*, both far from an intron-exon boundary (see figure 1.2) and in areas of low sequence conservation in

vertebrates, neither rs9349407 nor rs10948363 are likely to be functional. However, the linkage disequilibrium (LD) block suggests other downstream variants in high linkage with these tag SNPs that may be causative. Both GWAS SNPs are found in a large 165kbp LD block which encompasses the entire *CD2AP* gene (figure 1.2). Examining 1000 Genomes pilot data reveals 68 variants in high linkage disequilibrium with rs9349407 (figure 1.3).

The biological relevance of *CD2AP* in AD pathology remains to be tested. However, the gene is expressed in the brain and seems to maintain BBB integrity which is known to be damaged in Alzheimer's disease (Wisniewski and Kozlowski 1982) and also during human aging (Montagne et al. 2015). Additionally, given the diverse functions of the protein in innate immunity, vesicle trafficking and actin remodeling, it is not difficult to see how this gene complements the new pathways identified as being involved with AD pathology (Morgan 2011). Alternatively, *CD2AP* may contribute to Alzheimer's disease through changing disease risk for other known AD risk factors such as cardiovascular disease, hypertension via renal disease or even through neurovascular damage which is known to be a common comorbidity of AD.

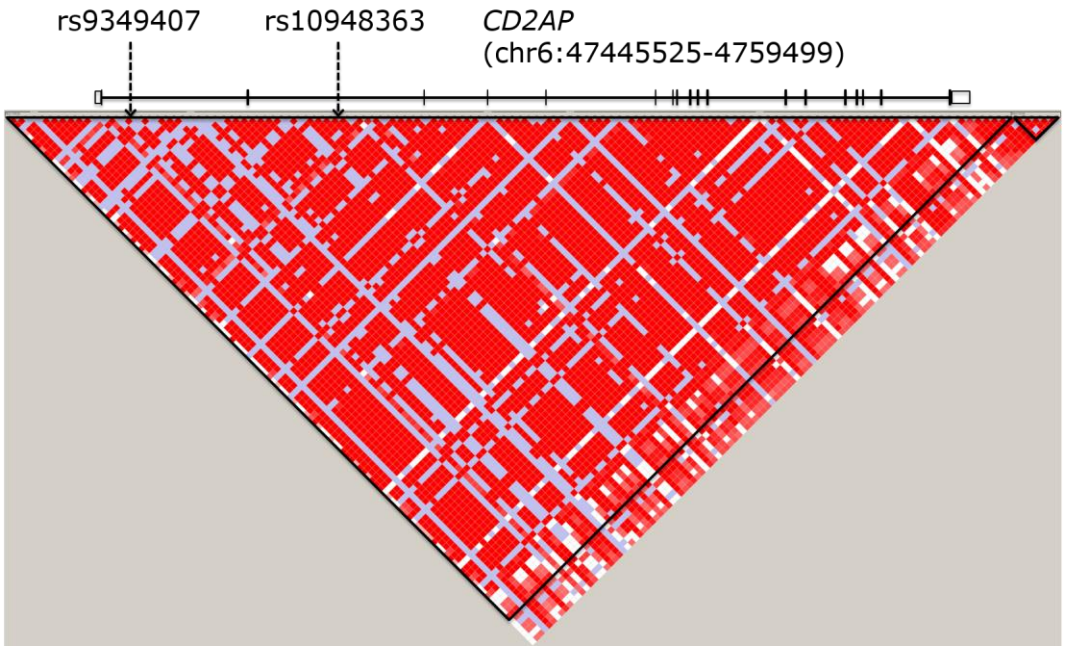

Figure 1.2. Linkage disequilibrium (LD) measured in D' surrounding *CD2AP* and the GWAS SNP, rs9349407. The SNP falls in a 165kbp LD block that spans *CD2AP*. LD plot generated in Haploview using HapMap Data (Rel 28, Phase II + III, Aug '10). LD blocks calculated by Haploview, with the intensity of red indicating the strength of the LD measured by D'.

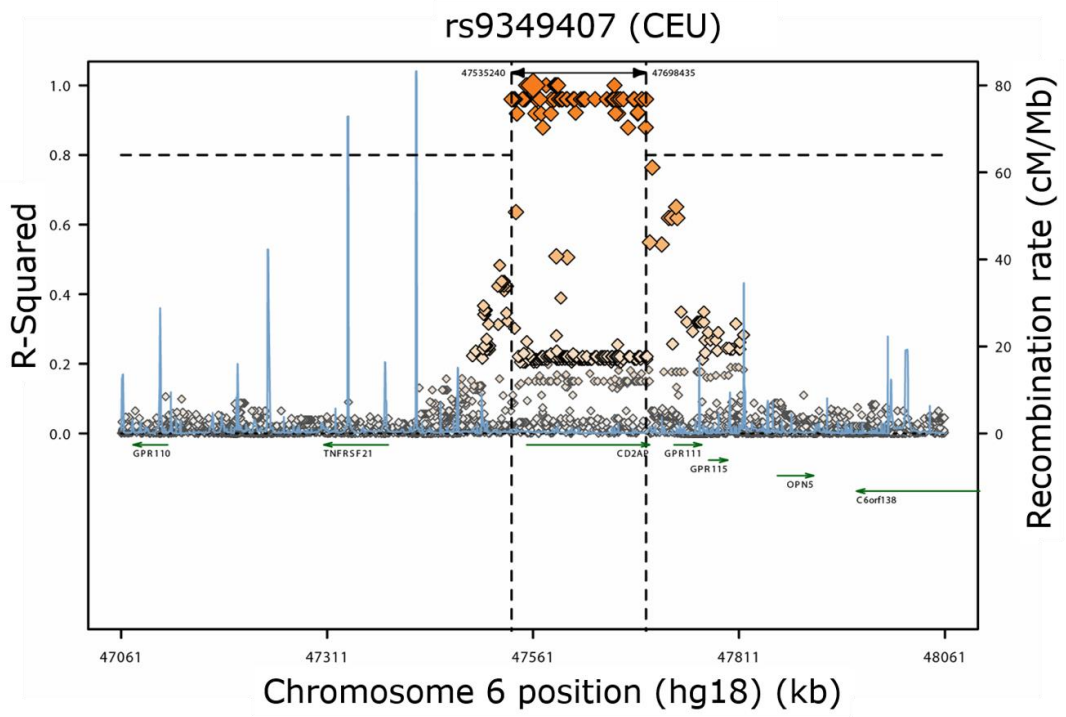

Figure 1.3. Linkage disequilibrium (LD) surrounding the *CD2AP* GWAS SNP rs9349407. Plot shows variants from 1000 Genomes pilot project 1 in LD with the index SNP rs9349407 (shown as the larger orange diamond). The smaller orange diamonds show 68 variants in high linkage (r^2>0.8) with the tag SNP within the 165kbp LD block identified in Figure 1.2. Vertical axes indicate LD ("R-Squared") shown by diamonds and recombination rate (cM/Mb) shown by the blue line graph. Plot generated with SNAP (www.broadinstitute.org/mpg/snap/ldplot.php) (Johnson et al. 2008).

1.9. Erythropoietin-producing human hepatocellular carcinoma (*EPHA1)*

EphA1 is a receptor tyrosine kinase (RTK) first identified in the erythropoietin-producing human hepatocellular carcinoma cell line in 1987 (Hirai et al. 1987). Later found to belong to a large family of 16 ephrin receptors (Eph Nomenclature Committee. 1997), EphA1 is involved in cell adhesion and cellular organization and is expressed primarily in epithelial tissue. However, receptors in the ephrin family are widely expressed in the brain and central nervous system during development and, due to their ability to signal bidirectionally, are thought to play an integral role in synaptic plasticity (Chen et al. 2012; Hruska and Dalva 2012). Ephrin receptors are divided into two classes, A and B according to their extracellular domains and the particular ligands they bind. EphA receptors primarily bind glycophosphatidylinositol (GPI) anchored ephrin-A ligands, however ligand binding across the two classes is promiscuous (Kullander and Klein 2002).

This 17.8kbp reverse transcribed gene is located on chromosome 7q34. Transcriptional regulation of *EPHA1* is not well documented, however, expression appears to be epigenetically modified by a promoter associated CpG island that occurs in exon 1 and intron 1 of the gene (Dong et al. 2009; Herath et al. 2009). *EPHA1* has 18 exons and one validated transcript of 3369bp (NM_005232) translating a 976 amino acid protein (NP_005223). The protein, EphA1, contains an extracellular ephrin-A ligand binding domain followed by a Fibronectin type 3 domain, an internal tyrosine kinase domain and a sterile alpha motif (SAM) domain specific to the EphA subfamily (The UniProt Consortium 2012).

EPHA1 was confirmed as a potential genetic risk factor for Alzheimer's disease in the 2011 GWAS studies (Hollingworth et al. 2011; Naj et al. 2011) through the SNP, rs11767557 ($p = 6x10^{-10}$ OR = 0.9 CI = 0.86-0.93); n = 54359). However, suggestive evidence for this gene was previously reported in 2010 (rs11771145, $p = 1.7 x10^{-6}$) (Seshadri et al. 2010). This association has subsequently been replicated in a cohort of 6835 Caucasians ($p = 5x10^{-4}$) (Carrasquillo et al. 2011). In 2013 the meta-analysis confirmed *EPHA1* association with LOAD with the tag SNP rs11771145, ($p = 1.1x10^{-13}$, OR = 0.90 CI = 0.88-0.93); n = 74046) (Lambert et al. 2013).

The associated SNP, rs11767557 is a T to C change located in an intragenic region 3154bp upstream from *EPHA1* (figure 1.4) in an area of low conservation, so it is unlikely to be functional. There are three other SNPs in high LD with rs11767557, none of which fall in areas of high conservation (figure 1.5). The other associated SNP, rs11771145, is G to A change also located upstream from *EPHA1,* 1623bp further upstream than rs11767557 (see figure 1.4) and also falls in an area without conservation. In the 1000 Genomes phase 3 European cohort the two SNPs are in LD (D' = 0.68, r^2 = 0.23), rs11767557 with a MAF of 21% and rs11771145 with a MAF of 36%. Both SNPs falls in a 46kbp LD block that spans *EPHA1* and two other genes *ZYX* (zyxin) and *EPHA1-AS1* (figure 1.4).

Zyxin is a LIM domain protein involved in actin regulation, particularly at cell junctions (Hirata et al. 2008), therefore it is not unreasonable to consider that the GWAS signal could be reflecting a functional variant in *ZYX*, given the involvement of synaptic dysfunction and cell membrane processes in AD

(Morgan 2011). The other gene, *EPHA1-AS1* (EphA1 antisense RNA 1), is a little known long noncoding RNA with one RefSeq transcript (ENST00000429289). However, both associated SNPs fall in the gene. *EPHA1-AS1* appears to be highly expressed across a range of tissue types (Uhlen et al. 2015). Long noncoding RNAs (lncRNAs) play important regulatory roles and are involved in brain development, aging signaling pathways and synaptic transmission (Qureshi and Mehler 2012). LncRNAs have been implicated in mouse models of Alzheimer's disease, with *BACE1*-AS found to regulate BACE1 (β-secretase) expression (Faghihi et al. 2008) and *Sox2OT* found to be a potential biomarker for LOAD (Arisi et al. 2011). Consequently a role for *EPHA1-AS1* in Alzheimer's disease or related pathways remains to be discovered.

EphA1 has largely been investigated for its part in increasing the invasion capabilities of many cancers (Hafner et al. 2004; Dong et al. 2009). However, could dysfunctional ephrin receptor signaling be responsible for increasing susceptibility to Alzheimer's disease? EphA1 certainly functions on pathways that are known to be involved in AD pathology, including cell adhesion. An increasing body of evidence suggests EphA1 receptor forward signaling inhibits cell adhesion and migration by negative regulation of integrin and the associated integrin-linked kinase (ILK) (Yamazaki et al. 2009; Miao and Wang 2012).

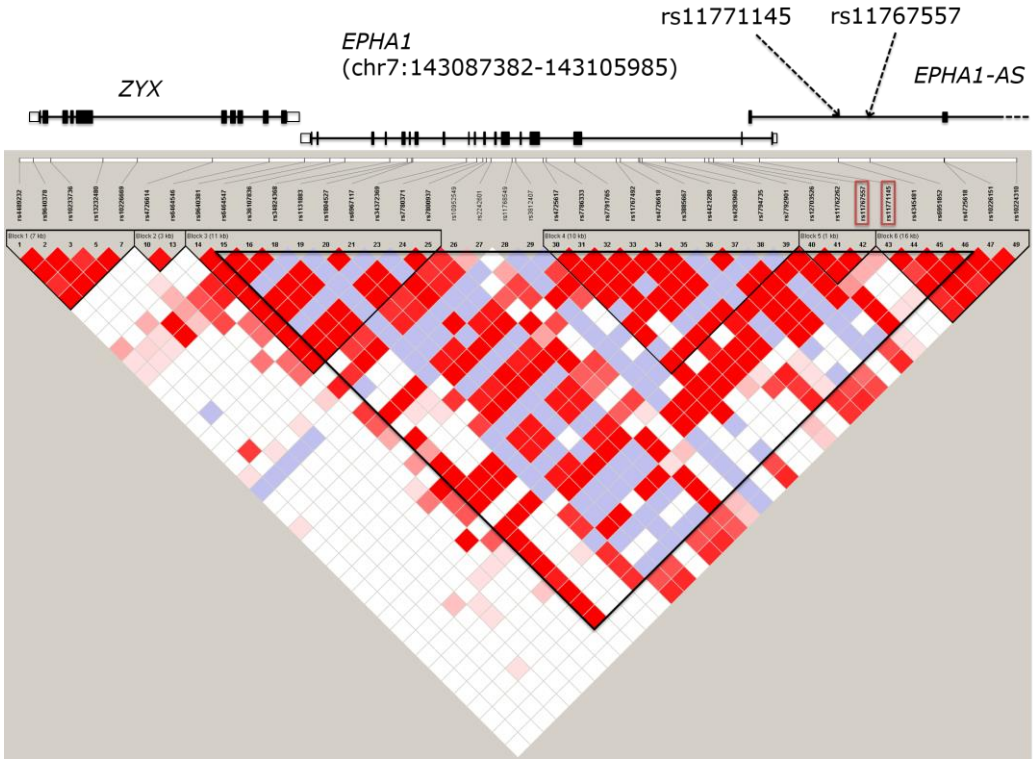

Figure 1.4. Linkage disequilibrium (LD) measured in D' surrounding *EPHA1* and the GWAS SNPs, rs11767557 and rs11771145 (indicated by red rectangles). The GWAS variants falls in a 13kbp LD block which includes *EPHA1* and upstream sequence which is part of the noncoding RNA, *EPHA1-AS*. The variants falls within a larger LD block (46kbp) which encompasses part of *ZYX*. LD plot generated in Haploview using HapMap Data (Rel 28, Phase II + III, Aug '10). LD blocks calculated by Haploview, with the intensity of red indicating the strength of the LD measured by D'.

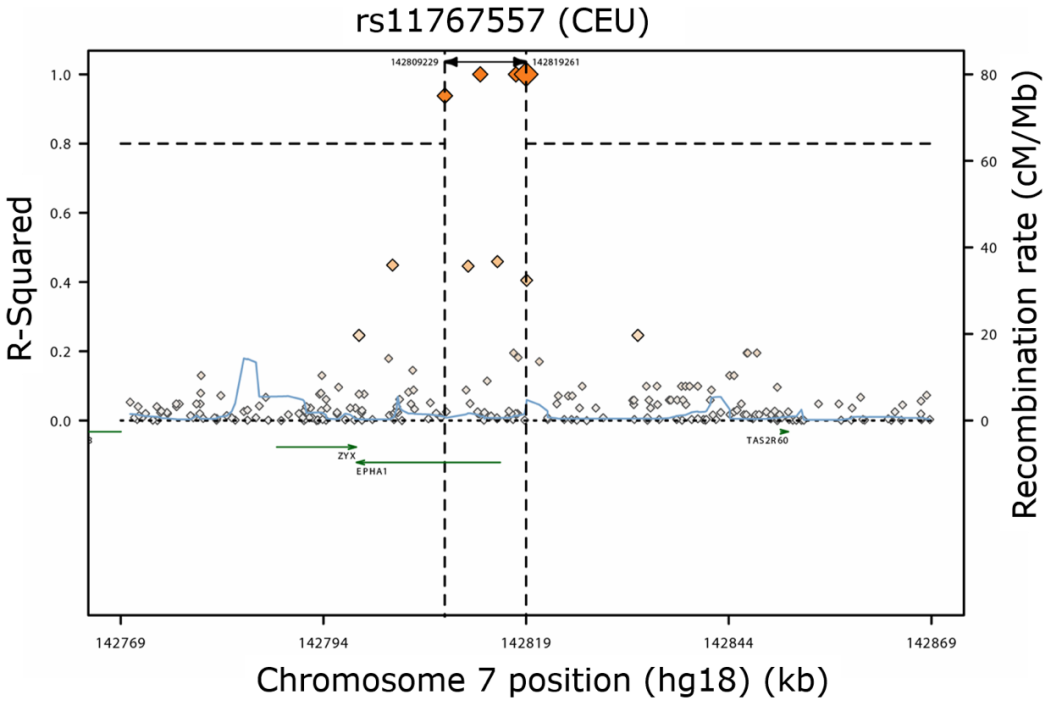

Figure 1.5. Linkage disequilibrium (LD) surrounding the *EPHA1* GWAS SNP rs11767557. Plot shows variants from 1000 Genomes pilot project 1 in LD with the index SNP rs11767557 (shown as the larger orange diamond). The smaller orange diamonds show three variants in high linkage (r^2>0.8) with the tag SNP found in a 10kbp region. Vertical axes indicate LD ("R-Squared") shown by diamonds and recombination rate (cM/Mb) shown by the blue line graph. Plot generated with SNAP (www.broadinstitute.org/mpg/snap/ldplot.php) (Johnson et al. 2008).

1.10. Sialic acid binding immunoglobulin-like lectin-3 (*CD33*)

CD33 or siglec-3 (sialic acid binding immunoglobulin-like lectin-3) is expressed on the cell surface of myeloid immune cells. This 67kDa type I transmembrane receptor was discovered in the 1980s through antibody profiling of myeloid cells (Griffin et al. 1984). Gene expression profiling revealed that this gene is involved in innate immunity and endocytosis (Crocker et al. 2007). In a clinical setting, CD33 is a marker of myeloid leukaemia (Robertson et al. 1992) and has proven to be a useful therapeutic target for pharmacological intervention using gemtuzumab ozogamicin, a humanized antibody to CD33 (Linenberger 2005).

CD33 is a 14.9kbp gene located on chromosome 19, specifically 19q 13.3 (Peiper et al. 1988). It is located within a gene cluster containing related siglecs that was thought to have arisen in an ancient inverse gene duplication (Cao et al. 2009). *CD33* does not have a TATA promoter sequence, however a minimal promoter 198bp upstream of the start codon has been found that contains an SP1 and PU1 site (Bodger and Hart 1998).
The gene is alternatively spliced into three RefSeq validated transcripts. Isoform1 or CD33M (NM_001772.3) is the longest transcript at 1466bp and encodes a 364 amino acid protein (NP_001763.3) containing a signal peptide, an Ig-like V-type domain, an Ig-like C2 type domain, a transmembrane domain and in the C-terminus, two immunoreceptor tyrosine-based inhibitory motifs (ITIMs) (The UniProt Consortium 2012). The shortest transcript, isoform2 or CD33m (NM_001082618.1) is 1085bp as it lacks exon 2a of Isoform1, resulting in a shorter protein of 237 amino acids that is missing the Ig-like V-type sialic binding domain. Isoform3 (NM_0011776008.1) is 1108bp and contains an alternative last exon and 3'UTR compared with isoform1. Isoform3 transcribes a shorter protein at 310 amino acids long.

The protein, siglec-3, belongs to a large family of CD33-related siglecs that all contain amino-terminal V-set immunoglobulin domains facilitating sialic-acid recognition, varying C2-set immunoglobulin domains and ITIMs in the cytoplasmic tail (Crocker et al. 2007). Unique in the immunoglobulin superfamily, siglecs bind sialylated carbohydrates rather than proteins. Possibly as a result of their substrate, the CD33-related siglec super family appears to be rapidly evolving, potentially in response to host changes in sialic acids driven by pathogens (Cao and Crocker 2011).

The first association of *CD33* with Alzheimer's disease was discovered by Bertram et al. 2008 in a family study of 1376 Caucasian families with at least two affected family members and age of onset of greater than 50 years. The SNP, rs3826656, located 1.76kbp upstream of the start codon of *CD33* was significantly associated ($p = 4x10^{-6}$) under a dominant inheritance model. Then in 2011, two GWAS studies, Naj et al. 2011 and Hollingworth et al. 2011 confirmed the association of *CD33* with AD through the SNP, rs3865444 ($p = 1.6x10^{-9}$, OR = 0.91(0.88-0.93); n = 48589) found 0.4kbp upstream of the gene, although the association with the previous SNP, rs3826656 failed to replicate (Naj et al. 2011). The protective association for rs3865444 was then replicated by an independent group (Carrasquillo et al. 2011). The most recent meta-analysis in 2013 replicated the association for rs3865444 in the first stage of the study ($p = 5.1x10^{-8}$, OR = 0.91 (0.88-0.94); n = 54162), but this was not replicated in the second stage (n = 19884) (Lambert et al. 2013). In non-Caucasian cohorts, AD associations with both rs3826656 (Deng et al. 2012) and rs3865444 (Yuan et al. 2012) have been found in Han Chinese populations, although the effect of rs3826656 was only found in ApoE ε4-positive individuals. While in an African-American cohort, the disease association with rs3865444 was not replicated (Logue et al. 2011).

Both rs3865444 and rs3826656 are located in areas with no evidence of sequence conservation. However they both fall within a LD block that extends 10kbp upstream of the 5' end of *CD33* and contains no other potential candidate genes (figure 1.6). Four SNPs in strong LD with rs3865444 are found within the gene, *CD33* (figure 1.7). However, due to the architecture of the region, it is possible that the GWAS signal is reflecting functional polymorphisms found in a larger (~150kbp) LD block of weak disequilibrium downstream of *CD33* (Medway and Morgan 2013).

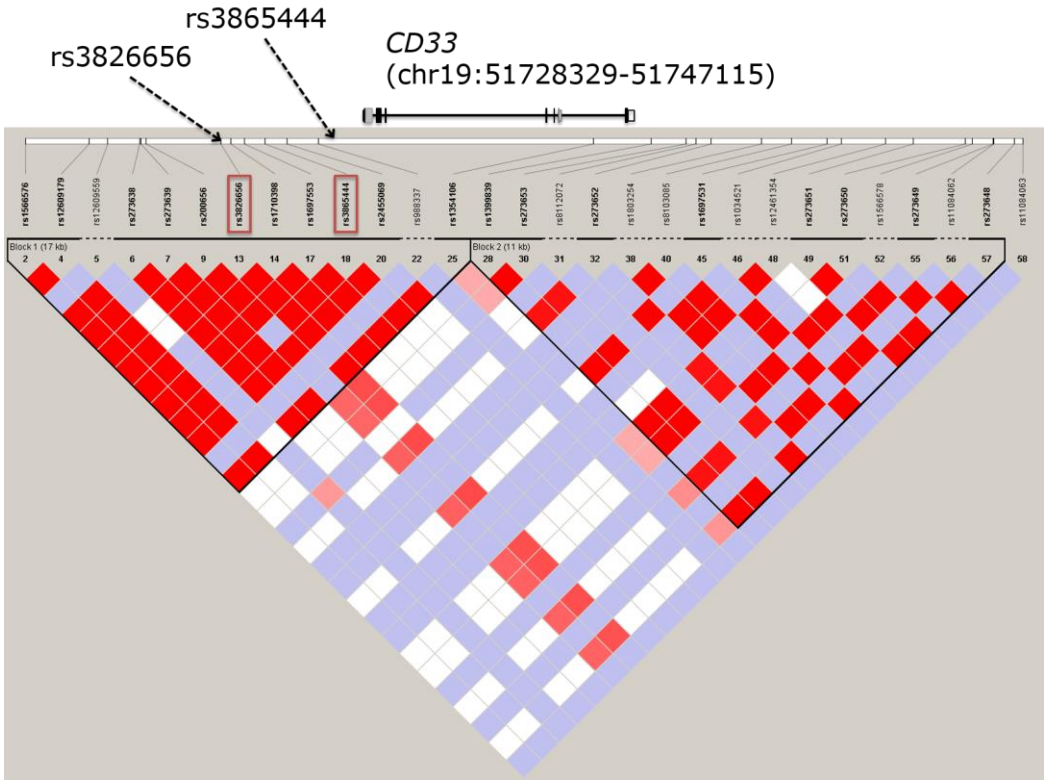

Figure 1.6. Linkage disequilibrium (LD) measured in D' surrounding *CD33* and the GWAS SNPs, rs3865444 and rs3826656 (indicated by red rectangles). The GWAS variants fall in a 17kbp haplotype block which includes part of *CD33* and upstream sequence. LD plot generated in Haploview using HapMap Data (Rel 28, Phase II + III, Aug '10). LD blocks calculated by Haploview, with the intensity of red indicating the strength of the LD measured by D'.

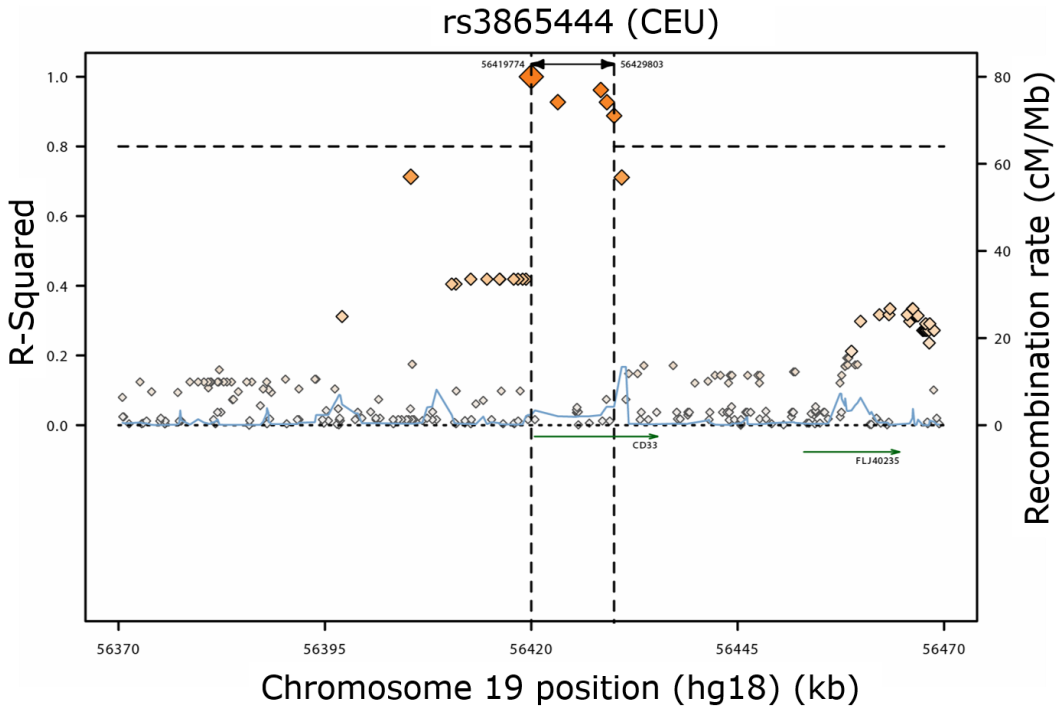

Figure 1.7. Linkage disequilibrium (LD) surrounding the *CD33* GWAS SNP rs3865444. Plot shows variants from 1000 Genomes pilot project 1 in LD with the index SNP rs3865444 (shown as the larger orange diamond).The smaller orange diamonds show four variants in a 10kbp region in high linkage (r^2>0.8) with the tag SNP. Vertical axes indicate LD ("R-Squared") shown by diamonds and recombination rate (cM/Mb) shown by the blue line graph. Plot generated with SNAP (www.broadinstitute.org/mpg/snap/ldplot.php) (Johnson et al. 2008).

The siglecs have not been thought to be involved in neurodegenerative disease, however the main siglec expressed by cortical microglia, the CD33-related siglec, siglec-11, has been shown to be activated by amyloid plaques. During amyloid aggregation, plaques become surrounded by sialylated glycoproteins, imitating the cell surface glycocalyx and thus activating the immunosuppressive siglec-11, allowing the plaques to evade microglial detection (Salminen and Kaarniranta 2009). However the opposite has also been found as there is evidence to suggest that CD33-related siglecs can reduce neuron loss, reducing the inflammatory response of microglia therefore reducing neurotoxicity (Wang and Neumann 2010).

In terms of the role *CD33* may play in Alzheimer's disease, elevated gene expression levels of *CD33* in the parietal lobe has been found to be associated with AD and also with increased cognitive decline measured by clinical dementia rating (CDR) (Karch et al. 2012). However, no association was found between the GWAS SNP and gene expression levels (Karch et al. 2012). Further examination by Griciuc et al. 2013 found that *CD33* is overexpressed in microglia of the brain in AD cases and that this expression positively correlates with insoluble Aβ levels and Aβ plaque load, while in culture, uptake and clearance of soluble Aβ by microglia was inhibited by siglec-3 (CD33) (Griciuc et al. 2013). This was confirmed by Bradshaw et al. 2013 who found increased cell surface expression of CD33 in monocytes, decreased endocytosis of Aβ peptide and increased numbers of activated human microglia (Bradshaw et al. 2013). The protective GWAS SNP was not associated with a decrease in transcript levels, but with decrease in protein product as detected by antibodies, suggesting that the functional SNP may influence translation but not stability of the messenger RNA (mRNA) (Griciuc et al. 2013). Another study confirmed the correlation of CD33 expression with microglial genes (*CD11b* and *AIF*-1), and CD33 expression was increased with AD and with the *CD33* risk allele, rs3865444C (Malik et al. 2013). This risk allele was in complete linkage disequilibrium with another SNP, rs12459419 in exon 2 of the main isoform. The SNP rs12459419 was found to modulate exon 2 splicing efficiency, generating transcripts without exon 2 therefore generating protein without the IgV domain (Malik et al. 2013). This protein would be unable to bind sialic acid. This was confirmed by Raj et al 2014 (Raj et al. 2014), who found that the CD33 risk allele (C) was associated with increased expression of exon 2 and alterative splicing was suggested as

the main mechanism of genetically driven differential expression of CD33 cell surface expression in both European and African-American individuals.

This involvement of the siglec-binding domain in Alzheimer's disease is interesting as there are antibody treatments available which specifically target CD33, through chemotherapy treatments developed for acute myeloid leukemia (AML). Examining the crossover between genetics in LOAD and AML it was found that the splicing SNP (rs12459419) which causes exon 2 skipping in the brain in AD also causes exon 2 skipping in leukocytes from AML patients (Malik et al. 2015). In a dose dependent manner, the minor allele rs12459419T decreased AD odds ratio by 0.1 and decreased CD33 expression by a quarter (Malik et al. 2015). This opens the possibility that humanized CD33 antibodies (e.g. lintuzumab) could be used to inhibit CD33 as a treatment option for AD.

Therefore, in spite of the lack of replication in the 2013 meta-analysis (Lambert et al. 2013), it seems highly likely that *CD33* is involved in AD through the immune system pathway. Immunity being one of the newer pathways identified as playing a role in AD along with other AD associated genes, *TREM2*, *CLU, CR1, ABCA7, MS4A* and *EPHA1* (Morgan 2011).

1.11. Project aims

This project aimed to identify and assess rare variants at three gene loci *CD2AP*, *EPHA1* and *CD33* identified by GWAS to be associated with LOAD disease risk. For each of the three gene loci, targeted next generation sequencing was used to sequence the entire GWAS locus as identified through linkage disequilibrium (Chapter 3). Coding and noncoding variants were prioritized for causality using both functional annotation and potential association with LOAD (Chapter 4). In Chapter 4, numerous *in silico* annotation tools were used to provide functional support for potential causative variants. Potential variants were then assessed for association with LOAD using an independent imputed GWAS dataset. Following variant prioritization, variants with predicted functionality and association were put forward for further investigation using two laboratory methods. In Chapter 5, minigene assays were used to examine the functionality of two variants predicted to affect splicing in *CD2AP* and *EPHA1*. The potential *cis*-regulatory effects of five untranslated variants from the three genes and a frameshift

variant in *CD33* were explored using allelic expression imbalance in brain tissues and B-lymphoblast cell lines (Chapter 6). Lastly the results of the study and the approaches used throughout this project are discussed in further detail in Chapter 7, and avenues for further research are proposed.

2. General Methods

This section contains detailed information on methods which are used repeatedly throughout this thesis. The chapter is split into two sections, one for laboratory methods and the other for bioinformatics or computer based methods.

2.1. Laboratory

The general “wet laboratory” experimental methods used throughout the thesis are described and justified in this section. Where any technique is specifically relevant to a particular chapter it is presented in the methods section of that chapter.

2.1.1. DNA extraction

Two methods of DNA extraction were used in this project depending on the quality of DNA required for the downstream experiments. For the NGS project and for any DNA extracted from whole blood or brain tissue from LOAD patients or control samples for the Nottingham DNA bank, DNA was extracted using phenol chloroform. For all cell culture and bacterial work, DNA was extracted using QIAamp DNA Blood mini kit (Qiagen) following standard manufacturer’s instructions.

Both whole blood samples and brain tissue were extracted using phenol chloroform. All reagents were obtained from Qiagen apart from the phenol chloroform which was from Sigma. For the brain tissue, extraction was performed using approximately 0.5cm^3 of tissue which had been stored at -80°C for a wide range of years. The sample was kept cold on dry ice and manually homogenised with a scalpel prior to being incubated overnight in a 1.5ml Eppendorf tube with 50µl proteinase K and 500µl AL lysis buffer at 50°C with constant shaking at 2.5 g to homogenise the sample. For whole blood, 2ml was pipetted into a 15ml Falcon tube with 20µl RNaseA, 2ml AL lysis buffer and 200µl proteinase K, inverting the sample before adding each reagent. Samples were incubated at 56°C for 10 minutes to homogenise the sample.

Following sample homogenisation, an equal volume of cold phenol chloroform was added to the sample (500µl for brain tissue, 4ml for whole blood). The

sample was mixed thoroughly before centrifuging. Brain tissue samples were centrifuged in a bench top centrifuge for 5 min at full speed (20000 g), while blood samples were centrifuged at 4°C at 6500 g for 15 min in a Harrier 18/18 swing bucket rotor. The phenol chloroform separated into three distinct phases following centrifugation with the top, clear phase containing the DNA, a middle white phase containing the proteins and a lower red phase containing other cellular debris. The clear phase was carefully pipetted into a new tube and an equal amount of phenol chloroform was added. The sample was mixed and centrifuged again, the brain tissue at full speed for 5 min and the blood samples at 6500 g for 15 min.

The top clear phase was again removed to a clean tube and 1/10 of the total volume of 3M sodium acetate (at pH 5.2) was added to obtain a ratio of 1:9 of sodium acetate to sample. The sample was mixed before an equal volume of ice-chilled 95% ethanol was added. The tube was inverted several times to encourage the precipitation of the DNA which was sometimes visible as a fluffy white precipitate. The sample was centrifuged for 15 min either at 20000 g for brain tissue or at 6500 g for the blood samples. The ethanol was discarded carefully so as not to dislodge the DNA pellet and a further volume of 70% ethanol was added to wash the DNA (500µl for brain tissue or 1ml for blood). The sample was treated as before and the ethanol removed. The 70% ethanol wash step was repeated and the ethanol was again removed. The DNA pellet was air dried at room temperature for 10-20 min to ensure the evaporation of any remaining ethanol. The pellet was then resuspended by heating the tube to 50°C for up to one hour in 100µl AE buffer. The DNA was stored at -20°C or at -80°C for longer storage periods.

The DNA extracted from either method was quantified using a nanodrop (ND1000) spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and the purity was assessed using the 260/280 ratio (which should be around 1.8 for DNA). For the NGS study (Chapter 3), two additional quality control steps were required. DNA degradation was assessed by loading the sample on a 1% agarose gel stained with ethidium bromide and run at 80V for at least 40 min. An additional quantification step was also required using a more accurate assessment of concentration. The Quant-iT dsDNA Broad Range Assay Kit (Invitrogen) was used following manufacturer's protocol. This method uses

PicoGreen fluorescence to measure the concentration of double stranded DNA present in a sample. This is more accurate than the nanodrop which will measure all DNA in a sample, both double and single stranded. All samples were assessed in triplicate to compensate for any bias in technique. Accurate assessment of concentration is essential for accurately sequencing from the pooled DNA samples. For each pool of 12, 500ng of DNA was used per sample (total DNA quantity of 6µg per pool). See Chapter 3, section 3.2.1 for further details on the NGS sample preparation.

2.1.2. RNA extraction

RNA extraction poses different issues to DNA extraction due to the high proportion of RNases found in the general environment. It is very important to work in a clean environment, free of RNases and also keep the RNA cold to minimise degradation. For the allelic expression imbalance study (Chapter 5) several methods of RNA extraction were tested using the brain tissue samples. However for all other experiments using cell cultures (in Chapter 4 and Chapter 5), RNA was extracted using RNeasy kit (Qiagen) following standard protocol with the additional on column DNase step. All RNA was stored at -80°C.

2.1.3. Primer design

All primers were designed using Primer 3 (v. 0.4.0) (http://frodo.wi.mit.edu/primer3/) and the appropriate reference sequence obtained from NCBI Genome browser. All primers are checked for SNPs falling within their primer binding sites using SNPCheck (https://ngrl.manchester.ac.uk/SNPCheckV3/snpcheck.htm) and potential non-specific products were identified using Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). If the reference sequences used for the primers appeared to indicate repetitive sequences (particularly if the primers were designed for intronic regions) then the sequence was submitted to RepeatMasker (http://www.repeatmasker.org/) to identify any repetitive regions. These repetitive regions were excluded from the possible sites for the primers.

When the PCR was designed to amplify complementary DNA (cDNA) samples, primers were designed to amplify across exons if possible. Alternatively the primer site was located across an exon junction. This would

help to reduce the possibility of amplifying genomic DNA which may be present in small quantities in any cDNA sample.

2.1.4. PCR and RT-PCR

Polymerase chain reaction (PCR) revolutionised molecular biology and has today allowed the development of a wide range of techniques including next generation sequencing. Given the wide use and knowledge of PCR, this thesis will not go into detail of this methodology. All PCRs in this project were carried out on Veriti 96-Well Fast Thermal Cycler (Applied Biosystems). Two different commercially available Taq and PCR reagents were used during this thesis due to a change in University approved vendors. Either NEB (New England BioLabs Inc, UK) or Roche standard PCR reagents were used as specified in the relevant projects/chapters. For both reagents, PCR reactions were set up using 10-100ng template DNA in total volumes of 30µl with 1X PCR buffer, 0.2 mM of each dNTP, 1U Taq DNA polymerase, 1pmol/ µl sense and antisense primer. Reactions were performed in a Veriti 96-Well Fast Thermal Cycler (Applied Biosystems) as follows: an initial denaturing step of 94°C for 2 min, 30 cycles of 94°C for 30 sec to denature the DNA, optimized annealing temperature for 30 sec, 72°C for 1 min for extension, and a final extension step of 72°C for 7 min.

To maintain the integrity of RNA and also to transform it into a molecule with wider applicability, it is converted into complementary or cDNA using reverse transcriptase prior to PCR and sequencing. This step is known as RT-PCR. For all RT-PCR steps in this thesis, the AffinityScript Multiple Temperature cDNA synthesis kit (Agilent) was used according to manufacturer's protocol. In total volumes of 20µl, 2µg total RNA was amplified using 25ng/µl oligo dT or 15ng/µl random primers. Reactions were set up as follows: 2µl 10X AffinityScript RT buffer, 25mM of each dNTP, 1 U RNase block and 1µl reverse transcriptase (RT). RT negative controls containing no enzymes were run simultaneously. Reactions were performed on a TRIO Thermoblock (Biometra) at 25°C for 10 min, 42°C for 1 hour, 70°C for 15 min, with a final hold at 10°C. The cDNA was stored at -20°C until required.

Amplified PCR products were separated by electrophoresis at 90V for 20 min in a 1% agarose gel stained with ethidium bromide (EtBr) in 1XTAE buffer and visualized by UV transillumination unless otherwise noted.

2.1.5. Sanger sequencing

As a gold standard method for identifying sequence variants, Sanger sequencing is still the best. Where PCR products were sequenced the following protocol was followed. PCR product (5 µl) was cleaned with ExoSAP-IT (2 µl) by incubating at 37°C for 15 min before deactivating the enzyme at 80°C for 15 minutes. ExoSAP-IT consists of two enzymes, Exonuclease I which degrades residual single stranded DNA (e.g. primers) and Shrimp Alkaline Phosphatase (SAP) which dephosphorylates residual dNTPs. Sequencing was performed with ABI Prism BigDye Terminator cycle sequencing ready reaction kit (v3.1) and 5pmol/µl sequencing primer. Following the sequencing reaction, samples were cleaned by running through Biosystems Performa DTR Gel Filtration cartridges before sequencing on an ABI3130x1 genetic analyser (Applied Biosystems) by the Molecular Diagnostics department.

Received electropherograms were checked visually in Chromas Lite (http://technelysium.com/au/). DNA sequences were exported as FASTA files and aligned to the reference sequence in Clustal Omega (v1.2.1) to check sequence identity and also enable calling of genotype of variants of interest.

2.2. Bioinformatics

This section predominantly describes and expands upon the bioinformatics programs used in the NGS chapters (Chapter 3 and 4); however these methods are also used in other chapters which will refer back to this section.

2.2.1. Power calculations

Two different power calculations were used in this thesis. The first calculated the power required to detect variants with a certain minor allele frequency (MAF) for the NGS study (Chapter 3) and the second calculated the power of the genotyping study to detect an association with LOAD.

2.2.1.1. Power calculations for detecting variants

The power to detect variants with MAF of between 1 and 5% was calculated using the formula below:

$$n \times \log(0.99) = \log(1 - \mathrm{POWER})$$

where n is the number of chromosomes (in this study, $n = 192$).

The sample size of the NGS study (n = 192) gives 86 to 99% power to detect variants with a MAF between 1 and 5%. It should be noted that this power

calculation does not take into account the error rates of the sequencing platform.

2.2.1.2. Power calculations to detect a disease association

It is important to ensure that the sample size which is available for genotyping would be expected to have sufficient power to detect an association with Alzheimer's disease prior to embarking on laboratory work. The power of the genotyping association study (Chapter 4, Section 4.2.4) to detect an association with LOAD was calculated using QUANTO (v1.2.4). The user inputs for QUANTO were specified as a matched case-control study, using the sample numbers available for the Nottingham genotyping samples (1000 cases and 970 controls) with MAF ranging from 0.01 to 0.05 increasing by 0.01 and assuming a log additive inheritance model and a 5% Type I error rate.

Figure 2.1 indicates the expected power for SNPs with minor allele frequencies (MAF) ranging from 0.01 to 0.05, with three odds ratios (OR) 1.2, 1.5 and 1.8. The figure illustrates the difficulty in detecting an association for variants with low MAF and low odds ratio (of 1.2) in the Nottingham genotyping samples. In order to investigate the effects of these variants, a far larger number of samples would be needed. However we make the assumption that rare variants found by our NGS studies (MAF between 1% and 5%) will have larger effect sizes (OR above 1.5) than typically seen in GWAS.

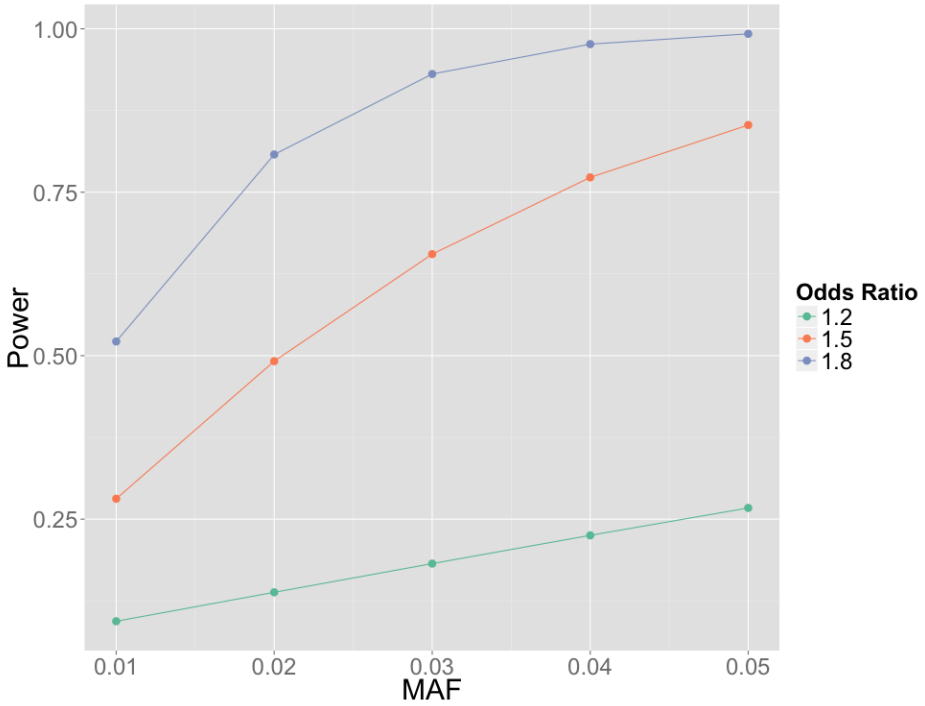

Figure 2.1. The power to detect a disease association in the Nottingham genotyping samples (1000 cases and 970 controls). The minor allele frequency (MAF) for a given SNP is shown on the x-axis and the power to detect an association on the y axis. The power for three differing odds ratios is shown. Graph produced using ggplot2 package in R.

2.2.2. PLINK

Plink v 1.07 (Purcell et al. 2007) was used for formatting files for the imputation and association testing of the independent GWAS dataset in Chapter 3 and also for the genotype association testing following the genotyping of prioritized missense variants in Chapter 4.

To perform a basic association test in Plink for the KASP genotyping data, the data was formatted for Plink. MAP, PED and COVAR files were generated as tab delimited text documents as described in figure 2.2. MAP and PED files were saved with the same file name, just different extensions in the same folder location.

Two association tests were used in this thesis, Fisher's exact test and logistic regression (Chapter 4, section 4.2.4 and 4.3.4). Fisher's test was selected as is able to reliably association test small sample sizes that are expected in the homozygous alternative (mutant) allele group when genotyping variants with MAF < 5%. Logistic regression is able to determine the impact of covariates

on the association and was used to provide further insight on covariate involvement in the genotyping association studies in Chapter 4 (section 4.2.4 and 4.3.4).

To run a Fisher's exact test in Plink the following commands were used:

```
$plink --file <inputfile> --fisher --1 --allow-no-sex --noweb --ci 0.95 --covar <covarfile> --out <outputfile>
```

While the following commands were used to run a logistic regression in Plink:

```
$plink --file <inputfile> --logistic --1 --allow-no-sex --noweb --ci 0.95 --covar <covarfile> --out <outputfile>
```

Where:

--file specifies the input file

--logistic or --fisher specifies the association test to be run

--1 indicates that the controls and cases are coded 0 and 1

--ci specifies the 95% confidence intervals be calculated

--covar specifies the covariate file to be used

--out specifies the name of the output file

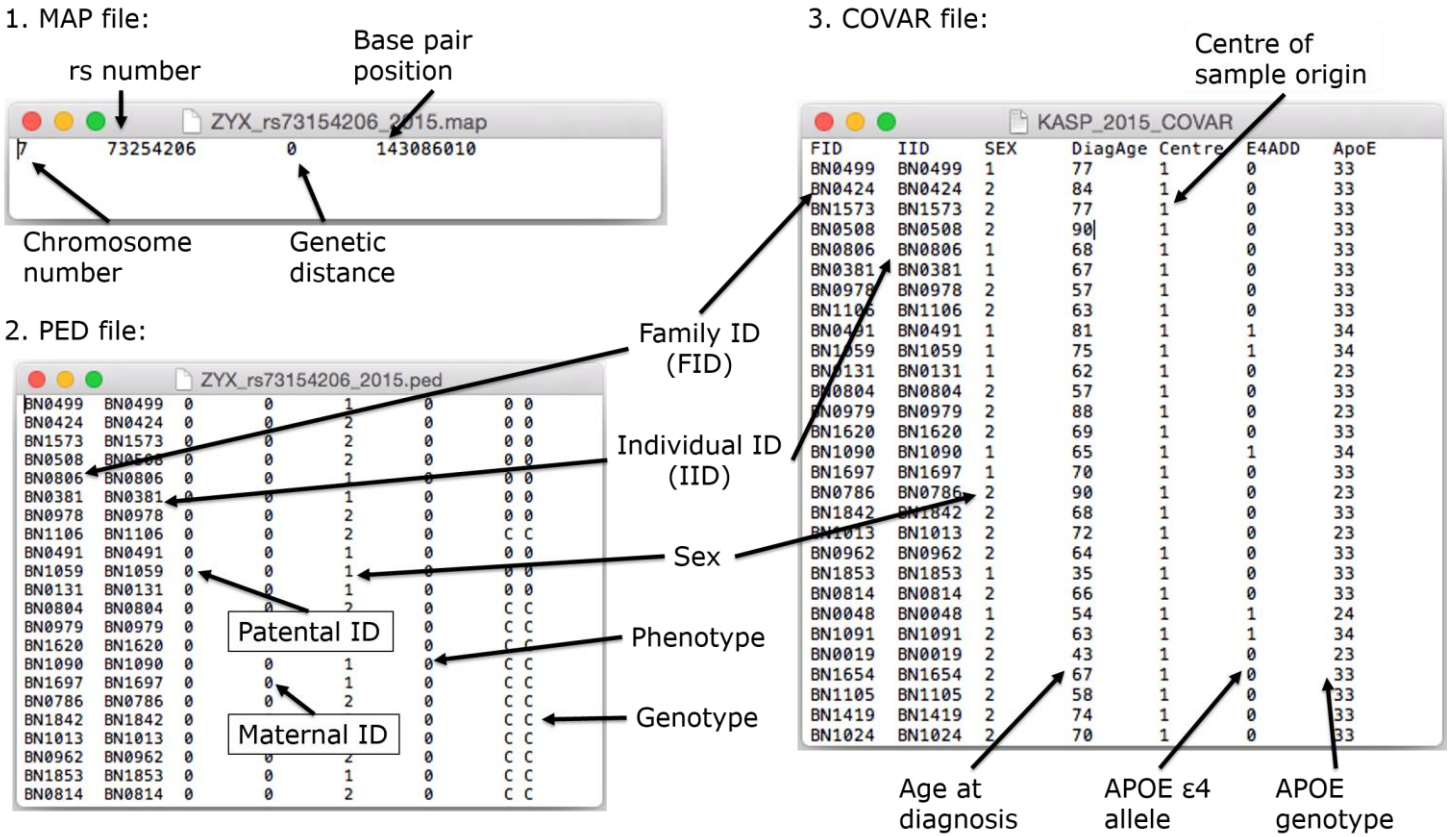

Figure 2.2. Example MAP (1), PED (2) and COVAR (3) files for Plink. 1. MAP files contained the following information: Chromosome, rs number or SNP identifier, Genetic Distance and Base-pair position. 2. PED files contained the following information: Family ID, Individual ID, Paternal ID, Maternal ID (both 0 as unknown), Sex (1 for male, 2 for female), Phenotype (0 for control, 1 for case) and Genotype (A,G,C,T or 0 for missing). 3. COVAR files contained the following information (where available): Family ID, Individual ID, Sex (1 for male, 2 for female), Age at diagnosis, Centre of sample origin (listed as a number), APOE ε4 allele number and APOE ε allele genotype.

2.2.3. NGS file formats

As the entire PhD project is based around the analysis and prioritisation of next generation sequencing (NGS) data, the file formats used in the NGS data analysis in Chapter 3 and Chapter 4 are expanded on here.

2.2.3.1. FASTQ, SAM and BAM files

The FASTQ files returned from the Illumina sequencing run were converted to SAM and BAM files through the alignment process. FASTQ files have evolved from the FASTA file format which was originally used to store nucleotide sequences as text files. In addition to the nucleotide sequence, FASTQ files store the associated Phred (or mapping) quality score for each nucleotide in the sequence. To keep the Phred score as a single character to save space the ASCII printable characters 33-126 (for Sanger format FASTQ) or 64-126 (for Illumina format FASTQ) are used to store the Phred score from 0 to 93 (or 62 for Illumina scores) (Cock et al. 2010).

Alignment programs like BFAST (see Chapter 3 section 3.2.3.2 Alignment) input FASTQ files and output SAM (Sequence Alignment Map) files. SAM files are tab delimited text files containing a header section, starting with '@' and an alignment section containing information on the alignment such as mapping position, mapping quality, orientation of the sequence and position of the mate read among others. Due to the large amount of information stored in a SAM file, it is often preferable to compress the file before any downstream analysis. BGZF compression of the SAM file is used to create a BAM (Binary Sequence Alignment Map) file. Additionally, to allow quick retrieval of alignment regions, BAM files can be indexed, generating a BAI file.

2.2.3.2. VCF files

All variants called from NGS and exome sequencing studies are stored in variant call format or VCF files. The specifications for these files are maintained by the Global Alliance for Genomics and Health Data Working group file format task team. This group manages file formats for SAM, BAM and VCF file formats. VCF files are standardized text files which contain SNP, indel and structural variant calls and are formatted in two parts, the header and the variant call records. The header lines are preceded by ## and provide information such as VCF version as well as any filters which may have been applied to the data and any additional annotation which may have been applied to the file (figure 2.3).

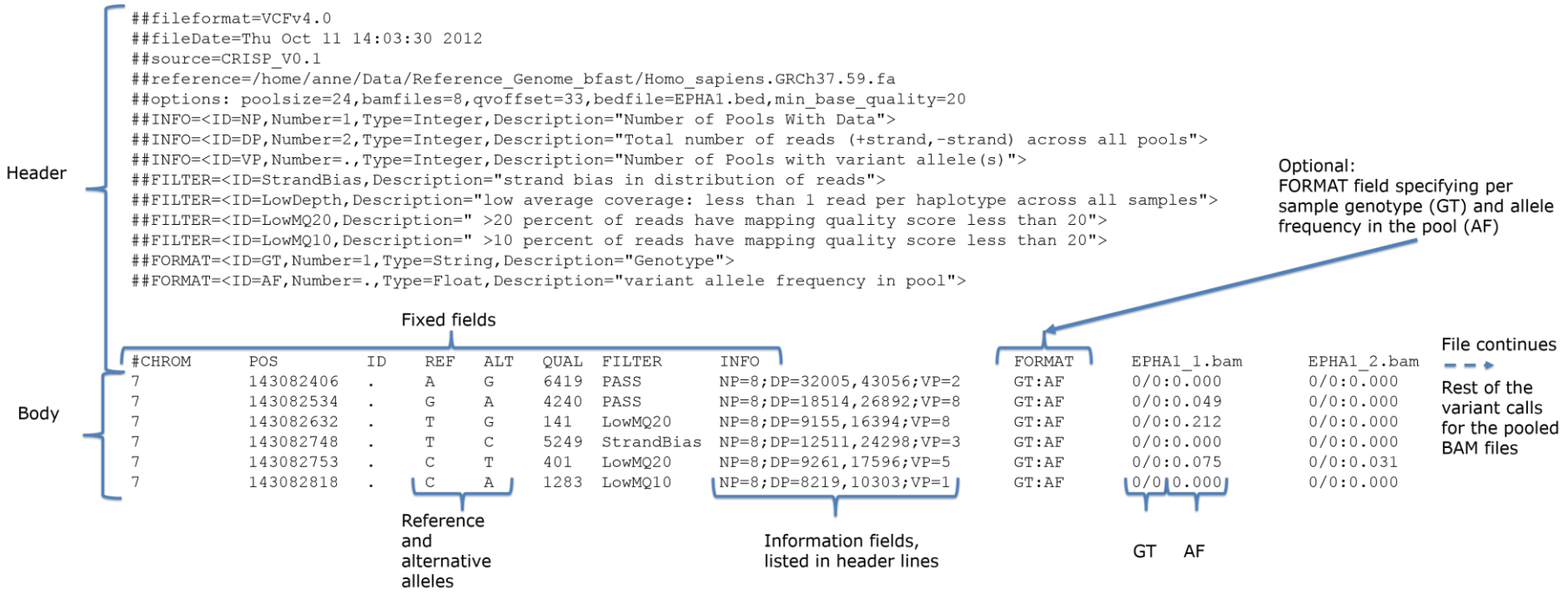

Figure 2.3 VCF file format. A simplified VCF file for *EPHA1* is shown as an example. The first eight columns are mandatory for a VCF file (shown above under Fixed fields), namely, chromosome (CHROM), position (POS), ID number (ID), reference allele (REF), alternative allele (ALT), quality score (QUAL), filter (FILTER) and information (INFO). The header lines shown by ## provide additional information on any other fields included, and detail information provided in the INFO and FILTER columns. The genotype (GT) and allele frequency (AF) are shown for the BAM files for pool 1 and 2, the next six columns in the file lists the rest of the BAM files (not shown). For pooled data the AF indicates whether individuals in that pool carry the alternative variant. For example although second variant (chr7 143082534) shows the genotype (GT) as 0/0 (or reference allele/reference allele), the AF is 0.049 indicating some individuals in the pool contain the alternative allele. This is also indicated in the INFO column where VP = 8 showing that all eight pools contain the alternative allele.

2.2.4. Linkage calculations

Two methods of calculating linkage disequilibrium were used in this thesis. To bulk process the variants output by the NGS study, a Perl program written by Dr. Christopher Medway (November 2012) was used - LD_calculator.pl (available on the Human Genetics laboratory server, University of Nottingham) which uses VCFtools (Danecek et al. 2011) and tabix (Li 2011) to calculate the LD between a specified SNP and a list of variants of interest.

The second method of calculating LD also used tabix, 1000 Genomes phased haplotype data, the VCFtools Perl script, vcf-subset, VCFtools and Plink. This was used to generate phased LD calculations between a particular pair of SNPs of interest. The following instructions were followed and changed according to the chromosome of the gene of interest. For simplicity, the calculations for the *EPHA1* gene on chromosome 7 are shown. First, the 1000 Genomes phase1 genotypes were extracted for the gene of interest using tabix:

```
$tabix -hf
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/201105
21/ALL.chr7.phase1_release_v3.20101123.snps_indels_svs.
genotypes.vcf.gz 7:143082302-143110465 >
EPHA1raw1kGSNPs.vcf
```

This command downloaded a specific section of the variation in the chromosome 7 VCF file and outputted it to a file called "EPHA1raw1kGSNPs.vcf". The commands underlined in bold were the areas which were changed according to the gene of interest.

The European (EUR) gene frequencies were extracted from the genotyping file using the VCFtools Perl script, vcf-subset. First, a tab delimited list of samples in the EUR populations was created by downloading the integrated_call_samples file (from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/phase1_integrated_calls.20101123.ALL.panel). Next, the "grep" command was used to extract the EUR samples. This samples list file was used for all chromosomes:

```
$grep EUR
Documents/integrated_call_samples_v3.20130502.ALL.panel
| cut -f1 > Documents/EUR.samples.list
```

Then the Perl program vcf-subset from the VCFtools package (http://vcftools.sourceforge.net/) was used to extract the genotypes from the 1000 genomes data as follows:

```
$perl vcf-subset –c Phase1_Samples_GBR_CEU
EPHA1raw1kGSNPs.vcf > EPHA1_CEU_GBR.vcf
```

The VCF file generated contained the 1000 genomes data for the EUR samples. This file was also used in Chapter 3, section 3.2.3.5 to compare allele frequencies from 1000 genomes with the variants called from the NGS data. Next, VCFtools was used to convert the VCF file to the Plink file format:

```
$vcftools --vcf EPHA1_CEU_GBR.vcf –plink-tped --out
EPHA1_CEU_GBR
```

And lastly, Plink was used to calculate the linkage disequilibrium between two specific variants of interest:

```
$plink –tfile EPHA1_CEU_GBR –ld rs6967117 rs1804527
```

Plink outputs the LD for the two variants and also indicates which variants are in phase with each other. Therefore this method was preferable when further information on the LD patterns for a few variants is required such as the two splicing variants (rs6967117 and rs1804527) used in this example. (See Chapter 5 for further detail on the splicing variants in *EPHA1*).

3. Deep sequencing Alzheimer's disease associated genes, *CD2AP*, *EPHA1* and *CD33*.

3.1. Introduction

GWAS has been fundamental in identifying disease risk genes for complex diseases. However, the associated SNP identified in a GWAS study is rarely the functional disease-causing SNP (Schaub et al. 2012). There may be many variants in close proximity that are in linkage disequilibrium (LD) with the GWAS SNP (International HapMap Consortium 2005; Altshuler et al. 2010) making it difficult to untangle the causative variant. Additionally, the experimental design of GWAS means that it only detects common variants that associate with disease. Following GWAS it has been found that the associated genetic loci still do not explain all of the heritability of complex diseases. It is suggested that rare variants of large effect sizes will be responsible for this "missing heritability" (Manolio et al. 2009). These reasons have led to the use of next generation sequencing (NGS) to uncover rare variants associated with disease by documenting all variants which may be found in the haplotype block tagged by the GWAS variant.

At the time of the NGS project, the three most widely accessible NGS platforms were SOLiD (Life Technologies), HiSeq 2000 (Illumina) and FLX (Roche/454) (See the reviews (Metzker 2010; Glenn 2011)). All these second generation NGS platforms follow a similar general protocol (Metzker 2010; Mardis 2011; Mardis 2013). A fragmented library is created from the target DNA which is covalently attached to platform-specific adapter sequences. The library is then amplified on a solid surface before the sequence is read at the same time as being extended. With a good cost per megabase of sequence generated ($0.10), highest output (>1000 million reads per run) and the lowest error rate (~0.1%) there is good support for the Illumina platform (Glenn 2011). Illumina uses sequencing by synthesis (see figure 3.1). Clusters of library fragments are amplified using fluorescently labelled nucleotides which temporarily terminate the sequence. The library fragments are covalently attached to a glass slide in a flow cell with eight microfluidic channels which contain complementary adapter sequences to the library adapters. Using bridge amplification (see figure 3.1), a precisely diluted library is then amplified on the flow cell surface, producing clusters of the same sequences which are

read by a laser. All four nucleotides, labelled with a different fluorophore, are present at each replication step. Following sequence detection, the terminator is removed from the nucleotide and the process is repeated.

Using NGS to whole genome sequence (WGS) a large cohort of LOAD patients and controls would enable the discovery of new rare genetic risk variants across the genome. However, the sample sizes and sequencing depth needed for sufficient power are cost prohibitive, even with the reduction in cost of next generation sequencing techniques (Koboldt et al. 2013). Therefore, target capture or enrichment (TE) is a good alternative to specifically sequence GWAS loci already identified as being associated with a particular disease. There are several methods of TE; however all are based around two techniques. The first is PCR amplification with multiplex primer pairs ligated to platform-specific adapters. The other method is hybrid capture in which the whole genome is fragmented and hybridized to complementary sequences (probes) covalently linked to biotin moieties (Koboldt et al. 2013). This allows for selection of targeted sequences through a second capture step using streptavidin-coated magnetic beads. The hybridisation method has distinct advantages over the PCR amplification method. TE using PCR can be difficult to multiplex (Wang et al. 1998; Cho et al. 1999) and there are limitations to the length of product which can be amplified (Barnes 1994). Additionally, the PCR itself can introduce additional errors in the sequence and introduce amplification bias resulting in underrepresentation of AT-rich and GC-rich regions in target sequences (Metzker 2010).

Next generation sequencing (NGS) GWAS disease-associated regions using TE has become an important next step in the search for disease-causing genetic variants (Kilpinen and Barrett 2013). In LOAD in particular, several NGS studies have already been undertaken to document the variants found in linkage disequilibrium regions surrounding disease-associated genes uncovering new variants which associated with disease (Cruchaga et al. 2012; Jin et al. 2012; Lord et al. 2012). Several different targeted sequencing approaches have been undertaken, from looking at genes associated with familial early onset Alzheimer's disease (Cruchaga et al. 2012; Jin et al. 2012) to the first GWAS genes, *CLU*, *PICALM* and *CR1* (Lord et al. 2012). However, NGS studies for the genes discovered in later GWAS have not yet been performed.

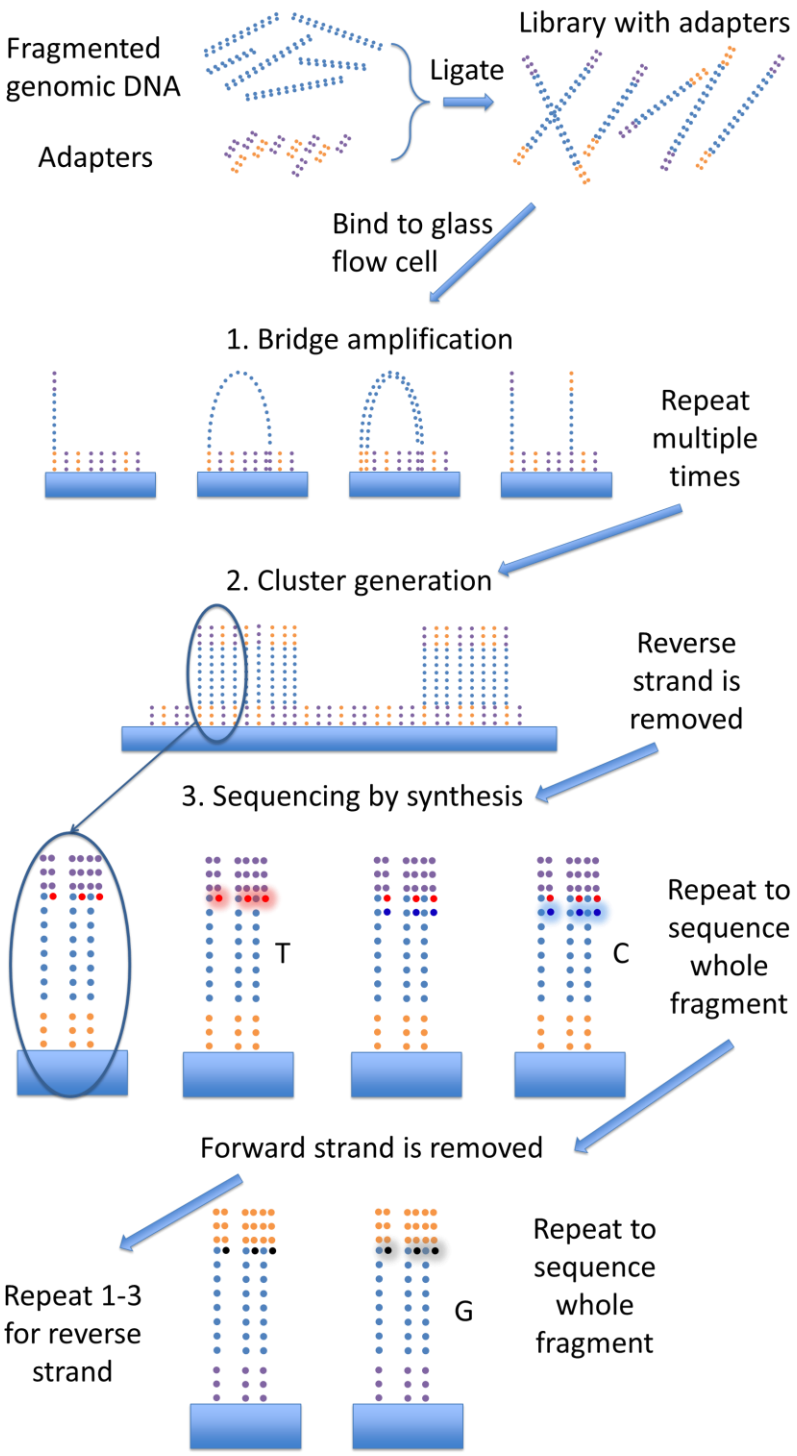

Figure 3.1. Sequencing by synthesis using paired end primers. The fragmented genomic library is represented by blue dots, while the paired adapter sequences are shown as purple (adapter sequence and forward primer) and orange dots (adapter sequence and reverse primer). The glass microfluidic flow cell is shown as a blue rectangle. Bridge amplification (1) is used to generate clusters (2) of the same fragment which are then sequenced by synthesis (3) first for the forward and then for the reverse sequence. When sequencing by synthesis, the fluorescently-labelled nucleotides are added and read one at a time using reversible terminator chemistry. Adapted from (Mardis 2013) and from the Illumina website (http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.html accessed August 2014).

3.1.1. Aims

This project investigates three of the genes uncovered by the 2011 GWAS, *CD2AP*, *EPHA1* and *CD33* using next generation sequencing of 96 LOAD patients to identify potential variants in the LD blocks surrounding the GWAS SNPs. Following variant identification (or calling), two independent methods will be used to validate the variants identified. These variants will be taken forward in the next chapter for annotation for functionality and confirmation of disease association.

3.2. Methods

J. Lord and J. Turton performed initial work on the deep sequencing project. Namely: sample preparation (J. Lord, J. Turton), initial QC of the data (J. Turton) and alignment (J. Turton).

3.2.1. Sample preparation

All participants gave informed consent and a local Ethics Committee approved the study. DNA was extracted from 96 CERAD post mortem confirmed AD case samples from The University of Nottingham Brain Bank and Manchester Brain Bank using phenol chloroform following standard laboratory protocol (See General Methods) and suspended in 100µl Qiagen AE buffer. Sample demographics for the 96 samples are presented in table 3.1, with a full sample list and demographic information provided in Appendix 1.

Table 3.1. Sample demographics for the next generation sequencing project. Number of samples (N). Mean age is the age of onset (for cases) or age of sampling (for controls) with the standard deviation in brackets. The three *APOE* epsilon alleles are shown as E2, E3 and E4.

Centre	**N**	**Mean Age**	***APOE* allele (%)**			**Sex (%)**	
			E2	**E3**	**E4**	**M**	**F**
Manchester	46	64.69 (10.56)	6.9	67.2	25.9	44	56
Nottingham	50	74.3 (12.61)	8.6	57.1	34.3	50	50

Concentration and quality of the samples were initially assessed by measuring on a Nanodrop 1000 Spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and by electrophoresis on 1% agarose gel stained with ethidium bromide. Figure 3.2 shows an example of the agarose gel quality check. Any sample showing degradation on the gel or RNA/protein contamination (A260/A280 ratio <1.8 or >2.0) was rejected.

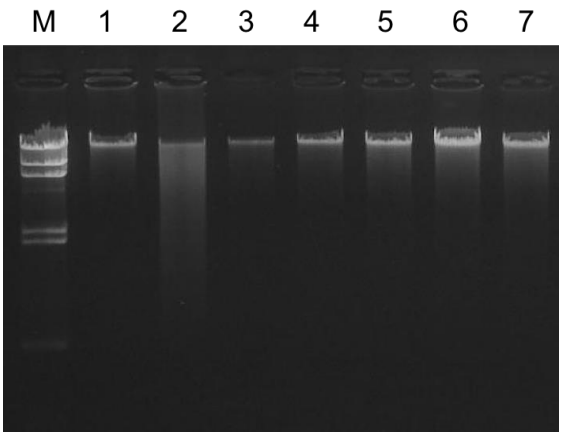

Figure 3.2. Example of gel for assessing quality of extracted genomic DNA. Lane M is a λHindIII DNA ladder. Lanes 1 and 3-7 show high quality, high molecular weight DNA, while lane 2 shows a degraded DNA sample which would have been rejected from the NGS project.

Samples passing the initial QC step were quantified using the more sensitive method; Invitrogen Quant-IT dsDNA Broad Range Assay kit (NY, USA) to obtain accurate concentrations for pooling. Pooled sequencing of groups of individuals is more cost-effective than sequencing individuals separately. The accuracy of pooled sequencing relies on the number of samples which are pooled, depth of sequencing (more than 50 times), the use of longer paired-end reads (>75 bp), proper filtering of high quality variants (mapping quality greater than 50) and an appropriate variant detection algorithm (Schlötterer et al. 2014).

Samples were combined into eight pools of 12 samples with equimolar concentrations to avoid bias. When deciding on pool size it is important to consider the possibility of distinguishing a true singleton variant from the error rate of the NGS sequencing platform. In this study a single allele would be found in 1 chromosome out of a possible 24, resulting in 4.17% of the reads generated for the pool. This is greater than the predicted error rate for the Illumina HiSeq 2000 which is expected to fall between 0.1 and 1% (Lou et al. 2013). For this study, samples were pooled for a total concentration of 6µg per pool (500ng per sample). The sample size of this study (96) gives 86 to 99% power to detect variants with a minor allele frequency (MAF) between 1 and 5%. For full details on the power calculation please see section 2.2.1 in the General Methods Chapter 2.

3.2.2. Target enrichment

For this project, target enrichment (TE) was performed using SureSelect (Agilent). Two hybridisation methods were available when this NGS project was undertaken, solution-based capture and array-based capture each with commercially available preparation kits. While array-based capture is faster and less involved than capture in-solution, it does require access to expensive equipment and between 10 and 15µg initial DNA regardless of how much sample will be targeted for enrichment. Solution-based methods do not have these drawbacks (Mamanova et al. 2010), so the SureSelect kit was selected.

Target enrichment using SureSelect Custom MP3 kit (Agilent) was designed online using eArray to selectively target loci implicated in GWAS. About 1.04MB of across 13 genes or loci were targeted (J. Turton, PhD thesis): *ABCA7, BIN1, MS4A, VAMP1, VAMP2, TRIM15, SPARCL1, CD2AP, CD33, EPHA1, IDE_KIF11_HHEX, APOE* and *CR1*. Loci were enriched for potential functional regions as follows. All design and prediction of functional loci performed by J. Turton, August 2010 unless otherwise specified (See J. Turton, PhD thesis for full information).

Possible micro RNA (miRNA) binding sites were assessed using www.microrna.org (Betel et al. 2008), and Target scan 5.1 (Lewis et al. 2005). Potential transcription factor binding sites (TFBS) were assessed using the Transfac Matrix Database v7.0 (Matys et al. 2006). Coordinates were also obtained for messenger RNA transcripts (GenBank) and spliced human expressed sequence tags (UCSC genome browser track). The LD for each gene was assessed using SNAP (SNP Annotation and Proxy Search) (Johnson et al. 2008) and HapMap Release 28, Phase II and III, August 2010 (International HapMap Consortium 2003; International HapMap Consortium 2005). And lastly, Evolutionary Conserved Regions (ECRs) were enriched for using the ECR Browser (accessed October 2010) (Ovcharenko et al. 2004).

When designing the enrichment, repetitive elements of the genome were masked using Repeat Masker (RepBase 9.11, RM database version 20050112) and Window Masker (Morgulis et al. 2006) to avoid targeting non-specific sequence. Following the design of the baits through eArray, 69.89% of the total target of all 13 genes or loci was predicted to be covered which

equates to 832825 bp of sequence. The summary of targets and baits for the genes investigated in this project is shown in table 3.2.

The pooled DNA samples were sent to Agilent Technologies (Santa Clara, CA) to generate the enriched library using the designed baits. Following enrichment, the library was sent to Source BioScience (Nottingham, UK, September 2011) for sequencing using 100bp paired-end sequencing on the Illumina HiSeq 2000 with base calling performed with the Illumina Casava 1.9 software.

Table 3.2. eArray bait design and target summary for *CD2AP, EPHA1* and *CD33*. Gene coordinates are in hg19 format, bp is the base pairs which could be targeted while target bp is the actual bases targeted after repeat masking. Table reproduced with permission from J. Turton, PhD thesis.

Locus	**Coordinates**	**bp**	**Number of baits**	**Target bp covered**	**% bp baited**
CD2AP	6:47427281-47601015	173734	4250	120861	69
EPHA1	7:143082382-143110385	28003	928	24275	87
CD33	19:51718317-51748546	30229	632	18690	62

3.2.3. Next generation sequencing analysis

The next generation sequencing pipeline was followed (figure 3.3). The majority of next generation sequencing data analysis programs are written to be implemented using a terminal command line in a Linux operating system. Therefore for this project the next generation sequencing analysis was undertaken in Ubuntu, versions 12.04 to 14.04 as the system updates were released. To streamline data processing, shell scripts were written to automate analysis where possible. A shell script is a text file containing a header line:

```
#!/bin/bash
```

along with a sequence of commands to be executed by the Unix shell or command line interpreter.

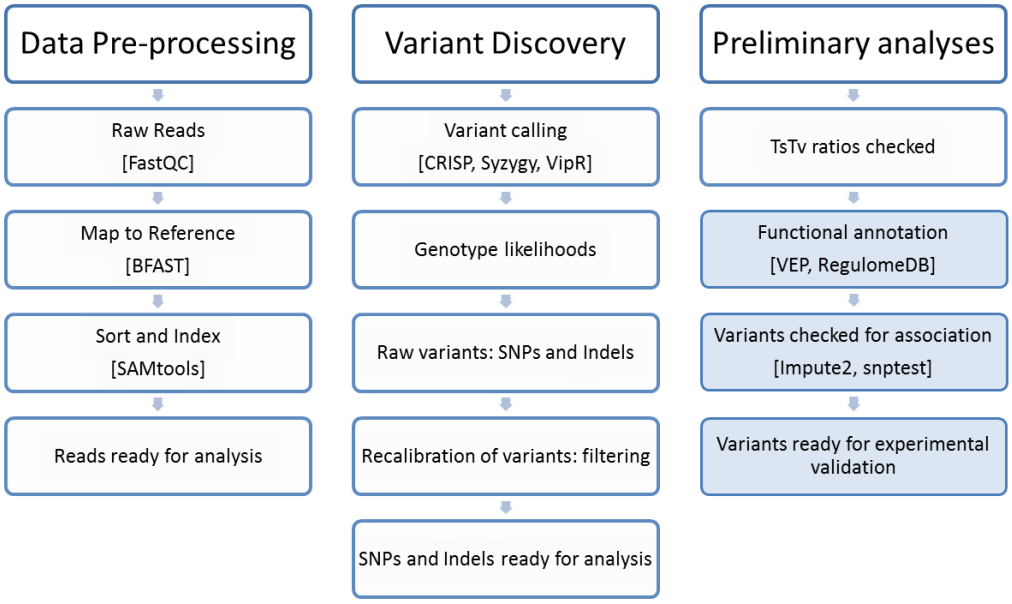

Figure 3.3. The NGS bioinformatics pipeline followed in this project. All programs used at the relevant steps are indicated in square brackets. TsTv is the transition transversion ratio for all biallelic SNPs. The last three steps are shown shaded and described in Chapter 4. Adapted from Genome Analysis Toolkit Best Practices DNA-Seq overview (https://www.broadinstitute.org/gatk/guide/best-practices accessed 30/09/2014)

3.2.3.1. Initial quality control of data received

Raw data received from Source BioScience was quality control (QC) checked by using FastQC (Babraham Bioinformatics) to determine the quality of the data. Sequence reads were interleaved prior to this QC step using an in-house Perl script written by C. Medway to orient all reads in forward orientation prior to alignment (see Appendix 2).

FastQC, a java-based program, provided information on sequence quality and content, allowing the identification of any potential issues prior to further analysis. While this program does provide a useful first pass on the quality of the sequencing run, it does have a few limitations which should be taken into account when interpreting the results. For example in order to reduce memory requirements reads are trimmed to 50bp and only the first 100 000 reads are taken into account for the Duplicate Sequences and Overrepresented Sequences modules. Additionally, for Duplicate Sequences, a non-enriched library is assumed which is violated by the target enriched library used in this NGS project.

The FASTQ files were imported into the program and it was run in standalone interactive mode for each pair of the pooled sequence data. The results were viewed in an html browser.

3.2.3.2. Alignment

Aligning next generation sequencing data is computationally demanding. There are a multitude of next generation alignment programs available, each with different strengths and weaknesses. The fastest alignment programs reply on first indexing either the sequence reads or the reference sequences or both. The majority of these programs fall into two main groups, hash-table based or suffix tree based (Li and Homer 2010). BFAST was selected for read alignment as it was specifically designed for resequencing projects with paired-end data from Illumina (Homer et al. 2009).

BFAST uses a hit and extend Smith-Waterman alignment to generate a score matrix, and Burrows-Wheeler transform methods that index the reference genome prior to searching. The index for the reference genome is provided as a string of 0s and 1s which define the bases in the read which are considered when aligning to the index or the key size. It was determined that a key size of 22 should be used for reads greater than or equal to 40bp.

First the reference genome was prepared. J. Turton provided the reference genome in GRCh37 format. The reference genome was indexed with:

```
$ bfast index -f Homo_sapiens.GRCh37.59.fa -m <mask> -w
14 -I 1 -n 8
```

Options mean:

-f : the reference file

-m : binary mask

<mask>: each of the 10 masks from table 3.3 would be used here in succession.

-w 14: width of the mask (hash or seed width) to be used

-i 1: index number - will change depending on the mask used

-n 8: use all eight processors

Table 3.3 Masks used for indexing reference genome for BFAST. M1-10 indicate the ten masks used. Only the first 14 numbers were used in each mask (indicated before the brackets). Each time the reference genome is indexed, only bases for mask positions with "1" are stored.

M1	11111111111111[1111]
M2	11111011101110[10100101011011111]
M3	10111101011010[0101100001101000111111]
M4	10111001101001[1001001111010100010111111]
M5	11111011011101[111011111111]
M6	11111110010100[1000101111101110111]
M7	11110101110010[1000101011010100111111]
M8	11110110101101[1001100000101101001011101]
M9	11110110100010[0011010110010110011010011 1]
M10	11110100101101[10101110010110111011]

The alignment was produced in two parts. Initially candidate alignment locations (CALs) were identified for each read using the indexed reference genome. The interleaved files with all reads in the forwards orientation were used to generate the alignments. To illustrate the commands used, only the commands for pool one is shown in the below examples, however a shell script was written to automate the alignment process for all pools.

```
$bfast match -f Homo_sapiens.GRCh37.59.fa -r
Sample1_R1R2.fastq -A 0 -n 8 -T /temp/ -t >
Sample1_R1R2_CAL_new
```

Options mean:

-f : the reference file

-r : Reads.fastq

-A 0: Nucleotide space

-n 8: use all eight processors

-T: Temporary folder

-t: provide time stamp

Next, the reads within each CAL are aligned a second time using gapped local alignment to identify the best match:

```
$bfast localalign -f Homo_sapiens.GRCh37.59.fa -m
/Sample1_R1R2_CAL_new -A 0 -M 500 -n 8 -t >
Sample1_aligned
```

Options mean:

-f : the reference file

-m : CALs file

-A 0: Nucleotide space

-M 500: CAL limit (default)

-n 8: use all eight processors

-t: provide a time stamp

Following local alignment, a SAM file was generated. The post-processing stage was altered to obtain correct pairing using the following commands:

```
$bfast postprocess –f refgenome.fa –I samplealigned.baf
–a 3 –A 0 –S 0 –P 2 –n 8 –O 1 –o –t > outputfile.sam
```

Options mean:

-a 3: choose uniquely the alignment with the best score (allows rescue of ambiguous reads through choosing the best aligned read in the pair)

-A 0: nucleotide data

-S 0: reads are on the same strand

-P 2: no positioning specified

-n 8: use all eight processors

-O 1: output a SAM file

3.2.3.3. Post-alignment processing

BFAST outputted aligned files as sequence alignment/map (SAM) files. These were converted from SAM files to binary sequence alignment/map (BAM) files using SAMtools (Li et al. 2009). Again, only pool one is shown in the examples, but the process was repeated for all pools.

```
$samtools view -S Sample_1_aln_a3S0P2.sam -bo
Sample_1_aln_a3S0P2.bam
```

Converting the files to BAM format reduces the file size to expedite downstream analyses. Following conversion, SAMtools was used again to sort and index the individual BAM files:

```
$samtools sort Sample_1_aln_a3S0P2.bam
Sample_1_aln_srt_a3S0P2
```

Finally, the individual gene regions were split out from the sorted and indexed aligned BAM files for further analysis and variant calling. The commands for CD2AP for pool one are shown but this was repeated for all genes for all pools using shell scripting. Gene coordinates used to excise the genes are shown in table 3.4

```
$samtools view Sample_1_aln_srt_a3S0P2.bam
chr6:47427281-47601015 -bo CD2AP_1.bam
```

To ensure that all the enriched sequence area is captured, an additional buffer region of 80bp was included for each gene region (Table 3.4).

Table 3.4. Details of the sequence split out for each gene.

Gene	**Chromosome**	**Start position**	**End position**	**Size (kbp)**
CD2AP	6	47427281	47601015	173.734
EPHA1	7	143082382	143110385	28.003
CD33	19	51718317	51748546	30.229

Following the separation of the individual gene sequences, the files were again sorted and indexed and the general characteristics and quality of the alignment assessed via the flagstat function in SAMtools and by the C program SAMstat (Lassmann et al. 2011). It is important to check the quality of the alignment, as this will affect any downstream analyses such as variant calling. SAMstat assesses the mapping quality of the alignment, giving an indication of how well the reads are aligned. Mapping quality (or Phred quality score) is an integer representative of error rate: $-10 \log_{10} P$(mapping error) (see also Chapter 2, section 2.2.3).

Post-processing usually also involves two additional steps, namely the removal of duplicate reads and the realignment of reads around indels. Duplicate reads were not removed in this analysis as the alignment involved pooled data. Therefore it is possible that the duplicates could be due to several individuals in the pool having the same read at that position (A. Altmann pers. comm.). Reads were not realigned around indels as previous analysis has determined that this does not significantly affect downstream

analyses (J Turton pers. comm.). Additionally, realignment would require several months of computer processing time as a result of the high read depth remaining due to not removing duplicate reads.

3.2.3.4. SNP/variant discovery

Pooled data presents unique issues for SNP discovery. Variant discovery software needs to be able to identify rare variants from pooled sequencing results. With error rates of NGS runs currently between 1-5%, it can be difficult to distinguish single copy alleles in a pool. Therefore it is important to use software specifically designed to deal with pooled data

CRISP (Comprehensive Read Analysis for Identification of Single Nucleotide Polymorphisms (SNPs) from Pooled sequencing) (Bansal et al. 2010) is a Python and Perl based program that detects genuine variants by comparing allele counts across several pools and assessing the probability of observing multiple variant calls across several pools due to sequencing error. The program incorporates the distribution of reads and the size of pools to remove false positives. CRISP outputs a VCF (variant call format) file containing the variants identified, quality and confidence scores.

Syzygy has also been specifically designed for calling pooled sample data and also outputs a VCF file (Rivas et al. 2011). This Python 2.6 based program relies on Bayesian likelihood (see Eddy 2004) to call variants with the posterior probability taking into account strand bias, sequence context and depth coverage.

VipR is a script written for the statistical program, R which uses the Skellam distribution to identify SNPs that have significantly different allele frequencies in at least two pools (Altmann et al. 2011). Like CRISP and Syzygy, VipR has been specifically designed for variant calling resequencing pooled data.

All SNP calling software above output VCF files which is a generic file format containing information on the SNPs called by the software and was developed by 1000 Genomes (Danecek et al. 2011). For more information on VCF files please see the General Methods section (Chapter 2, Section 2.2.3).

3.2.3.5. Validating variants

Three methods of validation were used. First variants were filtered according to the best practice variant detection as suggested by the Genome Analysis

Toolkit. Secondly, variant frequencies were compared to frequencies generated using LOAD samples sequenced on other sequencing platforms and to the publicly available 1000 Genomes database. Lastly, the transition transversion (TsTv) ratios were checked against 1000 Genomes data.

3.2.3.6. Filtering called variants

The initial variants called by CRISP were then filtered on quality (Q>30), quality per depth (Q/D >2) and on the proximity of homopolymer runs (HP>6) following the best practice variant detection suggested by the Genome Analysis ToolKit pipeline (GATK v4.0, Broad Institute (McKenna et al. 2010)) and by Altmann et al (Altmann et al. 2012). Filtering is recommended to remove false positive SNP calls.

3.2.3.7. Comparison to Exome Sequencing and 1000 Genomes variants

The majority (all but seven) of the samples in the deep resequencing study were also sent to UCL as part of an exome sequencing project. The variants called by CRISP were compared to the variants called in the exome sequencing project to validate the NGS variants.

Tabix, bgzip and VCFtools were used to obtain 1000 Genomes rs numbers and frequency information for the variants called by CRISP. Data from northern European populations (CEU and GBR) from 1000 Genomes specific to each of the three gene regions was downloaded using tabix, compressed and indexed using bgzip. SNPs and allele frequencies specific to the European population were extracted using VCFtools (Danecek et al. 2011). The 1000 Genomes SNP frequencies were then compared to the SNP frequencies calculated by CRISP using Spearman correlation in SPSS v21.

Additionally, the transition transversion (TsTv) ratios for the gene regions were calculated using VCFtools (using the --TsTv-by-count function) for the biallelic variants called in the NGS dataset before and after filtering and compared to the TsTv ratios calculated for the same gene regions from the 1000 Genomes data.

3.3. Results

3.3.1. Next generation sequencing analysis

3.3.1.1. Quality of the reads

All pools passed six of the ten FastQC modules, Per Base Sequence Quality, Per Sequence Quality Scores, Per Base Sequence Content, Per Base GC Content, Per Base N Content, Sequence Length Distribution, and Overrepresented Sequences. However, all pools failed the Duplicate Sequences module, and warnings were raised for all pools for Per Sequence GC Content and Kmer Content. While pools 2, 5, 6 and 7 raised warnings for Per Base Sequence Content. An average of 245.5 million reads per pool was obtained and pools had an average of 41% GC content.

Table 3.5. Summary statistics for FastQC analysis on raw interleaved files. Each of the eight pools was run separately for each of the ten modules in FastQC. Total reads are given per million. A ✓ indicates that the module was passed while a **-** indicates the module generated a warning and a **×** indicates that the module was failed.

Pool	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Total Reads (/million)	219.8	250.3	225.4	256.7	242.1	277.3	233.6	258.5
%GC Content	42	41	42	41	41	41	41	41
Per Base Sequence Quality	✓	✓	✓	✓	✓	✓	✓	✓
Per Sequence Quality Scores	✓	✓	✓	✓	✓	✓	✓	✓
Per Base Sequence content	✓	-	✓	✓	-	-	-	✓
Per Base GC content	✓	✓	✓	✓	✓	✓	✓	✓
Per Sequence GC Content	-	-	-	-	-	-	-	-
Per Base N Content	✓	✓	✓	✓	✓	✓	✓	✓
Sequence Length Distribution	✓	✓	✓	✓	✓	✓	✓	✓
Duplicate Sequences	×	×	×	×	×	×	×	×
Overrepresented Sequences	✓	✓	✓	✓	✓	✓	✓	✓
Kmer Content	-	-	-	-	-	-	-	-

3.3.1.2. Quality of the alignment

Approximately 99% of reads were mapped correctly in all genes according to the flagstat function of SAMtools (table 3.6) and a similarly high percentage (97-98%, table 3.6) were also properly paired to their mate read. The majority

of aligned reads for the three genes also had a good mapping quality of greater than 30, which corresponds to an error rate of 0.001 (table 3.6).

Table 3.6. Quality of the alignment as assessed by the flagstat function in SAMtools and the program SAMstat. Values are listed as percentage of sequence reads for both analyses.

Gene	**Flagstat**		**SAMstat**
	Mapped correctly	**Properly paired**	**Mapping Quality >= 30**
CD2AP	99.75	98.25	93.64
EPHA1	99.62	97.59	92.26
CD33	99.73	98.12	79.03

To assess the success of the TE, two metrics were calculated for the sequencing study, the specificity of the enrichment and the enrichment factor (see Mertes et al. 2011).

1) The specificity of the enrichment:

$$\frac{Total\ number\ of\ reads\ in\ TE\ region}{Total\ \ reads\ obtained} = \frac{1209121676}{1949915034} = 62\%$$

2) And the enrichment factor was also calculated, first by calculating the average depth of coverage at the targeted region.

$$\frac{Reads\ mapping\ to\ targeted\ region\ \times\ read\ length\ (bp)}{Total\ size\ of\ targeted\ region\ (bp)}$$

$$= \frac{1209121676 \times 100}{832825}$$

$$= 145183.2 \times$$

Then by calculating the approximate average depth of coverage in the rest of the genome (using $3x10^9$ bp as the approximate size of the human genome).

$$\frac{Total\ reads\ mapping\ to\ rest\ of\ \ genome\ \times\ read\ length\ (bp)}{size\ of\ \ genome\ (bp)}$$

$$= \frac{(1949915034 - 1209121676) \times 100}{3 \times 10^9}$$

$$= 24.7 \times$$

Finally the enrichment factor was calculated.

$$\frac{Average\ depth\ of\ \ coverage\ at\ the\ targeted\ region}{Average\ depth\ of\ \ coverage\ at\ the\ rest\ of\ the\ genome} = \frac{145183.2}{24.7} = 5877 \times$$

Therefore the sequencing of the targeted region of interest was 5877 fold higher than the sequencing of the rest of the genome.

3.3.1.3. SNP/Variant discovery

The three SNP calling programs tested, CRISP, vipR and Syzygy, all called similar numbers of SNPs, although Syzygy called nearly three times as many as vipR (Table 3.7). However, CRISP called all of the 106 variants also found by vipR, and 109 variants that were also called by Syzygy. Therefore this method was selected as the most appropriate, as it appears to maximize SNP discovery and minimize false positives. The output is also more useful than vipR, as the program calculates a range of descriptive statistics that allows the quality of the called variants to be assessed.

Table 3.7 Comparison of the SNP calling software, CRISP, vipR and syzygy using the NGS data for EPHA1 as an example. The total number of variants found as well as the number of variants found by the other programs are listed. The same settings were used for all programs. Low quality variants were assessed using the different statistics available in the different programs (The filter output for CRISP and flag output for Syzygy).

Program	**Total variants**	**Common variants**	**Low quality variants**
vipR	106	85	NA
CRISP	192	106	44
Syzygy	266	85	95

CRISP called a total of 1273 variants in the NGS data for the three genes, which were reduced to 831 following filtering (Table 3.8). Most of the variants called were intronic, as only 17 variants are coding or influence splicing and 21 are found in untranslated regions (UTRs).

Table 3.8. Summary of the variants found for the NGS data in *CD2AP*, *EPHA1* and *CD33* using CRISP. The number of variants before filtering are shown in brackets in the first column. The next column shows the average depth of coverage per chromosome for the study, then a breakdown of the variants are listed.

Gene	**Total variants (unfiltered)**	**Ave depth of coverage per chr**	**SNPs**	**Indels**	**Exonic**	**Splice**	**5' UTR**	**3' UTR**	**Total coding + UTR**
CD2AP	626 (953)	434.0	592	34	3	1	2	13	19
EPHA1	127 (192)	290.5	119	8	9	1	-	2	12
CD33	78 (128)	358.19	74	4	3	-	-	4	7

3.3.1.4. Validating variants

Comparing the variants which were found in both the NGS and the Exome sequencing study revealed 24 variants were called in both datasets out of a possible 38 coding and UTR variants called in the NGS study for the three genes (sum of total coding and UTR variants, last column table 3.8).

The variant frequencies calculated by CRISP appear to be good approximations, as confirmed by the significant positive correlation between the MAF estimates from CRISP and the MAFs observed in the 1000 Genomes data (Spearman Correlation Coefficient = 0.604, $p < 0.0001$).

Additionally, the TsTv ratio for variants called by CRISP from the three gene regions were comparable to the TsTv ratios obtained for the same regions from 1000 Genomes data (Table 3.9).

Table 3.9. Transition transversion (TsTv) ratios for the three gene regions pre- and post-filtering. The last column contains the TsTv ratios for biallelic SNPs from the same gene regions from 1000 Genomes

Gene	**TsTv Pre-filtering**	**TsTv Post-filtering**	**TsTv for 1000 Genomes**
CD2AP	1.761	2.273	2.248
EPHA1	1.904	3.423	2.934
CD33	1.174	2.083	2.171

3.4. Discussion

With the advent of GWAS, ten reproducible genetic associations for Alzheimer's disease were discovered and new pathways were implicated in the pathology of the disease. To gain further insight into three of the GWAS associated genes, *CD2AP*, *EPHA1* and *CD33*, the genes and surrounding LD region were sequenced using next generation technology, discovering over a thousand variants (table 3.8). These variants will be annotated for functionality and association tested in an imputed independent GWAS dataset in Chapter 4.

3.4.1. Quality of the raw reads

When considering the results of the ten modules run by FastQC, the warnings or fails should be investigated to ensure that the issues the program is flagging for the particular modules are valid for the NGS project being

undertaken. In this project the only module or parameter which failed entirely was Duplicate Sequences. The Duplicate Sequences module plots the percentage of duplicated sequences found in the pools and issues a failure if more than 50% of the sequences are non-unique. However, this is to be expected if the NGS project is using a target enriched library as was the case in this thesis. This violates the assumption of the Duplicate Sequences module by creating duplicates through TE.

For the warning messages, all pools flagged issues with Per Sequence GC Content and Kmer Content. The Per Sequence GC Content warning is raised when more than 15% of the reads deviate from the normal distribution as modelled for each particular file analysed, possibly indicating biases with the library preparation. Unfortunately, biases with GC content is a known issue for Illumina sequencing technologies (Dohm et al. 2008; Aird et al. 2011; Benjamini and Speed 2012; Oyola et al. 2012). Additionally, this metric could be being influenced by the high GC content of the sequencing primers. Therefore it is acceptable to ignore this warning. The Kmer Content only interrogates 2% of the library. The number of every possible combination of 7bp oligomers at each position in the library is measured and tested for significant deviations from even coverage. It is expected that this module will issue a warning when non-random priming is used as was the case in this NGS project. Four of the pools (2, 5, 6 and 7) raised warnings for Per Base Sequence Content. This module plots the proportion of each base across each position in the library and flags a warning if the difference between any of the four bases is greater than 10% for any position. For the four pools flagged with warnings, the difference was only greater for the first and last base in the read. Therefore this difference in base proportion does not represent a systematic bias in the library preparation and can be disregarded. As the raw interleaved files were deemed acceptable following QC with FastQC the data was taken forward for alignment.

3.4.2. Quality of the alignment

Using the custom settings for the post alignment processing of the data resulted in nearly 99% of the reads being properly mapped and paired (table 3.6). High quality alignment of NGS reads is particularly important as all downstream analyses will depend on a good alignment. Mismatched bases

introduced at this stage could later be interpreted incorrectly as variant sites, leading to false positive variant calls.

Additionally, it is important to check the TE of a sequencing project to ensure that sufficient targeting of the intended region has occurred to allow sequencing at sufficient depth for testing the aims of the sequencing project. This project had an enrichment factor of 5877 times, which is very high. This could stem from over sequencing as a result of targeting only 1.1Mb of sequence for the HiSeq2000 which can sequence well over 600Gb of sequence in a run. The average coverage per chromosome was also well over the general recommended coverage for resequencing projects which suggests that an average depth of between 35 to 50 times be used (see table 3.8) (Sims et al. 2014).

3.4.3. SNP/variant discovery

It is important to test different methods of SNP detection, as inherent biases in different algorithms can lead to the generation of false positives (Altmann et al. 2012). The three programs tested appeared to call similar SNPs, however CRISP provided the best balance between number of variants called and quality of variants (table 3.7). This is presumably due to the unique detection method of the program, comparing variants detected across all pools to reveal likely false calls (Bansal et al. 2010).

NGS has considerably high error rates, making it difficult to distinguish single copy alleles from false positive calls. Therefore alternative methods of verification should be used to validate any variants called from NGS data (Lord et al. 2012).

3.4.4. Validating variants

In any NGS study it is important to validate variants discovered by alternative means such as sequencing using a different method, comparing the minor allele frequency obtained in the NGS study to an independent database or using additional metrics such as transition transversion ratios which can give a general indication of the quality of sequencing and variant calling.
The majority of the samples used in this study were also sequenced in an exome sequencing study. This provided a good opportunity to validate exonic variants discovered in this dataset with those found in the exome sequencing study. Unfortunately, the exome sequencing study was not targeted for the

GWAS loci and did not achieve good sequencing depth for the three genes investigated in this NGS study. Therefore only 63% of the coding variants identified in this study were also identified in the exome sequencing study (14 out of 38). While it might be surprising that the exome sequencing missed these variants, WGS has been found to be more efficient at detecting mutations within the exome (Belkadi et al. 2015). Additionally, as this NGS study was targeted genome sequencing, the exome sequencing data did not provide any validation for calling variants in the noncoding regions of the genome. Luckily many of the variants identified in the NGS study have already been discovered in the 1000 Genomes study allowing comparisons to be made between the datasets. There was a good correlation between the minor allele frequencies found in the two datasets.

Additional confidence in the variants which passed filtering can be taken after comparing the TsTv ratios for the three gene regions. The TsTv ratios after filtering better approximated those found in 1000 Genomes. Calculating the TsTv ratios for the 1000 Genomes data for the region was used to generate a more accurate TsTv ratio for the gene regions sequenced. It is generally accepted that transitions occur twice as often as transversions in whole human genome sequencing (so a TsTv ratio of 2). However, TsTv does vary across the different regions of the genome with TsTv ratio for exonic sequence usually higher at between 2.8 and 3.0 (Le and Durbin 2011).

3.4.5. Conclusions

Using targeted next generation sequencing of pooled individuals is a cost effective way to identify novel rare variants in genetic loci which are associated with a particular disease. It is useful to know the biases which may arise from a particular next generation sequencing platform to enable sensible interpretation of initial quality control results. The unique qualities of pooled sequencing require specialised programs to call variants from the aligned data. However several good options exist for this. Following variant calling, filtering on various quality metrics as recommended by GATK provides a good method of reducing the number of false positive variants, thereby reducing the amount of work required in the downstream analyses annotating and prioritising variants. Given the large number of variants which still remain following filtering, the next task, undertaken in Chapter 4, will be successfully prioritising the variants for further experimental work.

4. Annotating and association testing NGS variants in *CD2AP*, *EPHA1* and *CD33* to identify potential functional variants.

4.1. Introduction

Next generation sequencing (NGS) has resulted in an almost exponential increase in the number of rare variants being discovered, both in the general population through projects such as 1000 Genomes and also in patients with disease through the efforts of consortia such as IGAP. It can be assumed that the majority of the variants identified will be neutral (i.e. non-deleterious). Consequently, the issue has shifted to become one of interpreting and prioritising the large number of variants discovered by NGS in order to distinguish disease-causing variants from neutral variation which occurs within a population. In the early days of NGS, new rare variants could simply be filtered by removing any variants found in the public dbSNP dataset (Koboldt et al. 2013). This is no longer a very sensible option as public databases now document large numbers of rare and deleterious variants (e.g. phase 3 data from the 1000 Genomes project) (Koboldt et al. 2013).

Luckily there are a large number of bioinformatics programs available that have been written in response to the problem of annotating and prioritising NGS data (reviewed in (Bao et al. 2014; Pabinger et al. 2014)). Two of the most popular are ANNOVAR (Wang et al. 2010b) and VEP (McLaren et al. 2010). They benefit from accessing online databases which are readily updated and so the most current information will be added to the annotation analysis. However there are still difficulties interpreting the data as any annotation program will only be as good as the experimental databases providing the annotation (McCarthy et al. 2014). It is important to verify annotations with additional *in silico* prediction tools to strengthen the support for further laboratory investigation for any particular variant. To illustrate, both ANNOVAR and VEP will annotate variants as splice variants if they fall within a specified number of bases of a known splice site. However, simple proximity to a splice site does not translate to effect. If the variant does not affect the consensus splice sequence it will not disrupt splicing. Additionally variants not annotated as splicing variants by the annotation tools could influence splicing if they alter or destroy an exonic or intronic splice enhancer or silencer sequence (see Chapter 5, figure 5.1).

It is also important to be able to relate the sequencing results back to the phenotypes of the individuals being sequenced. For example, just because a mutation is missense and predicted to be deleterious doesn't necessarily translate to causality. Healthy individuals have been found to carry several putative deleterious variants with no apparent effect (as revealed in Abecasis et al. 2010; MacArthur et al. 2012; Xue et al. 2012; Shen et al. 2013). Therefore, proving a causal link between a potential variant identified in a NGS study and the disease is often difficult (Guerreiro et al. 2014; MacArthur et al. 2014). This problem has been reviewed by MacArthur et al. (MacArthur et al. 2014). They recommend implicating the potential variant as being involved in the disease through combining evidence from a number of different areas to assess the support, rather than assigning causality outright (MacArthur et al. 2014).

When searching for disease-causative variants, an important criterion is that they associate with the disease of interest. Rare variant association testing is notoriously difficult, requiring large sample sizes in order to achieve sufficient power for a meaningful association test to be performed (Zuk et al. 2014). Imputation can provide a potential solution for association testing rare variants (defined in this thesis as variants with MAF <5 %). The large increase in the number of rare variants documented in the 1000 genomes project has provided researchers with a vastly improved reference panel for imputation. Imputation predicts genotypes not directly sequenced on a genotyping chip using a dense reference panel of genotypes (Marchini and Howie 2010). It has often been used in GWAS to increase the number of SNPs available for association testing. However, rare variants were usually excluded from downstream analyses due to the difficulties with imputing their genotypes (Pei et al. 2008; Huang et al. 2009; Sung et al. 2012). With the use of a reference panel from the 1000 Genomes Phase 1 dataset, Impute2 is able to reliably predict the genotype of rare variants (Howie et al. 2009; Howie et al. 2012). Therefore, imputation provides an attractive alternative for association testing rare variants which impute in previously used LOAD GWAS datasets (e.g. Asimit and Zeggini 2012).

4.1.1. Aims

This chapter functionally annotates the variants identified in the NGS chapter using *in silico* methods. Numerous *in silico* tools are available that can provide annotations on potential functional effects of a SNP, although many are unable to process SNPs in bulk.

Imputation in an independent dataset (2067 LOAD cases, 7376 controls) will then be used to find potentially associated SNPs. Imputation uses population genetic models to predict unobserved genotypes in a study sample set from a reference panel of individuals genotyped at a large number of SNPs, effectively increasing the number of SNPs which can be association tested.

Following prioritisation with functional annotation and association testing through imputation, missense variants identified in each of the three genes will be genotyped in an independent dataset to confirm disease association. Variants that are predicted to have a detrimental functional effect, and any that show association following imputation or genotyping will be put forward for further investigation.

4.2. Methods

4.2.1. Annotating variants

Ensembl's Variant Effect Predictor (VEP, previously SNP Effect Predictor, (McLaren et al. 2010)) is an online or Perl scripted tool that allows several SNPs to be checked simultaneously. The filtered VCF files for each of the three genes were annotated using the Perl API version of VEP (variant effect predictor) modified by C. Medway to include annotations from the Encyclopedia of DNA Elements (ENCODE) project downloaded from the UCSC genome browser website (http://genome.ucsc.edu/ENCODE/downloads.html, accessed Nov 2012 (Consortium 2011)) (Appendix 3).

The ENCODE annotations for DNase I hypersensitivity, TFBS, methylation and histone modification in relevant cell types (SK-N-SH, H1HESC, NHA, HepG2) were supplemented with UCSC tracks for conservation (both phyloP and phastCons), transcription factor binding sites (TFBS) from the TRANSFAC matrix database (v7.0), CpG islands, miRNA binding sites from TargetScan miRNA database and VISTA enhancers.

DNase I hypersensitivity sites are found by digesting DNA with DNase I to locate areas of chromatin which are nucleosome free and accessible and therefore undergoing active transcription. This can indicate regulatory areas such as promoters, enhancers, silencers and locus control regions (Crawford et al. 2006). TFBS are indicated through ChIP-seq, which have been applied by ENCODE to a specific variety of transcription factors (TFs) as were histone modifications and methylation status (Consortium 2011). The addition of the ENCODE annotations enabled the possible functional properties of the noncoding variants to be assessed.

Linkage disequilibrium (LD) between the variants and the original GWAS SNPs were calculated using an in-house Perl script (see Chapter 2, section 2.2.4), VCFtools, tabix and bgzip using 1000 Genomes project phase 1 data for the northern European cohort (CEU and GBR populations).

Additional *in silico* programs were used to assess the coding and UTR SNPs. The conservation of SNP locations across mammalian species were assessed using PhyloP and Phastcons data from UCSC genome browser's Tables function. All exonic variants were assessed for potentially disrupting exonic splicing enhancer sites using ESEfinder v3 (http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home) (Cartegni et al. 2003). For 3'UTR SNPs, potential miRNA binding sites were assessed using TargetScan v6.2 (http://www.targetscan.org), miRDB (http://mirdb.org/miRDB/) and MicroSNiPer (http://epicenter.ie-freiburg.mpg.de/services/microsniper/), while the structural effect of UTR SNPs were assessed using RNAsnp (http://rth.dk/resources/rnasnp/) (Sabarinathan et al. 2013a). All programs were accessed February 2013.

4.2.2. Imputation and association testing variants

To test for an association between the rare variants discovered from the NGS sequencing study and Alzheimer's disease, the merged genome wide association study (GWAS) datasets from the Alzheimer's Research UK and Mayo datasets (Carrasquillo et al. 2009) (2067 AD cases and 1376 controls, genotyped on HumanHap300v1 (Illumina)) and 6000 control cases from the publicly available WTCCC2 National Blood Service (NBS) and 1958 British Birth Cohort (genotyped on 1.2M chip (Illumina)) were imputed against 1000

Genomes reference panel, using default settings for Impute2, running with one phased reference panel (Howie et al. 2009). All programs and data were downloaded in April 2013.

Imputed datasets were association tested using snptest v2.4.1 (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html). This program is written by the same group as Impute2 and is designed specifically to association test datasets imputed with Impute2. The frequentist method with an additive inheritance model was selected for the association test. Genotype uncertainty from the imputed dataset was accounted for using the default threshold method. The effect of the different cohorts, the merged GWAS dataset and the WTCCC2 control dataset, was controlled for during the analysis.

Impute2 was selected as the imputation program for this study as it uses 1000 Genomes data (1000 Genomes Phase 1 integrated variant set (v3) NCBI build 37, accessed March 2013) as a reference panel, making it robust to population stratification. This makes it a useful tool for imputing rare variants which are usually more sensitive to population stratification. Impute2 separates the phasing of observed genotypes and the imputation of missing genotypes, allowing for more information to be incorporated at the phasing step. These two steps are repeated over 30 iterations using a Markov Chain Monte Carlo (MCMC) prediction model. SNPs are classified depending on whether they are genotyped in both datasets, T (typed SNPs in both study and reference datasets), or in just the reference dataset, U (untyped SNPs in the study dataset). Similar haplotypes are identified in the Reference panel by phasing SNPs found in both datasets (T) and comparing these to SNPs in the Reference panel. This allows the U SNPs to be predicted from the haplotypes identified in T, with the assumption that if the haplotype matches at T, it will also match at U (Howie et al. 2009).

The merged GWAS dataset and the WTCCC2 NBS and 58C datasets were provided in hg19 format (GWAS dataset lifted over by J. Lord, former PhD student, WTCCC2 dataset lifted over by C. Medway, former Post Doc). The chromosomes for the genetic loci which were analysed in this NGS study (chromosomes 6 (for *CD2AP*), 7 (for *EPHA1*) and 19 (for *CD33*)) were extracted in Plink and saved as .map and .ped files:

```
$plink --bfile Merged_data_hg19 --chr 19 --make-bed --
recode --out chr19_Merged --noweb --allow-no-sex
```

The commands used for chromosome 19 for the merged dataset are given above as an example. These steps were automated with shell scripting to repeat the process for all three genes and across all three datasets (GWAS merged, WTCCCC2 NBS and WTCCC2 58C datasets).

The .map and .ped files were then converted to .gens and .samples files using GTOOL (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html):

```
$./gtool -P --ped chr19_Merged.ped --map
chr19_Merged.map --og chr19_Merged.gens --os
chr19_Merged.samples
```

Impute2 requires the genotyping data in .gens and .samples format. The .samples header was edited to change the trait being examined to binary (P>B). The phenotype was also changed in the .samples file from Plink format (1/2 for cases and controls) to 0/1 and Centre was added as a discrete variable (D) for covariate analysis. Imputation in Impute2 was then performed using the target coordinates and a buffer region of +- 1Mb:

```
$impute2 -m genetic_map_chr19_combined_b37.txt -h
ALL_1000G_phase1integrated_v3_chr19_impute.hap.gz -l
ALL_1000G_phase1integrated_v3_chr19_impute.legend.gz -g
chr19_Merge.gens -align_by_maf_g -int XXXX -Ne 20000 -o
Gene_phased.impute2
```

Where:

-m is the fine scale recombination map for the region to be analysed

-h is the known haplotype file

-l is the legend file. Both the .hap.gz and .legend.gz are the 1000 genomes phase 1 reference data and were downloaded from Impute2 website (https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated).

-g is the .gens file created in GTOOL. The .samples file should also be in the same location.

-int specifies the imputation interval

(for *CD2AP*: 47427281-47601015, for *EPHA1*: 143082382-143110384 and for *CD33*: 51718317-51748546).

-Ne refers to the effective population size that the imputed data was sampled, with 20000 being the default value.

-align_by_maf_g is used where no strand file was specified.

For the GWAS dataset the strand file was downloaded (http://www.well.ox.ac.uk/~wrayner/strand/BDCHP-1x10-HUMANHAP300v1-1_11219278_C-b37-Source-Strand.zip) and was specified using the -strand option.

Following imputation, the success of the run was assessed using the concordance value which was output on the terminal screen upon the completion of a run. Successful imputation is indicated with concordance values >95%. Concordance measures the agreement between the imputed and actual genotypes as a check to verify the imputation process has completed successfully for any particular dataset. To assess the success of the imputation at any particular SNP, the info score was used. The info score indicates the certainty of the genotype prediction at a particular site and ranges from 0 to 1. A score of 1 indicates the SNP was present in both datasets and was imputed with complete certainty (Marchini and Howie 2010). As per convention, SNPs which were imputed with info scores lower than 0.4 were excluded from the association test.

SNPtest v2.4.1 was used to association test the imputed dataset as this is the default association testing program provided with Impute2. Assuming the imputation passed concordance, SNPtest v2.4.1 was run incorporating all three datasets using:

```
$ SNPtest_v2.4.1 -data
CD33_Merged_phased.impute2 chr19_Merged.samples
CD33_NBS_phased.impute2 chr19_NBS.samples
CD33_58C_phased.impute2 chr19_58C.samples
-o CD33_All_snptest.out -frequentist 1 -method
threshold -pheno phenotype -cov-name centre -overlap -
missing-code -9
```

The output from the association test in SNPtest v2.4.1 was a tab delimited text file containing various statistics for association for each SNP tested including odds ratios (OR), MAF of SNP in cases and controls and p-values.

4.2.3. Variant prioritisation

Variants were prioritised using the VEP annotation supplemented with the ENCODE datasets for conservation, TRANSFAC (TRANScription FACtor database), transcription factor binding sites, DNase hypersensitivity sites,

methylation (H3K4me1, H3K4me2, H3K4me3) and acetylation (H3K27ac). A variant was prioritised if there was evidence of these measures of functionality occurring at the variant's position. As epigenetic changes are known to be tissue specific, the methylation and acetylation marks were further assessed according to particular cell types deemed appropriate for Alzheimer's disease. Namely normal human astrocytes (NHA), neuronal glioblastoma cell line (SK-N-SH) and liver hepatocellular carcinoma cell line (HepG2). HepG2 was selected as this cell line has been documented to overproduce amyloid beta and so has been proposed as a potential cell model for Alzheimer's disease (Koudinov and Koudinova 1997). Additionally, variants were selected if they were in high LD with the GWAS variant (as measured by D').

Unfortunately, *CD33* was not well annotated by VEP or the ENCODE datasets used at the time. Therefore the only noncoding variant with any lines of annotation was put forward as potentially interesting and any missense or frameshift coding variants were prioritised for this gene.

Given the differences involved in annotating the noncoding variants (excluding UTR variants) these were prioritised using separate thresholds. Less experimental evidence is available to provide adequate annotation for noncoding variants. Therefore, noncoding variants were selected as those with more than 5 lines of evidence obtained from the ENCODE annotation and high LD (measured by D').

Coding and UTR variants were prioritised if they had more than 6 lines of evidence suggesting they may be functional and if they were in high LD with the GWAS variant (as measured by D'). For variants which were imputed successfully in the independent dataset, any which also had tentative association were put forward for functional assessment in the laboratory.

4.2.4. Genotyping prioritised variants

The prioritised coding and UTR variants were genotyped in an independent sample set to confirm the tentative LOAD association identified in the imputed dataset (table 4.1).

As initial genotyping work revealed the KASP assay as a more robust genotyping method (Braae et al. 2014), this was used for genotyping all

variants, apart from the frameshift variant rs201074739 as this was genotyped as part of C. Medway's Post Doc project using the Mayo Clinic (Jacksonville, Florida, USA) sample of 4050 cases and 4719 controls. Genotyping for the *EPHA1* variant rs34372369 was performed by the MSc. student Natalie Barker under my supervision.

Prior to developing a genotyping assay for a prioritised variant, the sample size needed for appropriate power to detect a disease association for the variant was estimated using QUANTO. Due to limited resources, only 1000 cases and 970 controls were available for genotyping. Therefore if this sample size would not provide adequate power to detect an association with disease, there would be little point in proceeding with the analysis. Indeed variants with a minor allele frequency less than 1% were excluded from association testing for this reason. Please Chapter 2, section 2.2.1 for a full account of power calculations.

Table 4.1. List of prioritised variants selected for association testing. Important qualifying criteria were more than 7 lines of evidence with the databases used (last column), apart from CD33 (as described in section 3.2.2.8.2). MAF is the CRISP calculated minor allele frequency in the NGS study. Ref refers to the reference allele and Alt is the alternative allele called in the database. If the variant is found in the 1000 Genomes database, the rs ID number is provided. The consequence of the variant as annotated by VEP is shown. Support is the number of lines of evidence from the prioritised section. AD shows the p-value for the variant from the imputed dataset. P is the power to detect an association given the sample size of 1000 cases and 970 controls, apart from the CD33 variant which assumes a sample size of 4000 cases and 4000 controls. Power assumes an odds ratio of 1.5.

Gene	**Chr**	**Location**	**MAF**	**Ref**	**Alt**	**rsID**	**Consequence**	**Support**	**AD**	**P (%)**
CD2AP	6	47445540	0.057	C	G	rs111766401	5_prime_UTR_ variant	7	0.03	90
EPHA1	7	143088823	0.064	C	T	rs1804527	synonymous_ variant	8	0.05	94
EPHA1	7	143092269	0.040	G	A	rs34372369	missense_variant	7	-	78
ZYX	7	143086010	0.012	C	T	rs73154206	missense_variant	7	0.03	74
CD33	19	51729104-51729107	0.024	CC GG	-	rs201074739	frameshift_variant, feature_truncation	0*	0.04	99

The Nottingham DNA bank Taqman/KASP genotyping plates with 1000 cases and 970 controls were used (See table 4.2 for sample demographics). Consent was previously obtained for all samples and a local ethics committee approved the study. DNA was extracted using a standard phenol-chloroform method (See Chapter 2, section 2.1.1).

The samples are plated out in 96-well optically clear plates with 20ng DNA in each well and with at least three blank wells with no DNA on each plate for the no-template controls (NTC). Additionally, a Sanger sequence-verified, heterozygous positive control for the variant being genotyped was included on each reaction plate. The majority of the plates used for the association study were generated as part of this validation project.

Table 4.2. Sample demographics for the Nottingham genotyping plates. The samples were obtained from 7 centres from UK and Europe, Bonn (Germany), Belfast, Manchester, Leeds, Nottingham, Oxford and Southampton (UK). The average age at diagnosis (cases) or sampling (controls) is indicated in the first column (Ave Age). The percentages of the three *APOE* epsilon alleles are shown as is the percentage of male and females in the study samples.

Samples	**N**	**Ave. Age (years)**	***APOE* allele (%)**			**Gender (%)**	
			E2	**E3**	**E4**	**M**	**F**
Cases	1000	74.6	4.8	61	34.2	40	60
Controls	970	71.3	9.1	76.6	14.3	47	53

For designing the KASP assay, the SNP of interest and 100bp upstream and downstream sequence was annotated with known variants using 1000 genomes browser and repetitive sequence was masked using RepeatMasker (http://www.repeatmasker.org/). The annotated sequence was submitted to LGC Limited via their SNP submission template. The custom KASP assay was then designed using the Kraken software system (LGC Limited) and validated with LGC's additional assay validation service.

KASP genotyping reactions were carried out in 10 µL total volumes containing 1X KASP Master Mix, 1X Custom KASP Assay and 20 ng DNA. Plates were cycled on a Veriti 96-Well Fast Thermal Cycler (Applied Biosystems) at the following conditions: 94°C for 15 min, then 10 cycles of 94°C for 20 sec, 61°C for 1 min (dropping by 0.6°C per cycle), then 26 cycles of 94°C for 20 sec, 55°C for 1 min followed by 9 cycles of 94°C for 20 sec and 57°C for 1 min. End point fluorescent readings were taken post-cycling on a MX3000P (Stratagene).

Following genotyping, the association with Alzheimer's disease was tested using a Fisher's exact test and also logistic regression in Plink v1.07 (Purcell et al. 2007) (see also Chapter 2, section 2.2.2). The association tests were run both without and with the covariates age of diagnosis (or age of sampling for controls), *APOE* genotype, centre where the samples originated and gender.

Fisher's exact test was selected as it is robust to small values in any comparison group. Very few or even no genotypes are expected in the homozygous alternative (mutant) group. Logistic regression was also run as it is able to determine the influence that each of the covariates has on the model and so provides further insight into the association.

4.3. Results

4.3.1. Variant annotation

For the 18 coding variants, functional predictions from VEP revealed seven missense mutations across the three GWAS genes, and an additional two missense mutations in zyxin (*ZYX*), the gene adjacent to *EPHA1* (table 4.3). No missense variants were found in *CD2AP*. Three of the coding variants are novel. The two suggestive splicing variants and all exonic variants were investigated using ESEfinder (table 4.3). The two splicing variants, *EPHA1* SNP, rs6967117 slightly reduces the binding affinity of serine/arginine-rich splicing factor 1 (SRSF1) (Wt score = 5.19 mut score = 4.76), while the *CD2AP* SNP, 6:47544253 A>G causes a SRSF5 site to be lost.

The 21 UTR variants are annotated in table 4.4. For these variants, the results of the RNA secondary structure predictions from RNAsnp are also presented. For the RNAsnp predictions, a p score <0.2 is indicated as suggestive of RNA folding (Sabarinathan et al. 2013b). One of the 5'UTR variants in *CD2AP*, rs1056434 shows a possible structural change (D=0.309, p=0.0682), and one of the 3'UTR variants in *CD2AP*, rs1043276 shows a significant structural change (D= 0.398 p=0.01), while several other 3'UTR variants in *CD2AP* (rs36077218, rs35361796 and rs35274349) and one in *CD33* (rs147493755) and in *ZYX* (rs144114027) are suggestive (p<0.2). Seven 3'UTR variants overlap with potential miRNA binding sites as predicted by TargetScan (ver 6.2), mirDB and MicroSNiPer. Three overlap with two miRNA binding sites, rs151064033 (miRNAs, hsa-miR-888-5p, hsa-miR-618) and rs141029774 (hsa-miR-200a, hsa-miR-141-3p) in *CD2AP* and rs1803254 (miRNAs, hsa-miR-374c-3p, hsa-miR-154-3p) in *CD33.*

The locations of the coding and UTR variants within each of the transcripts for the three GWAS genes are shown below in figure 4.1. The coding variants for *CD2AP* appear to cluster between exon 6 and 11, while for the remaining genes, the variants are more spread out.

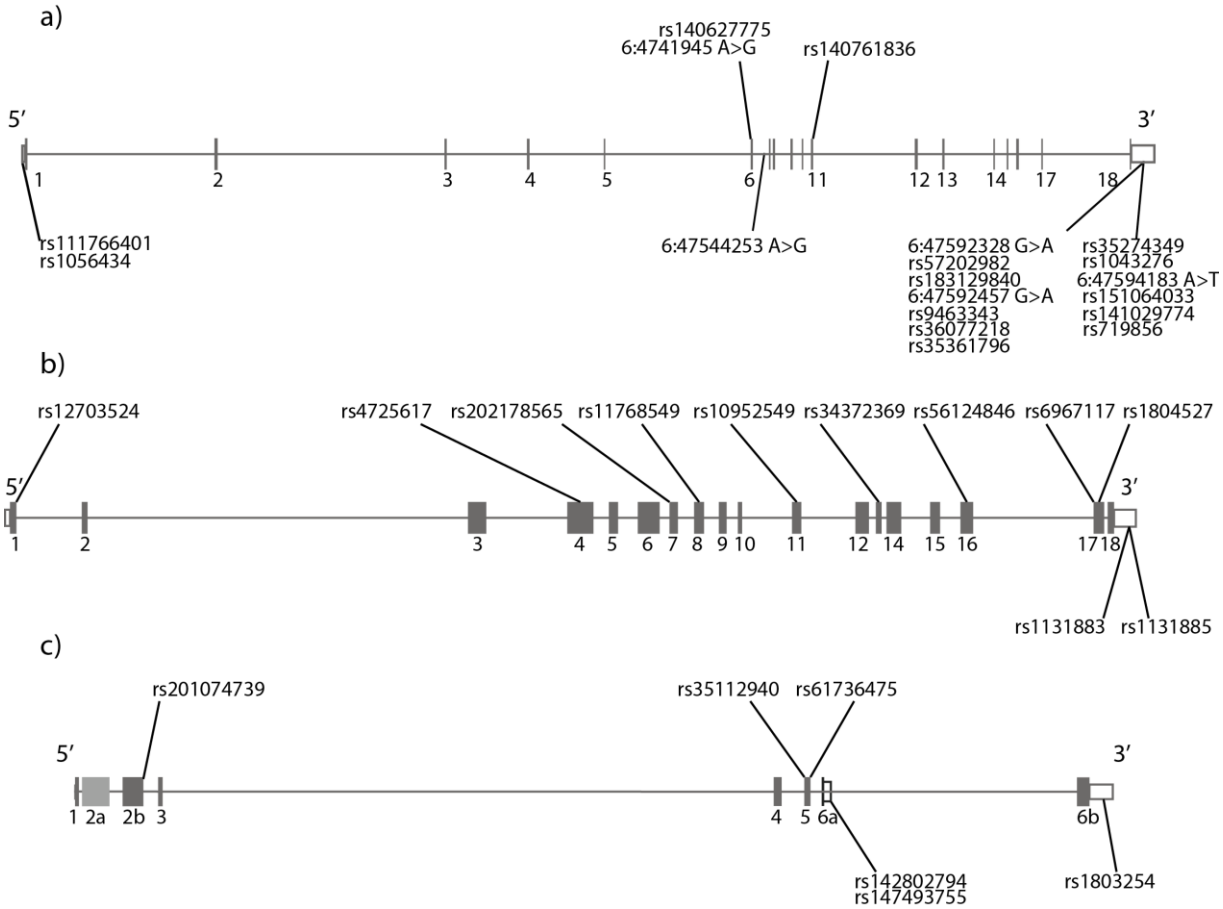

Figure 4.1. Overview of the variant positions in the validated transcripts for a) *CD2AP* (ENST00000359314), b) *EPHA1* (ENST00000275815) and c) *CD33*. For *CD33*, all three isoforms are shown on one diagram, as isoform 2 (ENST00000421133) differs from isoform 1 (ENST00000262262) by the exclusion of exon 2a, while isoform 3 (ENST00000391796) differs from the first two isoforms by having an alternative last exon and 3'UTR (exon 6a).

Table 4.3. *In silico* functional predictions of the coding and splice variants found in *CD2AP, EPHA1, ZYX* and *CD33*. The reference and alternative alleles for *EPHA1* refer to the forward strand (note that this gene is coded from the reverse strand). The 1000 genomes allele frequency for the Northern European population from Utah (CEU) is shown under 1kG CEU MAF. The GWAS SNPs used for LD calculations are rs9349407 (*CD2AP*), rs11767557 (*EPHA1* and *ZYX*) and rs3865444 (*CD33*).

Gene	chr	Location	Ref	Alt	Conservation		CRISP MAF	1kG CEU MAF	LD (D')	Consequence	Co-located Variation	Exon	Protein change	Pathogenicity (PolyPhen)	Splicing (ESEfinder)
					PhyloP	PhastCons									
CD2AP	6	47541945	A	G	1.05949	1	0.007	-	-	synonymous	-	6	T	-	-
	6	47541954	C	T	-0.157961	0.96063	0.011	0.01	0.278	synonymous	rs140627775	6	S	-	-
	6	47544253	A	G	-1.95791	0	0.005	-	-	splice variant, intron	-	Before 7	-	-	SRSF5 site lost
	6	47549785	A	G	0.0483307	0.511811	0.006	-	-	synonymous	rs140761836	11	L	-	-
EPHA1	7	143088823	T	C	-0.451094	0.661417	0.937	0.91	0.748	synonymous	rs1804527	17	P	-	SRSF2 and SRSF6 sites introduced
	7	143088867	T	C,A	0.652213	0.992126	0.060	0.09	0.748	splice variant, missense	rs6967117	17	M/L	Benign	Reduces SRSF1 site score by 0.49
	7	143090844	G	A	1.36318	0.992126	0.004	-	-	synonymous	rs56124846	16	H	-	-
	7	143092269	G	A	2.63561	1	0.040	0.05	-1	missense	rs34372369	13	P/L	Probably damaging	SRSF1 site lost
	7	143093538	G	A	0.219433	0.992126	0.794	0.78	0.783	synonymous	rs10952549	11	L	-	SRSF1 site lost
	7	143095153	C	T	2.44459	0.976378	0.016	0.01	1	missense	rs11768549	8	R/Q	Benign	-
	7	143095499	G	A	2.50739	0.913386	0.007	-	-	missense	rs202178565	7	P/L	Probably damaging	SRSF6 site introduced
	7	143097100	A	G	0.54152	0	0.937	0.91	0.874	missense	rs4725617	4	V/A	Benign	-
	7	143105830	C	A	0.180598	0.992126	0.559	0.39	-0.856	synonymous	rs12703524	1	A	-	-
ZYX	7	143085969	G	A	2.485	0.992126	0.08	-	-	missense	-	8	R/H	Possibly damaging	SRSF1 site lost
	7	143086010	C	T	0.170157	0.952756	0.012	0.03	0.896	missense	rs73154206	8	R/W	Benign	-
CD33	19	51729104-51729107	CCGG	-	1.65606	0	0.02	-	-1	frameshift (deletion)	rs201074739	2b	-	-	-
	19	51738917	G	A	0.139402	0	0.23	0.22	0.981	missense	rs35112940	6	G/R	Benign	SRSF2 site lost
	19	51738920	T	C	-1.37045	0	0.01	0.01	-1	missense	rs61736475	6	S/P	Benign	-

Table 4.4. 5' and 3' UTR variants called by CRISP in *CD2AP, EPHA1, ZYX* and *CD33* annotated using VEP and an in house script to incorporate ENCODE data. The reference and alternative alleles for *EPHA1* refer to the forward strand orientation. The first two 3'UTR variants for *CD33* refer to transcript isoform 3 ENST00000391796), while the last refers to isoform 1 (ENST00000262262). The 1000 genomes allele frequency for the Northern European population from Utah (CEU) is shown under 1kG CEU MAF. The GWAS SNPs used for LD calculations are rs9349407 (*CD2AP*), rs11767557 (*EPHA1* and *ZYX*) and rs3865444 (*CD33*). miRNA binding was assessed using TargetScan (ver 6.2), mirDB and MicroSNiPer. RNA folding was assessed using RNAsnp.

Gene	chr	Location	Ref	Alt	Conservation		CRISP MAF	1kG CEU MAF	LD (D')	Position	Co-located Variation	cDNA position	Transcription factor binding sites	miRNA binding	RNA folding
					PhyloP	Phast Cons									
CD2AP	6	47445540	C	G	0.549	0	0.057	0.06	-1	5'UTR	rs111766401	16	13 including Pol2, Pol2-4H8, TBP	-	D= 0.136 (p=0.2130)
	6	47445789	A	C	0.808	0	0.550	0.58	-1	5'UTR	rs1056434	265	19 including Pol2, Pol2-4H8,TBP	-	D= 0.309 (p=0.068*)
	6	47592328	G	A	0.707	1	0.010	0.93	-	3'UTR	-	2741	-	None	D= 0.003 (p=0.9134)
	6	47592330	T	A	-1.655	1	0.280	-	-	3'UTR	rs57202982	2743	-	None	D= 0.034 (p=0.5273)
	6	47592415	C	G	-1.719	0.055	0.011	-	-	3'UTR	rs183129840	2828	-	None	D= 0.059 (p=0.3657)
	6	47592457	G	A	0.674	0	0.029	-	-	3'UTR	-	2870	-	None	D= 0.097 (p=0.2158)
	6	47592459	A	G	0.739	0	0.572	-	-	3'UTR	rs9463343	2872	-	None	D= 0.087 (p=0.2450)
	6	47592461	A	G	-0.749	0	0.294	-	-	3'UTR	rs36077218	2874	-	None	D= 0.105 (p=0.1942)
	6	47592463	A	G	0.675	0	0.267	-	-	3'UTR	rs35361796	2876	-	None	D= 0.108 (p=0.1990)
	6	47592465	A	G	-0.878	0	0.291	-	-	3'UTR	rs35274349	2878	-	None	D= 0.137 (p=0.1415)
	6	47594002	T	C	-1.155	0	0.694	0.7	1	3'UTR	rs1043276	4415	-	None	D= 0.398 (p=0.010*)
	6	47594183	A	T	0.516	0	0.065	-	-	3'UTR	-	4596	-	None	D= 0.021 (p=0.6312)
	6	47594204	A	T	0.516	0	0.019	0.01	-	3'UTR	rs151064033	4617	-	hsa-miR-888-5p, hsa-miR-618	D= 0.023 (p=0.6245)
	6	47594318	G	A	0.545	0	0.011	0.002	-	3'UTR	rs141029774	4731	-	hsa-miR-200a, hsa-miR-141-3p	D= 0.002 (p=0.9484)
	6	47594722	G	A	0.545	0.354	0.247	0.24	-1	3'UTR	rs719856	5135	CEBPB	hsa-miR-194-5p	D= 0 (p=0.9975)

Table 4.4 continued.

Gene	chr	Location	Ref	Alt	Conservation		CRISP MAF	1kG CEU MAF	LD (D')	Position	Co-located Variation	cDNA position	Transcription factor binding sites	miRNA binding	RNA folding
					PhyloP	Phast Cons									
EPHA1	7	143088526	T	C	0.395	0	0.440	0.31	-0.792	3'UTR	rs1131885	3042	Pol2, Pol2-4H8	None	D=0.026 (p=0.651)
	7	143088531	C	T	0.395	0	0.030	0.05	0.851	3'UTR	rs1131883	3037	Pol2, Pol2-4H8	hsa-miR-3622b-5p	D=0.006 (p=0.909)
ZYX	7	143087934	G	A	0.075	0	0.005	0.001	-	3'UTR	rs188278148	2223	-	None	D=0.0522 (p=0.5172)
	7	143087956	G	A	-0.049	0.015	0.033	0.03	-1	3'UTR	rs11552744	2245	-	hsa-miR-1915-3p	D=0.0019 (p=0.9604)
	7	143088076	C	T	-0.542	0	0.046	0.04	0.851	3'UTR	rs144114027	2365	-	None	D=0.2111 (p=0.1188)
	7	143088085	G	A	-0.357	0	0.169	0.22	0.187	3'UTR	-	2374	-	None	D=0.0043 (p=0.9264)
CD33	19	51739189	G	A	0.434	0	0.008	0.01	1	3'UTR	rs142802794	1076	-	None	D=0.0382 (p=0.5514)
	19	51739215	T	C	0.434	0	0.011	0.02	-1	3'UTR	rs147493755	1102	-	hsa-miR-4474-5p	D=0.0811 (p=0.1707)
	19	51743144	G	C	0.778	0	0.039	0.08	0.775	3'UTR	rs1803254	1136	-	hsa-miR-374c-3p, hsa-miR-154-3p	D=0.097 (p=0.2578)

Table 4.5. Noncoding variants called by CRISP with more than 6 lines of functional evidence as annotated using VEP and an in house script to incorporate ENCODE data. The reference and alternative alleles for *EPHA1* refer to the forward strand orientation. The GWAS SNPs used for LD calculations are rs9349407 (*CD2AP*), rs11767557 (*EPHA1*) and rs3865444 (*CD33*). Transcription factor binding sites indicated by CHIP-seq (ENCODE) and TRANSFAC database. DNase I hypersensitivity cluster score shown under DNase. CpG islands (CpG column) were measured in SK-N-SH cells using the Methyl 450 array (indicated as Meth450) and reduced representation bisulfite sequencing (indicated by RRBS). Methylation (1,2,3) indicates the cell line and whether methylation was found at H3K4me1, H3K4me2 or H3K4me3. While the last column, H3K27ac indicates the cell line and whether this lysine residue was acetylated.

Gene	**chr**	**Location**	**Ref**	**Alt**	**Conservation**		**CRISP MAF**	**1kG CEU MAF**	**LD (D')**	**Gene location**	**Co-located Variation**	**Transcription factor binding sites**	**TRANSFAC**	**DNase**	**CpG**	**Methylation (1,2,3)**	**H3K27ac**
					Phylo P	**Phast Cons**											
CD2AP	6	47444749	G	T	-1.361	0	0.009	0[+]	-1	upstream	rs75968487	-	-	8	(Meth450)	NHA (1,2,3), HepG2 (1,2,3)	NHA, HepG2
	6	47487213	C	G	-1.743	0	0.004	0.01	-	intron 2-3	rs140627775	11 including FOXA2	-	9	(Meth450)	HepG2 (1,2)	HepG2
	6	47544253	A	G	-3.321	0	0.005	-	-	splice region, intron 6-7	-	-	-	-	-	HepG2 (1)	-
EPHA1	7	143085969	G	A	3.748	1	0.008	-	-	downstream	-	Pol2	Tax/Creb	23	-	NHA (1,2)	-
	7	143086010	C	T	0.942	0.984		0.03	0.831	downstream	rs73154206	-	myogenin/ NF-1, NFY	23	(RRBS)	NHA (1,2)	-
ZYX	7	143090844	G	A	-0.223	0	0.004	<0.01	-	downstream	rs56124846	HA-E2F1	-	69	-	NHA (1) HepG2(1,2,3)	HepG2
	7	143092269	G	A	1.730	1	0.041	0.05	-0.888	downstream	rs34372369	GATA-2, c-Fos	-	57	-	NHA(1), HepG2 (1,2,3)	HepG2
EPHA1 -AS	7	143106086	A	T	0.429	0.047	0.009	-	-	intron 1-2	-	7 including CTCF	-	41	-	NHA (1,2) HepG2 (1,2,3)	-
CD33	19	51729459	C	T	-1.878	0	0.004	-	-	Intron 3-4	-	PU.1	-	3	-	-	-

[+]rs75968487 is not found in European populations in the 1000 Genomes project.

Given the large number of intronic and noncoding variants (793), functional predictions for all variants are not shown here. The aim of the study is to predict rare (MAF < 5%) variants which may be responsible for causing disease. The functional predictions for the nine noncoding variants with MAF <5 % which show six or more lines of annotation are shown in table 4.5. Given the minimal functional support available for *CD33*, the only noncoding variant with any annotation is shown *CD33* 19:51729459 C>T.

The *CD2AP* intronic variant rs140627775 overlaps with an area where 11 transcription factor binding sites (TFBS) are found (table 4.5). Four variants (rs75968487 in *CD2AP*, rs56124846 and rs34372369 in *ZYX* and *EPHA1-AS* 7:143106086 A>T) overlap regions of potential promoter or enhancer activity, determined by the methylation marks at H3K4me1, H3K4me2 and H3K4me3, indicating the fourth lysine residue on histone H3 which can be mono- (me1), di- (me2) or tri- (me3) methylated. Three of these (rs75968487, rs56124846 and rs34372369) also overlap regions of active enhancer activity shown by H3K27ac (table 4.5).

4.3.2. Imputation and association testing variants

As potential positive controls, the initial results of the imputation were assessed by determining the association of the GWAS SNPs identified in the 2011 GWAS studies (table 4.6). None were significantly associated with LOAD.

The results of the association test for the coding and UTR variants are shown in table 4.7. Several variants could not be imputed. Of those that did successfully impute, however, all had good info scores. While none of the variants were associated with *p*-value that would withstand correction for multiple testing ($p <= 5 \times 10^{-8}$), several showed tentative association ($p < 0.05$).

As there are well over 700 noncoding variants, the results of the association test for this class of variants is again restricted to those with MAF <5% and more than 6 lines of functional annotation, apart from *CD33* as for table 4.5. The association results for these noncoding variants are presented in table 4.8. Only three variants were successfully imputed, and only the *EPHA1* SNP, rs73154206, showed tentative association with LOAD (OR = 0.340 (0.171-0.675), p = 0.05172).

Table 4.6. Association between common, GWAS tag SNPs and AD in the imputed dataset. The imputed MAF for AD cases and controls, OR and *p*-values were calculated using the frequentist association method, assuming an additive model, in snptest (v2).

Gene	**Variant**	**Info score**	**MAF AD**	**MAF Controls**	**OR (95% CI)**	***p*-value**
CD2AP	rs9349407	0.996	0.274	0.277	0.985 (0.911-1.06)	0.607
EPHA1	rs11767557	0.999	0.189	0.208	0.889 (0.814-0.972)	0.085
CD33	rs3865444	0.990	0.248	0.314	0.721 (0.661-0.786)	0.114

Table 4.7. Association testing coding and UTR variants in the imputed dataset. Results are given for Impute2 imputation and snptest association testing, using the frequentist method with an additive inheritance pattern and controlling for centre. The genome location, reference (Ref) and alternative (Alt) allele and SNP rs identifier are shown, as is the imputed MAF for AD cases and controls, OR and *p*-values. Tentatively associated SNPs are indicated in bold.

Variants	**Location**	**Ref**	**Alt**	**rs number**	**Info score**	**AD MAF**	**Control MAF**	**OR (95% CI)**	***p* value**
CD2AP coding	47541945	A	G	-	Not imputed				
	47541954	C	T	rs140627775	0.146284	0.00051	0.00045	1.137 (0.229-5.63)	0.694
	47544253	A	G	-	Not imputed				
	47549785	A	G	rs140761836	Not imputed				
CD2AP 5'UTR	**47445540**	**C**	**G**	**rs111766401**	**0.990227**	**0.0729**	**0.0706**	**1.035 (0.904-1.185)**	**0.0324**
	47445789	A	C	rs1056434	0.990057	0.41344	0.41613	1.011 (0.940-1.087)	0.1381
CD2AP 3'UTR	47592328	G	A	-	Not imputed				
	47592330	T	A	rs57202982	Not imputed				
	47592415	C	G	rs183129840	Not imputed				
	47592457	G	A	-	Not imputed				
	47592459	A	G	rs9463343	Not imputed				
	47592461	A	G	rs36077218	Not imputed				
	47592463	A	G	rs35361796	Not imputed				
	47592465	A	G	rs35274349	Not imputed				
	47594002	**T**	**C**	**rs1043276**	**0.994519**	**0.284509**	**0.290059**	**1.027 (0.949-1.111)**	**0.0459**
	47594183	A	T	-	Not imputed				
	47594204	A	T	rs151064033	Not imputed – insufficient allele counts				
	47594318	G	A	rs141029774	Not imputed – insufficient allele counts				
	47594722	**G**	**A**	**rs719856**	**0.982445**	**0.216575**	**0.23078**	**0.921 (0.846-1.003)**	**0.0012**
EPHA1 coding	143088823	T	C	rs1804527	0.999847	0.0615272	0.0682691	1.118 (0.968-1.289)	0.0588
	143088867	**T**	**C,A**	**rs6967117**	**0.996977**	**0.0584226**	**0.0682497**	**1.180 (1.020-1.366)**	**0.04401**
	143090844	G	A	rs56124846	Not imputed				
	143092269	G	A	rs34372369	0.96768	0.0246815	0.0472601	0.510 (0.409-0.636)	0.48374
	143093538	G	A	rs10952549	0.973268	0.136291	0.190895	1.495 (1.323-1.689)	0.63358
	143095153	C	T	rs11768549	0.946699	0.0065162	0.0161729	0.399 (0.265-0.599)	0.80583
	143095499	G	A	rs202178565	Not imputed				
	143097100	A	G	rs4725617	0.999794	0.0685562	0.0749503	1.101 (0.961-1.261)	0.0721
	143105830	C	A	rs12703524	0.966638	0.455212	0.4352	1.084 (0.995-1.181)	0.8964

Table 4.7 continued.

Variants	Location	Ref	Alt	rs number	Info score	AD MAF	Control MAF	OR (95% CI)	*p* value
EPHA1 3'UTR	143088526	T	C	rs1131885	0.96457	0.293873	0.365973	0.721 (0.649-0.800)	0.62272
	143088531	C	T	rs1131883	0.923087	0.0053648	0.0241339	0.218 (0.139-0.343)	0.28151
ZYX coding	143085969	G	A	-	Not imputed				
	143086010	**C**	**T**	**rs73154206**	**0.788445**	**0.0023328**	**0.0068269**	**0.340 (0.171-0.675)**	**0.05172**
ZYX 3'UTR	143087934	G	A	rs188278148	Not imputed				
	143087956	**G**	**A**	**rs11552744**	**0.865256**	**0.0071963**	**0.0214589**	**0.331 (0.219-0.498)**	**0.03791**
	143088076	C	T	rs144114027	0.849442	0.0113461	0.0198365	0.567 (0.411-0.782)	0.52203
	143088085	G	A	-	0.974277	0.159639	0.207483	0.725 (0.652-0.807)	0.96977
CD33 coding	**51729104-51729107**	**CCGG**	**-**	**rs201074739**	**0.6309**	**0.0005216**	**0.0028726**	**0.181 (0.044-0.752)**	**0.03962**
	51738917	G	A	rs35112940	0.95535	0.165103	0.199506	0.793 (0.725-0.879)	0.4346
	51738920	T	C	rs61736475	0.831117	0.0005216	0.014166	0.411 (0.261-0.645)	0.1198
CD33 3'UTR	51739189	G	A	rs142802794	Not imputed				
	51739215	T	C	rs147493755	Not imputed				
	51743144	G	C	rs1803254	0.944084	0.0155797	0.0594916	0.250 (0.184-0.341)	0.99306

Table 4.8. Association testing noncoding variants with functional support in the imputed dataset. Results are given for the Impute2 imputation and snptest association testing, using the frequentist method with an additive inheritance pattern and controlling for centre. The genome location, reference (Ref) and alternative (Alt) allele and SNP rs identifier are shown, as is the imputed MAF for AD cases and controls, OR and *p*-values. The tentatively associated SNP is indicated in bold.

Variants	**Location**	**Ref**	**Alt**	**rs number**	**Info score**	**AD MAF**	**Control MAF**	**OR (95% CI)**	***p* value**
CD2AP	47444749	G	T	rs75968487	0.999521	0.00243	0.00145	0.536 (0.227-1.271)	0.11479
	47487213	C	G	rs140627775	Not imputed				
	47544253	A	G	-	Not imputed				
EPHA1	143085969	G	A	-	Not imputed				
	143086010	**C**	**T**	**rs73154206**	**0.788445**	**0.0023328**	**0.00682695**	**0.340 (0.171-0.675)**	**0.05172**
ZYX	143090844	G	A	rs56124846	Not imputed				
	143092269	G	A	rs34372369	0.96768	0.0246815	0.0472601	0.510 (0.409-0.635)	0.48374
EPHA1-AS	143106086	A	T	-	Not imputed				
CD33	51729459	C	T	-	Not imputed				

4.3.3. Variant prioritisation

The functional annotation was combined with the results of the associated imputed dataset to generate a final prioritised list of variants which could be put forwards for experimental validation in the laboratory. Variants which were put forward for laboratory investigation and which were investigated in other chapters in this thesis are noted in the last column in both tables.

Again, as differing levels of annotation are available for the noncoding and coding and UTR datasets, the variants for these regions are presented in two separate tables (table 4.9 for the noncoding variants and table 4.10 for the coding and UTR variants).

Table 4.9. Prioritised noncoding variants following annotation for functionality and association in the imputed dataset. Samples identified in bold for *EPHA1* and *ZYX* are also found in the coding table as these two genes overlap (i.e. upstream variants for one gene are found in the other gene). Ref and Alt refer to the reference and alternative alleles, with MAF indicating the minor allele frequency called by CRISP. VEP defined variant location relative to the gene is shown under consequence. Number of lines of evidence of functionality (Func.) and p-value of association with LOAD in the imputed dataset (AD) are also shown. If lab work on the variant was written up in this thesis, the chapter number is indicated in the last column (Lab?). Two variants were prioritised with less than 6 lines of functionality as indicated by *.

Gene	**Location**	**Ref**	**Alt**	**MAF**	**rsID**	**Consequence**	**Func.**	**AD**	**Lab?**
CD2AP	47444749	G	T	0.009	rs75968487	upstream_gene_variant	6	-	-
	47487213	C	G	0.004	-	intron_variant	6	-	-
	47544253	A	G	0.005	-	splice_region_variant intron_variant	2*	-	5
EPHA1	143085969	G	A	0.008	-	downstream_gene_variant	7	-	-
	143086010	C	T	0.012	rs73154206	downstream_gene_variant	7	0.05	4
ZYX	143090844	G	A	0.004	rs56124846	downstream_gene_variant	7	-	-
	143092269	G	A	0.041	rs34372369	downstream_gene_variant	7	-	4
EPHA1-AS	143106086	A	T	0.009	-	intron_variant, nc_transcript_variant	7	-	-
CD33	51729459	C	T	0.004	-	intron_variant	2*	-	-

Table 4.10. Prioritised coding and UTR variants following annotation for functionality and checking association of variants which were imputed in the GWAS dataset. Samples identified in bold for *EPHA1* and *ZYX* are also found in the noncoding table as these two genes overlap (i.e. upstream variants for one gene are found in the other gene). Ref and Alt refer to the reference and alternative alleles, with MAF indicating the minor allele frequency called by CRISP. VEP defined variant location relative to the gene is shown under consequence. Number of lines of evidence of functionality (Func.) and p-value of association with LOAD in the imputed dataset (AD) are also shown. If lab work on the variant was written up in this thesis, the chapter number is indicated in the last column (Lab?).

Gene	Location	Ref	Alt	MAF	rsID	Consequence	Protein	Codons	Domain	Func.	AD	Lab?
CD2AP	47445540	C	G	0.057	rs111766401	5_prime_UTR_variant	-	-	-	7	0.03	4
	47594204	A	T	0.01	rs151064033	3_prime_UTR_variant	-	-	-	6	-	6
	47594318	G	A	0.002	rs141029774	3_prime_UTR_variant	-	-	-	6	-	6
EPHA1	143088823	C	T	0.06	rs1804527	synonymous_variant	914 (P)	ccA/ccG	SAM	8	0.05	4,5
	143088867	C	T	0.06	rs6967117	missense_variant, splice_region_variant	M900L	Atg/Ttg	Cytoplasmic	7	0.04	5
	143090844	G	A	0.004	**rs56124846**	synonymous_variant	872 (H)	caC/caT	Kinase	7	-	-
	143092269	G	A	0.041	**rs34372369**	missense_variant	P697L	cCg/cTg	Kinase	7	-	4
	143095153	C	T	0.016	rs11768549	missense_variant	R492Q	cGg/cAg	Fibronectin type-III 2	6	-	-
	143088531	C	T	0.030	rs1131883	3_prime_UTR_variant	-	-	-	6	-	6
ZYX	143085969	G	A	0.008	-	missense_variant	R475H	cGc/cAc	LIM zinc-binding 2	7	-	-
	143086010	C	T	0.012	**rs73154206**	missense_variant	R489W	Cgg/Tgg	LIM zinc-binding 2	7	0.03	4
	143087934	G	A	0.006	rs188278148	3_prime_UTR_variant	-	-	-	6	-	-
	143087956	G	A	0.033	rs11552744	3_prime_UTR_variant	-	-	-	6	0.03	6
	143088076	C	T	0.047	rs144114027	3_prime_UTR_variant	-	-	-	6	-	-
CD33	51729104-51729107	CCGG	-	0.024	rs201074739	frameshift_variant, feature_truncation	155-156	-	Ig-like V-set	0	0.04	4, 6
	51738920	T	C	0.007	rs61736475	missense_variant	S305P	Tca/Cca	Cytoplasmic	0	-	-
	51743144	G	C	0.039	rs1803254	3_prime_UTR_variant	-	-	-	1	-	6

4.3.4. Association testing prioritised variants

Following validation with a small number of general DNA samples and a known Sanger-verified heterozygous sample, all KASP assays were run with 1970 DNA samples - the entire KASP genotyping sample set. A number of samples failed for each of the assays possibly due to poor DNA quality or due to inaccurate pipetting of the samples when preparing the genotyping plates. Details of the samples run for each of the assays are provided in table 4.11.

The frameshift variant rs201074739 was genotyped by C. Medway as part of his Post Doc work at the Mayo Clinic (Jacksonville, Florida, USA). Therefore in the interest of not repeating the experiment, the results of the Sequenom genotyping for this variant is presented here.

Table 4.11. The final results from the genotyping study. The minor allele frequency (MAF) for the cases and the controls in the genotyped samples and the number of cases and controls successfully genotyped for each sample. The CD33 variant was genotyped at the Mayo Clinic* on a different set of LOAD samples.

Gene	**Variant**	**MAF cases**	**MAF controls**	**N cases**	**N controls**
CD2AP	rs111766401	0.068	0.071	994	968
EPHA1	rs1804527	0.058	0.066	992	970
EPHA1	rs34372369	0.050	0.065	1012[+]	970
ZYX	rs73154206	0.018	0.011	979	968
*CD33**	rs201074739	0.024	0.022	4050	4719

[+]More than 1000 samples were genotyped for this variant due to 12 additional initial LOAD samples used to validate the assay.

None of the genotyped variants were found to be associated with LOAD in the Nottingham KASP genotyping samples (or the Mayo Clinic samples for *CD33* rs201074739) through association testing using either Fisher's exact test (table 4.12) or logistic regression (table 4.13).

The association did not reach significance with the addition of covariates for either test (table 4.12, table 4.13). For two variants, EPHA1 rs34372369 and ZYX rs73154206 the p-value showed slight improvement with the addition of covariates in the Fisher's exact test (table 4.12). However, this was not reflected in the results of the logistic regression (table 4.13), indicating possible confounding in the Fisher's exact test results through one or more of the covariates, age of diagnosis (or age of sampling for controls), *APOE* genotype, centre where the samples originated and gender.

Table 4.12. Results of the Fishers exact association test without and with covariates (+cov). Covariates included were age of diagnosis (or sampling for controls), *APOE* genotype, centre of sample origin and gender. The odds ratio (OR) and 95% confidence interval (95 CI) are shown. The addition of covariates does not change the *CD33* variant OR or the p-value.

Gene	**Variant**	**OR (95 CI)**	**p value**	**+cov OR (95 CI)**	**+cov p value**
CD2AP	rs111766401	0.97 (0.75 - 1.24)	0.849	0.96 (0.74-1.24)	0.791
EPHA1	rs1804527	0.88 (0.68 - 1.14)	0.352	0.85 (0.65-1.12)	0.265
EPHA1	rs34372369	0.82 (0.63 - 1.08)	0.168	0.75 (0.14-0.57)	0.059
ZYX	rs73154206	1.55 (0.89 - 2.69)	0.134	1.62 (0.93-2.83)	0.094
CD33	rs201074739	1.09 (0.89-1.33)	0.393	1.09 (0.89-1.33)	0.393

Table 4.13. Results of the logistic regression for the variants without and with covariates (+cov). Covariates included were age of diagnosis (or sampling for controls), *APOE* genotype, centre of sample origin and gender. The odds ratio (OR) and 95% confidence interval (95 CI) are shown.

Gene	**Variant**	**OR (95 CI)**	**p value**	**+cov OR (95 CI)**	**+cov p value**
CD2AP	rs111766401	0.95 (0.75-1.22)	0.709	1.02 (0.77-1.34)	0.868
EPHA1	rs1804527	0.88 (0.68-1.14)	0.349	0.82 (0.61-1.11)	0.207
EPHA1	rs34372369	0.82 (0.63-1.08)	0.155	0.83 (0.60-1.13)	0.237
ZYX	rs73154206	1.56 (0.90-2.71)	0.111	1.43 (0.77-2.65)	0.252
CD33	rs201074739	1.09 (0.89-1.32)	0.390	1.1 (0.88-1.36)	0.388

4.4. Discussion

In this chapter, variants identified in the NGS study (Chapter 3) were annotated for functionality. Additionally, an independent GWAS dataset was imputed to test the identified and potentially functional variants for association with LOAD. Prioritised variants were put forward for additional functional study in Chapters 5 (splicing) and 6 (allelic expression imbalance). Four coding and one 5'UTR variant which had been put forward for further investigation were genotyped in an independent sample set to validate potential association with LOAD.

4.4.1. Functional variants of interest

In silico functional annotation of variants revealed 17 coding and splice variants and 21 variants in the UTRs of the genes. No missense variants were detected for *CD2AP* and although it has the largest gene sequence, only four synonymous variants were detected. This is probably because a large proportion of the gene is intronic (98.7% of 149kbp).

Of the five missense variants found for *EPHA1*, only two (rs34372369 and rs202178565) were predicted to be probably damaging by PolyPhen. Both are

rare (MAF <0.05) and in areas of high conservation (table 4.3). rs34372369 falls in the protein kinase domain of the protein, while rs202178565 falls in one of the fibronectin type III domains (The UniProt Consortium 2012). It is possible that both these variants could influence the functioning of the receptor making them both variants worthy of further investigation.

The two missense variants in *ZYX* are both found in exon 8, in the LIM domain of the protein, although only one is predicted possibly damaging (table 4.3). If the alternative alleles disrupt this domain, the variants may have an effect on the actin-modelling potential of the protein. The benign variant, rs73154206 is found at nearly twice the allele frequency in the 1kG population than in the NGS case population (table 4.3), possibly indicating a protective role. This variant was therefore selected for genotyping in the independent dataset to confirm the suspected disease association (see section 4.4.4).

The two splice variants are both predicted to influence binding of splicing regulatory proteins, with the intronic *CD2AP* 6:47544253A>G destroying an SRSF5 binding site, and the exonic *EPHA1* variant rs6967117 having a weak effect on an SRSF1 binding, however these results will need to be verified experimentally (Cartegni et al. 2003).

The only damaging coding variant in *CD33* is the frameshift variant, rs201074739, which is a four bp deletion in exon 2b of the gene (table 4.3). If the deletion transcript is translated it is predicted to produce a truncated protein that is missing the functional immunoglobulin (IG) and transmembrane (TM) domains (figure 4.2). Most mutations which are predicted to produce premature truncated proteins are targeted by nonsense mediated decay (Chang et al. 2007). Therefore to determine what effect, if any, this deletion may have on CD33, further investigation should be performed *in vivo*.

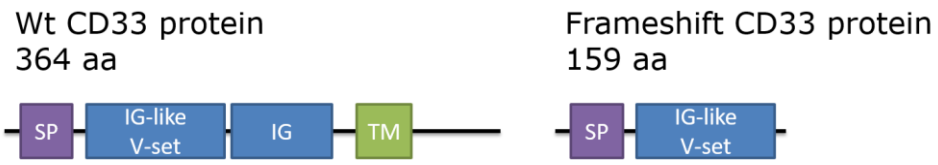

Figure 4.2 Predicted protein from the wildtype (Wt) CD33 isoform 1 and predicted protein from the same isoform but with the 4bp deletion (CCGG/-) rs201074739. The purple box, SP, is the signalling peptide, while the two IG and http://www.ebi.ac.uk/interpro/search/sequence-search domains are shown in blue boxes and the transmembrane domain (TM) in green. Predictions were made using ExPASy (http://web.expasy.org/translate/) and InterProScan (www.ebi.ac.uk/interpro/search/sequence-search) (both accessed May 2013).

CD2AP contained the most UTR variants (15, see table 4.4). The two 5'UTR variants fall in binding sites for several transcription factors, including various Pol2 subunits and TBP (TATA-binding protein) (table 4.4), and so could disrupt or modify the binding affinity of transcription factors at these sites. Additionally, the suggestive role of the common SNP, rs1056434, in changing the folding of the RNA at this site could potentially affect proteins binding at this location (Wan et al. 2011).

Of the 13 3'UTR variants in *CD2AP*, only three overlapped potential miRNA binding sites, rs151064033, rs141029774 and rs719856. However, detailed investigation into the miRNA regulation of *CD2AP* has not been undertaken, and given the size of the 3'UTR (which accounts for 56% of the coding DNA sequence) it is possible that there are other miRNA sites present that have not yet been added to the bioinformatics databases. Long 3'UTRs are commonly found in transcripts expressed in the brain, and it is suggested that they play a role in UTR based regulation (Ramsköld et al. 2009).

While over 700 noncoding variants were called, only nine had a MAF less than 5% and more than 6 lines of evidence, noted as overlap with a potential functional site following annotation (table 4.5). Two of these noncoding variants from the *EPHA1* gene loci were actually coding variants in *ZYX*, and two of the noncoding variants in *ZYX* were coding variants in *EPHA1*. The 3'UTR for *EPHA1* overlaps with the 3'UTR for *ZYX* as *EPHA1* is coded in the reverse orientation (see figure 1.4 and 1.5 in Chapter 1). Therefore the high scores for that area may be due to the functional constraints on the coding sequence for *ZYX* rather than on the noncoding region downstream of

EPHA1. Interestingly, as many as 13% of the genes in the human genome have been found to be overlapping in some manner (Makalowska et al. 2005).

The additional ENCODE annotations provided for the noncoding variants included methylation data for histone H3 at the fourth lysine residue (K4) and acetylation data at the 27th lysine residue (K27) from relevant cell lines, NHA and HepG2. Identifying these histone marks in combination allows the prediction of potential promoters or enhancers through H3K4 mono-, di- and trimethylation status, while the acetylation of H3K27 shows whether the site is actively being transcribed (Creyghton et al. 2010). Three variants (rs75968487 in *CD2AP* and both *ZYX* variants, rs56124846 and rs34372369) overlapped sites of active enhancer sequence as indicated by H3K4 and H3K27 histone marks in NHA and HepG2 cells (table 4.5). However, none of the variants had a DNase I hypersensitivity cluster score higher than 100. The DNase I hypersensitivity clusters combined data from 125 cell lines with a high score (above 100) indicating DNA which was "open" i.e. accessible to transcription factors and so sensitive to DNase I degradation.

4.4.2. Imputed associated variants of interest

The imputed dataset did not replicate the association of any of the GWAS SNPs for the three genes (table 4.6), indicating insufficient power to detect these associations, however the high info scores (all around 0.99) show that the variants have imputed well. The results of the association test should be interpreted with caution, as imputing rare variants (defined in this study as variants with MAF <=5%) is notoriously difficult (Zheng et al. 2012).

The 5'UTR SNP in *CD2AP*, rs111766401, is almost associated with AD (table 4.7), and is in linkage with the GWAS SNP (D' = -1, table 4.4). Only two 3'UTR SNPs were successfully imputed, rs1043276 and rs719856, both suggested to be associated, although the odds ratios spanned 1 (table 4.7). Taken together with the suggested functional roles for these SNPs (table 4.4), it is possible that these SNPs might be influencing *CD2AP* expression.

Three associated coding SNPs were found in *EPHA1*, two coding, rs1804527 (synonymous SNP), rs4725617 (missense, benign) and the splice variant, rs6967117. This lends further support to the possible role of this splicing

variant in AD, and suggests further investigation into the possible role of synonymous and benign variants is warranted.

The potentially associated missense SNP in *ZYX*, rs73154206, (OR = 0.340 (0.171-0.675), p = 0.05172) was the variant that was found at a higher frequency in the 1000 genomes population, lending support for further investigation into this gene (table 4.7). Zyxin (*ZYX*) is a LIM domain protein involved in actin remodeling (Hirata et al. 2008), that falls within the 46kpb LD block that includes the GWAS SNP. Therefore it is not unreasonable to consider that the GWAS signal may originate from a functional variant in *ZYX*, particularly given the involvement of synaptic dysfunction and cell membrane processes in AD (Morgan 2011).

The only coding variant in *CD33* with an association is the frameshift deletion found in exon 3 (rs201074739, OR = 0.181(0.044-0.752), p=0.03962, table 4.7). The deletion introduces a premature stop codon in the protein, resulting in a shorter protein being produced that lacks the IGV binding domain of the other three isoforms. This variant is also in LD with the GWAS SNP (D'=-1) and may in fact be the causal variant responsible for the recent finding that the GWAS SNP is associated with decreased levels of CD33 but similar levels of mRNA (Griciuc et al. 2013). It would be interesting to confirm this by verifying protein levels of the deletion mutation *in vitro*.

Only one of the three noncoding variants showed a possible association (table 4.8). This is the variant rs73154206, already discussed above as a potentially associated coding variant in *ZYX*, it also falls downstream of *EPHA1*.

4.4.3. Prioritising variants

Noncoding variants are notoriously difficult to annotate given the lack of functional information available for noncoding regions of the genome. This is reflected in the results of this project where nine noncoding variants had more than six lines of functional support and association compared to 15 coding variants (table 4.9 and 4.10). However four of the noncoding variants in *ZYX* and *EPHA1* overlapped coding variants in the other gene as already mentioned (see end of 4.4.1).

With the advent of annotation initiatives such as the ENCODE (Encyclopedia of DNA Elements) project, annotation of noncoding variants should improve. As much as 98.5% of the human genome is noncoding (Lander et al. 2001). Many GWAS studies have discovered tag SNPs in noncoding regions of the genome, linking noncoding DNA to disease associations and also biological functions through analysis of conserved noncoding elements (Boyle et al. 2012). A recent study investigating the selective pressures on noncoding regions of DNA in the human genome found that transcription factor binding sites and noncoding RNAs contain less SNPs when compared to neutral reference sequence, suggesting a measure of selective pressure is occurring at these sites (Mu et al. 2011).

Both noncoding and coding variants were selected for functional testing based on their annotation for functionality, strength of LD with the GWAS variant and whether they were tentatively associated in the imputed GWAS dataset.

4.4.4. Genotyped prioritised variants

Only variants which had a minor allele frequency high enough to have sufficient power to be detected in the LOAD dataset were put forward for genotyping in the Nottingham genotyping samples (table 4.1). Five coding region variants were put forward for genotyping. All variants apart from the frameshift variant in *CD33* had at least 7 lines of evidence suggesting they be followed up for disease association as possible functional variants.

Following genotyping, none of the variants were found to be significantly associated with LOAD at the 5% significance level (table 4.12 and 4.13). The power of the study to detect an association with LOAD was high assuming an OR of around 1.5. This OR was assumed given the hypothesis that the association signal detected by the GWAS SNP is masking a rare variant with a large odds ratio as has been found for other low frequency variants such as TREM2 and TREML2 (Guerreiro et al. 2013; Benitez et al. 2014). If this assumption is not correct, and the variants tested are associated with LOAD but at a lower OR then it is possible that the sample set used is not large enough to provide sufficient power to detect an association with LOAD (See Chapter 2, section 2.2.1). If the variants do have a smaller association with LOAD then more samples would be needed to be able to pull apart the

association. However, they certainly are not associated with disease with high OR and so the original hypothesis can be refuted.

The addition of covariates did not improve the association for any variants, and appeared to reduce tentative association shown by some variants, indicating possible confounding with either *APOE*, age of onset, gender or centre (table 4.12 and 4.13). All covariates included in the logistic regression were significantly associated with LOAD for all variants. This is unsurprising as *APOE* genotype is known to be the highest risk factor for LOAD. Age and gender are also known to increase LOAD risk (Ertekin-Taner 2007). However, centre was found to be significantly associated with LOAD in the dataset as some centres contributed only case or control samples.

4.4.5. Conclusions

Properly annotating the variants discovered in a NGS study is essential for ensuring the correct variants are prioritised for further time consuming and costly laboratory exploration. In this chapter two different approaches are used in combination to identify potential causative variants which associate with LOAD.

The initial annotation of variants with the VEP, ENCODE and additional datasets revealed nine noncoding variants and 15 coding variants with more than six lines of functional evidence and a minor allele frequency (MAF) of less than 5%. The MAF threshold was selected to reflect the original hypothesis of the study which was to uncover rare variants in the LOAD associated genes, *CD2AP*, *EPHA1* and *CD33*.

Imputation in the independent LOAD GWAS dataset allowed 22 of the rare coding variants and three of the noncoding variants identified in the NGS study to be tested for disease association. The results of the association test in the imputed dataset were combined with the functional annotation to prioritise potential functional variants for further investigation.

Prioritised variants with tentative association, and one additional missense variant in EPHA1 (rs34372369, table 4.11) were put forward for genotyping in an independent dataset. The selection of variants was limited to exclude those which would not have sufficient power to detect an association given the

current number of genotyping samples available (Chapter 4, 4.4.4). Unfortunately no variants were associated with LOAD in the Nottingham samples, indicating true lack of association or incorrect assumption of disease odds ratio during initial power calculations.

The two variants identified as potentially affecting splicing were put forward for further experimental investigation (Chapter 5). Variants in the 5' or 3'UTR and the frameshift variant in *CD33* were selected for allelic expression imbalance (AEI) investigation (Chapter 6)

5. Investigating splicing variants in Alzheimer's disease associated genes, *CD2AP* and *EPHA1*

5.1. Introduction

Following the bioinformatics annotation in Chapter 4 (section 4.3.1, table 4.3), two variants were identified in *CD2AP* (6:47544253A>G) and *EPHA1* (rs6967117) that were predicted to influence splicing.

Splicing is a post-transcriptional process where introns are removed from precursor messenger RNA (pre-mRNA) and the exons are joined to form mature mRNA. Splicing is facilitated through a large ribonucleoprotein (RNP) and protein complex, the spliceosome (Will and Lührmann 2011). The spliceosome recognises specific *cis* sequence elements that define the intron-exon boundaries (Licatalosi and Darnell 2006) (figure 5.1 (A)), namely, the 5' splice or donor site sequence (GU in mRNA), the 3' splice or acceptor site sequence (AG), the branch site, usually 18-40bp upstream from the 3' splice site and, in higher eukaryotes, a polypyrimidine tract (Y) (Will and Lührmann 2011; DeConti et al. 2013) (figure 5.1 (A)). Two catalytic steps are involved in the splicing process. First, the 5' sense strand of the intron is cleaved forming a branched or lariat intermediate. This intermediate is then cleaved at the 3' sense strand and the exons are ligated.

In genes with multiple exons, alternative splicing (AS) allows for different exons to be incorporated into the mature mRNA producing different transcripts and protein isoforms. This increases the diversity of the transcriptome and contributes to the increased functional complexity found in the higher organisms (Kim et al. 2007; Nilsen and Graveley 2010). In humans, over 90% of the roughly 23 000 protein-coding genes are alternatively spliced (Pan et al. 2008).

There are several different ways that transcripts can be alternatively spliced. Most involve cassette-type alternative exons. Other AS methods involve different 5' or 3' splice sites, mutually exclusive exons, intron retention or alternative promoters to be incorporated into the different transcript isoforms (Mills and Janitz 2012). AS to create different transcript isoforms is particularly important in tissues which are functionally complex such as the brain. Many

transcripts in the brain are alternatively spliced (Johnson et al. 2009), particularly during development, where different isoforms may be expressed in different brain regions and during different developmental times (Kang et al. 2011). AS transcripts in the brain also play a role shaping synaptic function and plasticity (Li et al. 2007; Grabowski 2011).

The proper regulation of AS is controlled by *cis*- and *trans*-acting factors which work to either repress (silence) or activate splice site selection (Matlin et al. 2005). These silencer and activator binding sequences can occur in either the exon (exonic splicing enhancers/silencers (ESE/ESS)) or intron (intronic splicing enhancers/silencers (ISE/ISS)). ESEs are degenerate sequences which are recognised by serine arginine (SR) proteins (Cartegni et al. 2002; Cartegni et al. 2003). ISS and ESS binding factors are less well known but seem to include heterogeneous nuclear ribonucleoproteins (Du and Rosbash 2002; Fairbrother et al. 2002; DeConti et al. 2013). It is the relative concentration of enhancers or silencers that determines whether the exon is skipped or included in the final transcript (Matlin et al. 2005). The correct regulation of AS is very important with nearly a third of all pathogenic mutations estimated to cause aberrant splicing (Lim et al. 2011; Sterne-Weiler et al. 2011). Pathogenic splicing variants are often not labeled as affecting splicing (for example if they are missense) therefore it is possible that even more pathogenic mutations could be found to affect splicing (Singh and Cooper 2012).

Sequence variation can lead to dysregulation of pre-mRNA splicing and contribute to disease pathogenesis through causing the complete skipping of an exon, intron retention, introducing a new splice site, activating pre-existing pseudo splice site or by changing the balance of splice isoforms (Baralle and Baralle 2005; Singh and Cooper 2012) (figure 5.1 (B)). Variants can also contribute to disease by altering the secondary RNA structure of a molecule thus changing the accessibility of enhancer and silencer sequences to RNA binding molecules. For example, *MAPT* (tau) has a donor splice site in a stem loop structure which prevents this site from being used (Hutton et al. 1998; Varani et al. 1999; Donahue et al. 2006).

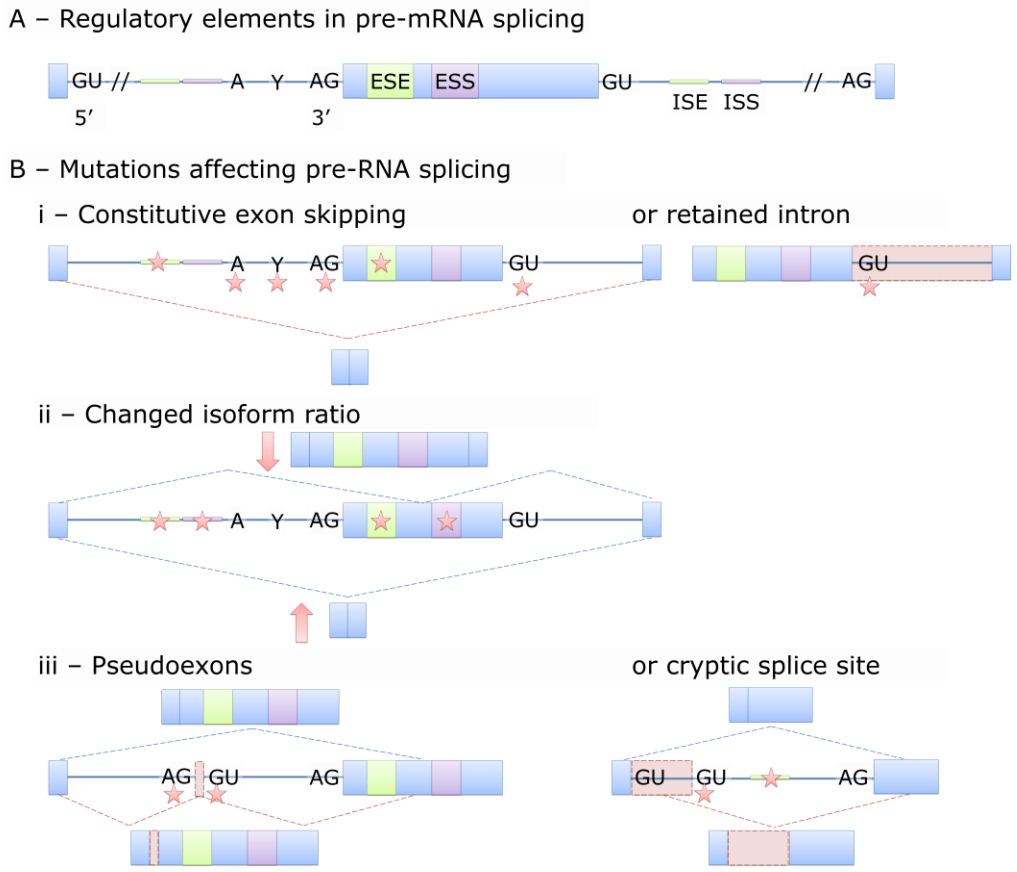

Figure 5.1. Regulatory splicing motifs and mutations which affect splicing in *cis*. Exons are represented as blue rectangles and introns are represented by a line. (A) There are several conserved sequences integral for correct splicing. The GU and AG 5' and 3' splice sites, polypyrimidine tract (shown as Y), branch point (A) and enhancer and silencer sequences in the exons (ESE, ESS) and introns (ISE, ISS). (B) Mutations (indicated by red stars) can affect these regulatory elements and cause disease through three main mechanisms. Aberrant splicing is indicated by dotted red lines with usual splicing indicated by dotted blue lines. (i) Mutations in the branch point, polypyrimidine tract, splice sites and enhancer sequences can cause exons to be skipped or introns to be retained. (ii) Mutations in enhancer or silencer sequences can decrease or increase alternatively spliced isoforms in a cell leading to disease. (iii) Mutations can create a pseudoexon in the intronic sequence or activate a cryptic splice site through changing the splice site sequence or creating a new enhancer sequence causing additional sequence to be included and altering the reading frame of the transcript. Figure adapted from (Baralle and Baralle 2005; Singh and Cooper 2012).

Dysfunctional alternative splicing has been linked to several neurodegenerative diseases. Mutations in the *MAPT* gene change the splicing of exon 10 in inherited frontotemporal dementia (FTD). More examples of AS involvement in neurodegenerative diseases include loss of function splice variants in *SMN* which have been found in spinal muscular atrophy (SMA) and amyotrophic lateral sclerosis (ALS) which has been linked to splicing errors in *EEAT2* and *SOD1* (Dredge et al. 2001). In LOAD, splicing mutations in familial or autosomal dominant genes have been linked to the disease. *PSEN1* and *PSEN2* have both had mutations identified that affect the AS of these genes (Cruts and Van Broeckhoven 1998; Tysoe et al. 1998; De Jonghe et al. 1999;

Sato et al. 2002). *APP* is known to be alternatively spliced into several isoforms (O'Brien and Wong 2011). Additionally aberrant splicing has been implicated in some of the late onset risk genes including *PICALM* and *CD33* (Schnetz-Boutaud et al. 2012; Malik et al. 2013; Raj et al. 2014). Therefore it is possible, given their influence on the binding of splicing regulatory proteins, that the identified splice variants in *CD2AP* and *EPHA1* may affect splicing in these genes which may lead to dysfunction and disease.

The two variants, a novel intronic *CD2AP* variant, 6:47544253A>G, destroying an SRSF5 binding site, and an exonic *EPHA1* variant, rs6967117, having a weak effect on an SRSF1 binding were identified in Chapter 4 (section 4.3.1, table 4.3). As the functional predictions are only through *in silico* tools, both potential splicing variants required experimental verification (Cartegni et al. 2003). As an initial screening method, *in silico* tools are a useful method to test putative splicing variants. However, they are not completely reliable at predicting functional results (Di Resta et al. 2014; Sharma et al. 2014). In particular, functionality of intronic variants such as the *CD2AP* 6:47544253A>G variant may not be correctly predicted as less is known about intronic variants which affect splicing (Jian et al. 2014). Therefore it is important to experimentally verify the effect of these variants on the splicing of their respective genes.

5.1.1. Aims

The aim of this project is to investigate the two splice variants 6:47544253A>G in intron 7 of *CD2AP,* and rs6967117 (also an A>G change) in exon 17 of *EPHA1* using the minigene assay. This assay allows both the reference and alternative alleles of a variant and the surrounding intronic and exonic sequences to be cloned into a splicing vector or plasmid (Cooper 2005) (figure 5.2). By comparing the splicing products transcribed in cell lines transfected with the plasmid containing either the reference or the alternative allele, the effect of each variant can be determined *in vivo*.

5.2. Methods

5.2.1. Bioinformatic investigation

The two splicing variants in *EPHA1* (rs6967117) and *CD2AP* (6:47544253A>G) were assessed using three different freely available *in silico* tools. First, to assess if the variants had any effect on the binding of exonic

splicing enhancers, ESEfinder was used (http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home accessed May 2013) (Cartegni et al. 2003). ESEfinder is able to identify potential ESEs that correspond to four common human serine/arginine protein binding sites (SRSF1 (SF2/ASF), SRSF2 (SC35), SRSF5 (SRp40) or SRSF6 (SRp55)). ESEFinder is also able to predict whether disease-associated variations could interrupt these ESEs. Limitations are that it is designed for exonic enhancer sequences and that it only tests a small number of SR proteins.

Secondly, to test if the variant changed the strength of splice site sequences, Berkley Drosophila Genome Project (BDGP) Splice Site Prediction by Neural Network (http://www.fruitfly.org/seq_tools/splice.html accessed May 2013) was used (Reese et al. 1997). This algorithm uses separate neural network recognizers for the acceptor (AG or 3' splice site) and the donor (GT or 5' splice site) sites to generate a score for potential splice sites (Reese et al. 1997).

Lastly, the effect of the variants on splicing isoforms were assessed using Human Splicing Finder (HSF) (http://www.umd.be/HSF3/ accessed July 2013) (Desmet et al. 2009). HSF includes several different algorithms and it is able to test the effect of variation on splice site, branch point and splicing enhancers and silencer consensus sequences.

5.2.2. Minigene assay investigation

A general overview of the methods used for investigating splicing using the minigene assay are shown in figure 5.2. Two minigene plasmids were created for each gene, one containing the reference allele (Wild Type) and the other containing the alternative allele (Mutant). The plasmids were transfected into two cell lines and the splicing products transcribed from the minigene plasmid were compared using RT-PCR and DNA sequencing.

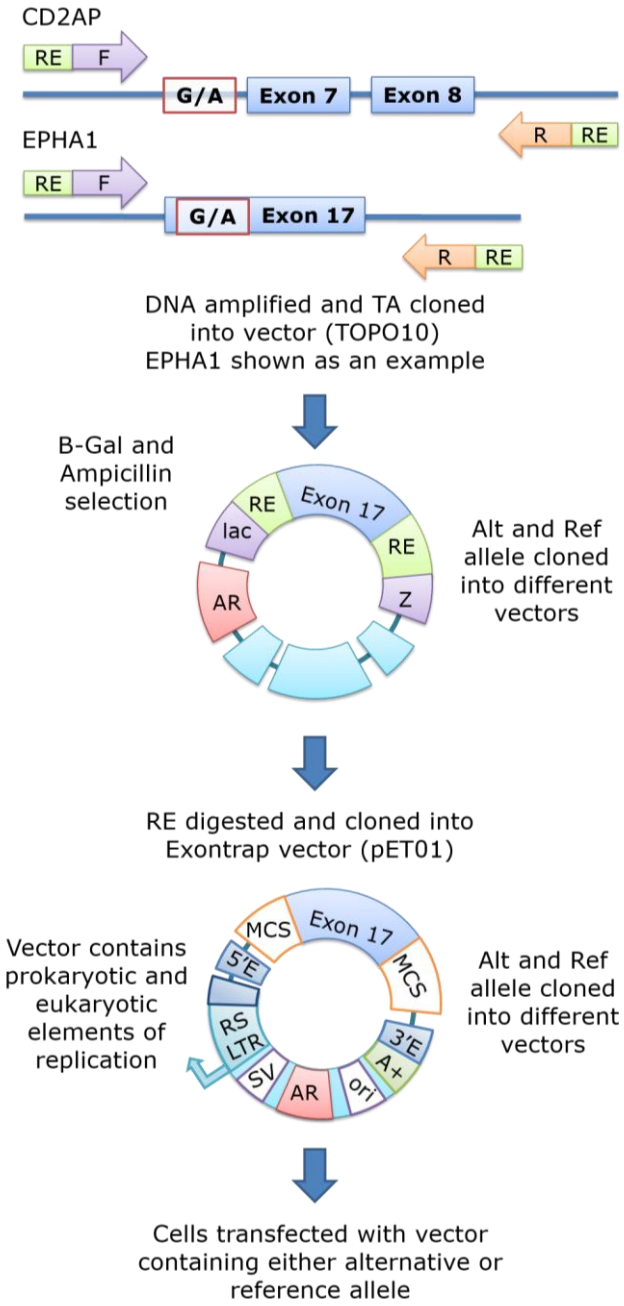

Figure 5.2. Overview of the cloning used to investigate splicing using the *EPHA1* variant as an example. Introns are shown as a line and exons are depicted as boxes. Genomic DNA from an individual heterozygous for the variant of interest is amplified with primers including restriction enzyme sites (RE) and TA cloned to isolate the reference (Ref) and alternative (Alt) alleles. The TOPO10 clone contains the exon of interest (Exon 17) flanked by the restriction enzyme sites (RE) and an ampicillin resistance gene (AR). This is inserted into the lacZ gene (both purple boxes, lac and Z). This vector is digested with restriction enzymes and cloned into the multiple cloning site (MCS) of the Exontrap vector (pET01). The cloned exon is flanked by 5' and 3' eukaryotic exonic sequences (5'E and 3'E) and a polyA tail (A+). This vector contains prokaryotic and eukaryotic elements of replication and can be transcribed in both bacterial and eukaryotic cells. A Rous Sarcoma Virus Long Terminal Repeat (RS LTR) promoter, a small portion of the eukaryotic phosphatase gene and the SV40 origin of replication (SV) allow the vector's replication in eukaryotic cells that express Wildtype T antigen. While a bacterial origin of replication from pBR322 (ori) and the ampicillin resistance gene (AR) allow the vector to be grown and selected for in bacteria.
TOPO10 vector map modified from TOPO TA Cloning product information (www.lifetechnologies.com/support accessed Oct 2013)
pET01 vector map modified from Exontrap product information (www.mobitec.com/cms/products/bio/04_vector_sys/exontrap.pdf accessed Oct 2013).

5.2.3. Primer Design, PCR and sequencing

Minigene primers (table 5.1) were designed based on the *CD2AP* and *EPHA1* reference sequences (NCBI Gene ID 23607 and 2041 respectively, accessed May 2013). Primers were designed as described in Chapter 2, section 2.1.3 with the addition of *Sal*I and *Xba*I restriction enzyme binding sites (see table 4.1 for primer sequences). PCR products were checked for *Sal*I and *Xba*I restriction enzyme sites using Webcutter v2.0 (http://bio.lundberg.gu.se/cutter2/). Primers were checked for single nucleotide polymorphisms (SNPs) as described (Chapter 2, section 2.1.3). All websites were accessed May 2013.

Table 5.1. Minigene primers for *CD2AP* and *EPHA1*. The restriction enzyme sites are indicated in bold.

Gene (rsID)	**Sense primer (location)**	**Antisense primer (location)**	**Product size with exon (bp)**	**Ta (°C)**
CD2AP (6:47544253A>G)	I6S (intron 6) ATAAGT***GTCGAC***A GGGAAAGCTGGG TAACTGT	I8AS (intron 8) GCTCTT***TCTAGA***TC CAGGGTGAATATAA GAACACTCT	775	60
EPHA1 (rs6967117)	I17S (intron 17) CCCAGA***GTCGAC*** ACTTCAGTGCTG GCTGCTG	I18AS (intron 18) CTGTGG***TCTAGA***AA GGGTGGGGGCATG AGT	317	62

All PCR apart from the PCR for the initial TA cloning product, were carried out using 10-100ng template DNA in total volumes of 30µl containing 1X PCR Buffer (NEB), 0.2mM of each dNTP, 1 U Taq DNA polymerase (NEB), 1pmol/µl of each primer. PCR reactions were performed in a Veriti 96-Well Fast Thermal Cycler (Applied Biosystems). PCR conditions for all reactions were as follows: an initial denaturing step of 94°C for 2 min, followed by 30 cycles of 94°C for 30 sec, optimized annealing temperature (Ta, see table 4.1) for 30 sec, 72°C for 1 min with a final extension step of 72°C for 7 min.

For the initial TA cloning reaction, a high fidelity Taq was used to reduce the likelihood of introducing errors into the DNA sequence during the PCR. This PCR was also carried out in total volumes of 30µl with 1X Fermentas Taq buffer with 2mM $MgCl_2$ (Thermo Scientific), 0.2mM dNTPs, 2.6 U Expand High Fidelity Taq (Roche) and 0.25pmol/µl of each primer. The PCR conditions for the high fidelity Taq were as follows: an initial denaturation step of 94°C for 2min, followed by 30 cycles of 94°C for 15 sec, 62°C for 30 sec and 72°C for

40 sec, plus 5 extra sec every cycle. A final extension step was then performed at 72°C for 7 minutes.

Amplified products were separated by electrophoresis at 90V for 20 min in a 2% agarose gel stained with ethidium bromide (EtBr) in 1X TAE buffer and visualized by UV trans-illumination.

PCR products were cleaned up, Sanger sequenced, evaluated and edited as described in Chapter 2, section 2.1.5.

Samples heterozygous for the splicing variants identified in *CD2AP* and *EPHA1* were obtained by sequencing samples from the pool containing the variant as called by CRISP, namely pool 1 for *EPHA1* rs6967117 and pool 7 for *CD2AP* 6:47544253A>G (see appendix 1). Sample AD236 (Nottingham) was identified as heterozygous for the *EPHA1* variant and sample M644 (Manchester) was heterozygous for the *CD2AP* variant. Identifying samples containing the variant of interest for *EPHA1* revealed a second exonic variant, rs1804527, 44 bp away from rs6967117 which was always found to be present with the splicing variant. The variant rs1804527 was investigated using ESEfinder in Chapter 4, section 4.3.1, table 4.3. Data from 1000 Genomes was accessed as described in General Methods (Chapter 2, section 2.2.4) to assess the linkage disequilibrium between the two variants.

Following PCR amplification, 3µl fresh PCR products was TA cloned with 1µl of TOPO vector, 1µl salt (1.2M NaCl and 0.06M $MgCl_2$ to a final concentration of 200mM NaCl and 10mM $MgCl_2$) and 1µl dH_2O using the TOPO TA Cloning system (Invitrogen, Life Technologies). TA cloned products were then transformed into One Shot TOP10 competent *E. coli* cells following manufacturer's instructions and 10-50µl transformation solution was spread on selective agar plates containing 50µg/mL ampicillin and 10µg X-gal (β-galactase substitute) and incubated overnight at 37°C. Six to eight white colonies were selected and cultured overnight in 5mL CircleGrow (MP) medium containing 50µg/mL ampicillin, at 37°C and shaken horizontally at 3 g. Plasmid DNA was extracted using QIAprep Spin Miniprep Kit (Qiagen) following manufacturers protocol. Plasmid DNA concentration was measured on the Nanodrop (ND1000) spectrophotometer. Between 150 and 300ng DNA was directly sequenced using the vector specific forward and reverse primers

M13F and M13R (provided by Invitrogen, Life Technologies) with Big Dye v. 3.1 and sequence electropherograms were checked as described above.

Plasmid DNA from clones containing the full insert with no sequencing errors were then cloned into a transformation vector, pET01 exon trap vector (MoBiTech). First the plasmid DNA samples and the vector DNA sample was digested with the restriction enzymes *Sal*I and *Xba*I (Fast Digest, Fermentas) in 20µl total volume following manufacturer's instructions. Digested products were electrophoresed on 3% agarose gel stained with SYBR-safe stain solution (Invitrogen), and visualized using blue-light transillumination. The insert band was excised from the gel using a clean scalpel blade and cleaned using Qiaquick spin columns (Qiagen) according to manufacturer's protocol. The DNA concentration was measured again and the amount of DNA required for the ligation reaction was calculated using the below equation, where 4461 is the size of the exon trap vector in base pairs (bp):

$$\left[\frac{100ng\ pET01\ \times\ size\ of\ insert\ in\ bp}{4461\ bp}\right] \times 3$$

The appropriate concentration of insert DNA (21ng for *EPHA1* and 52ng for *CD2AP*) and vector DNA (100ng) were ligated using T4 ligase. High efficiency NEB 5-alpha Competent *E. coli* cells were then transformed with vector pET01 containing the insert for each allele of the two SNPs using High Efficiency Transformation C2987I (NEB). A vector only control was also transformed to check the efficiency of the ligation reaction. Cells were plated out as before and single colonies were selected and grown overnight. Plasmid DNA was then extracted and sequenced following the same procedure as above to confirm the colonies contained an error free insert in the exon trap vector. When suitable colonies had been found, these were grown overnight in 5mL CircleGrow (MP) medium containing 50µg/mL ampicillin as described previously and the plasmid DNA was extracted using an Endo-free Kit: Endo-Free Plasmid DNA Midi Kit Centrifugation (OMEGA) for bacterial colonies containing the *EPHA1* reference and alternative sequences and Endotoxin-free plasmid DNA purification NucleoBond Xtra Midi EF (Macherey-Nagel) for bacterial colonies containing the *CD2AP* sequences. Extracted Endo-free DNA was sequenced again to confirm that the appropriate insert was present.

5.2.4. Cell culture and transfection

The cell culture and transfection for the *EPHA1* splicing mutation was undertaken by the BMedSci student, Lucy Millar under my supervision. The exontrap vectors containing all four inserts, *EPHA1* reference, alternative and *CD2AP* reference and alternative sequences were transfected into two mammalian cell lines. COS-7 and U-87MG cells for *EPHA1* and COS-7 and BE-2 cells for *CD2AP.* Cells were obtained from the European Collection of Cell Cultures (ECACC).

The COS-7 (CV-1 in Origin, carrying SV40) cell line is derived from African Green Monkey kidney cells through transformation with the Simian Vacuolating Virus 40 (SV40). This cell line is usually used for testing minigene splicing assays as it is easy to transfect and grow. U-87MG cells are established human astrocytoma and glioblastoma cell line with epithelial morphology derived from a 44 year old male Caucasian. The BE-2 cell line are established human neuroblastoma cells with neuroblast morphology derived from a 2 year old male Caucasian. The U87-MG and the BE-2 cell lines were selected for their similarity to neural cells, either astrocytes or neurons. This allowed the exontrap vectors to be subjected to *in vitro* environments more similar to the brain, the main diseased tissue in LOAD.

All cell culture reagents were acquired from Sigma. COS-7 cells were grown in Dulbecco's Modified Eagle Medium (DMEM). The U-87MG and the BE-2 cells were both grown in Eagle's Minimal Essential Medium (EMEM) supplemented with 1% Non-Essential Amino Acids (NEAA). For the U-87MG cells the media was supplemented with 1mM Sodium Pyruvate. Media for all cells were supplemented with a final concentration of 10% fetal bovine serum (FBS), 2mM L-Glutamine, 100 units/mL Penicillin, 100µg/mL Streptomycin and 2.5µg/mL Fungizone (Invitrogen). All cell lines were maintained at 37°C in a 5% CO_2, humid atmosphere. Additionally, all incubation steps mentioned below were carried out under these conditions.

Cells which were 90% confluent were detached using 3mL Trypsin (Sigma) which was inactivated with 3mL of the appropriate complete filtered medium after a 15 minute incubation. 200ul of media and cells were pipetted onto a hemocytometer for counting. $2.5x10^5$ COS-7 cells, $1.0x10^6$ U-87MG cells or $6.0\ x10^5$ BE-2 cells were transferred to two 60mm cell culture plates and 5mL

complete filtered media was added. Plates were incubated overnight and confluence was assessed after 24 hours. Transfection was performed when cells reached 50-60% confluence. Any cells which had not reached this confluence after 24 hours were given fresh media and incubated for an additional 24 hours.

The transfection was repeated twice in each cell line, starting with the COS-7 cells. Transfections were performed as follows: 1µg reference or alternative Endo-free plasmid DNA was added to 2ml filtered serum-free medium and 9µl Transfast (Promega) and vortexed before incubating at room temperature for 15 minutes. The old medium was removed from the cell culture plates which were then washed with PBS. Following incubation the DNA and Transfast mix was added to two plates, one containing plasmid DNA with the reference insert sequence, the other containing DNA with the alternative sequence. Cells were incubated for 1 hour after which 4mL complete filtered medium was added to each plate. Plates were then incubated for 24 hours. Total RNA was extracted from the cells using the RNeasy Mini Kit (Qiagen) following manufacturer's instructions. An additional DNase treatment using TURBO DNA-free (Ambion) was necessary to remove any additional plasmid DNA from the extracted RNA.

Complementary DNA (cDNA) was generated using the AffinityScript Multiple Temperature cDNA synthesis kit (Agilent) according to manufacturer's protocol. See Chapter 2, section 2.1.4 for full details.

The cDNA from the reference and alternative transfections were amplified by PCR using the vector specific primers, pET01S and pET01AS (provided by MoBiTech) which flank the cloned region of interest in the vector. PCR products were run on a 2% agarose gel to determine if the mutation affected splicing prior to Sanger sequencing as described previously (Chapter 2, section 2.1.5).

5.3. Results

5.3.1. Bioinformatic investigation

ESEfinder showed that the *EPHA1* variant, rs6967117 slightly reduces the binding affinity of serine/arginine-rich splicing factor 1 (SRSF1) (Reference score = 5.19, alternative score = 4.76), while the *CD2AP* variant, 6:47544253A>G causes a SRSF5 site to be lost (figure 5.3, table 5.2).

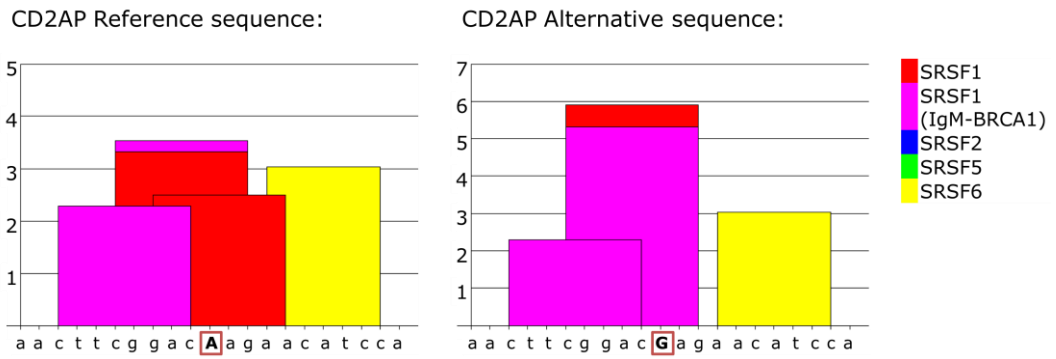

Figure 5.3. ESEfinder predictions for *CD2AP* variant 6:47544253A>G. The intronic variant causes a SRSF1 protein binding site to be lost, but strengthens the predicted binding affinity for another SRSF1 binding site. The key for the binding proteins predicted location is colour coded on the right hand side. The pink and red both indicate SRSF1 binding scores but by different prediction algorithms. The graphs show the predicted binding score on the y-axis and the sequence input for the reference and alternative alleles on the x-axis.

The BDGP predictor program showed that rs6967117 generated a new donor site with a weak score. This variant also causes a slight decrease in the score for the acceptor site for the alternative allele (from 0.70 for the reference allele to 0.62 for the alternative allele). This new donor site is unlikely to be functional as it falls near an existing acceptor site. For 6:47544253A>G, the alternative allele causes a very slight decrease in the acceptor site (from 0.83 for the reference allele to 0.80 for the alternative allele). See table 5.2 for the results.

Only the *EPHA1* variant showed any effect on splicing when the Human Splicing Finder (HSF) program was used. This variant was predicted to activate a cryptic intronic donor site (table 5.2). This confirms the prediction from the BGDP program.

Table 5.2. Splicing prediction results for the two variants from three bioinformatic programs. The table also lists the minor allele frequency (MAF) for the variant which was called from the NGS data using CRISP.

Gene	Variant (rsID)	CRISP MAF	ESEfinder – SRProteins	BDGP	HSF
EPHA1	7:143088867 (rs6967117)	0.060	Yes - Reduces score for SRSF1 site.	Generation of NEW donor site, 0.49, slightly increases acceptor site, by 0.08.	Activation of an intronic cryptic donor site. Potential alternation of splicing.
CD2AP	6:47544253	0.005	Yes – SRSF5 binding site is lost.	Very slightly decreases acceptor site, by 0.03.	No effect predicted.

5.3.2. Minigene assay

Initial sequencing of the *EPHA1* variant revealed that rs6967117 was always found with another variant present, namely rs1804527 (figure 5.4). Using VCFtools to calculate the LD in the 1000 Genomes data showed the two variants had an r^2 and D' of 1 and that the GG and AA alleles were in phase.

Following transfection into two cell lines, no difference was seen in the mRNA products from either COS-7 cells or U-87MG cells expressing the *EPHA1* insert for either allele when RT-PCR products were run on a 2% agarose gel stained with EtBr (figure 5.5). A product size of 398bp was produced for both *EPHA1* alleles in all cell lines. This translates to the expected product size of exon 17 (156bp) and two vector exons (180bp and 62bp respectively). Therefore the exon is included in both PCR products following transfection.

Additionally, no difference was seen in the mRNA products from COS-7 cells or BE-2 cells expressing the *CD2AP* insert (figure 5.6). A product size of 416bp was produced for both *CD2AP* alleles in all cell lines. This is the expected product size of exon 7 and exon 8 (79bp and 95bp) and the two vector exons. These results were confirmed by Sanger sequencing which also revealed no differences in RT-PCR products from the alternative or reference allele from either cell line.

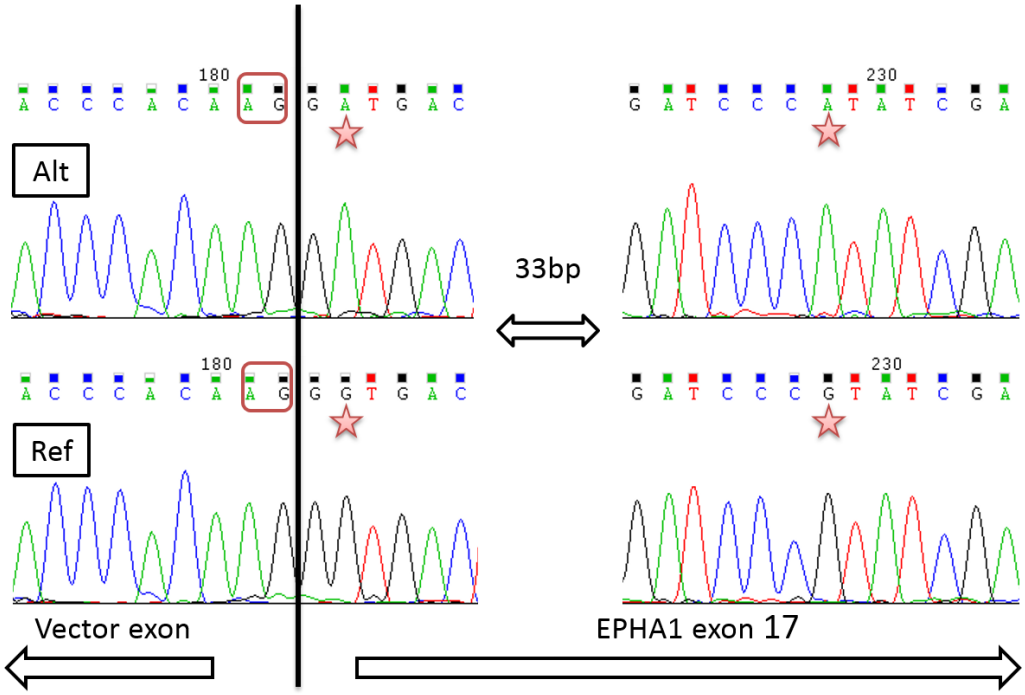

Figure 5.4. Electropherogram for the alternative and reference sequences from the minigene assay for the *EPHA1* variant. The star indicates the two alleles which are always found together. The red box indicates the AG 3' acceptor sequence.

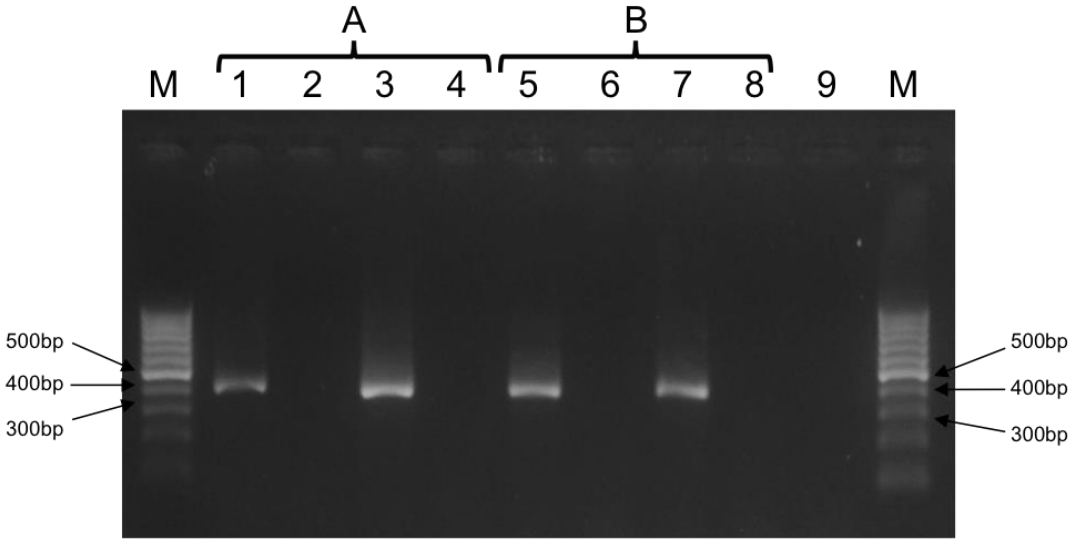

Figure 5.5. RT-PCR products from *EPHA1* mRNA expression in COS-7 (A) and U-87MG (B) cells run on a 2% agarose gel stained with EtBr. M shows the 100bp DNA ladder, while lane 9 is the non-template blank. Lanes 1 and 5 show the alternative allele RT(+) while lanes 2 and 6 show the alternative allele RT(-).for the respective cell lines. Lanes 3 and 7 show the reference allele RT(+) and lanes 4 and 8 show the reference allele RT(-). All RT(+) lanes show a band around 398bp (exon 17 (156bp) and the two vector exons (180bp and 62bp respectively). This shows that vector introns have been spliced out and that there is no difference between alternative and reference products from either cell line. Gel photo reproduced with permission from L. Millar.

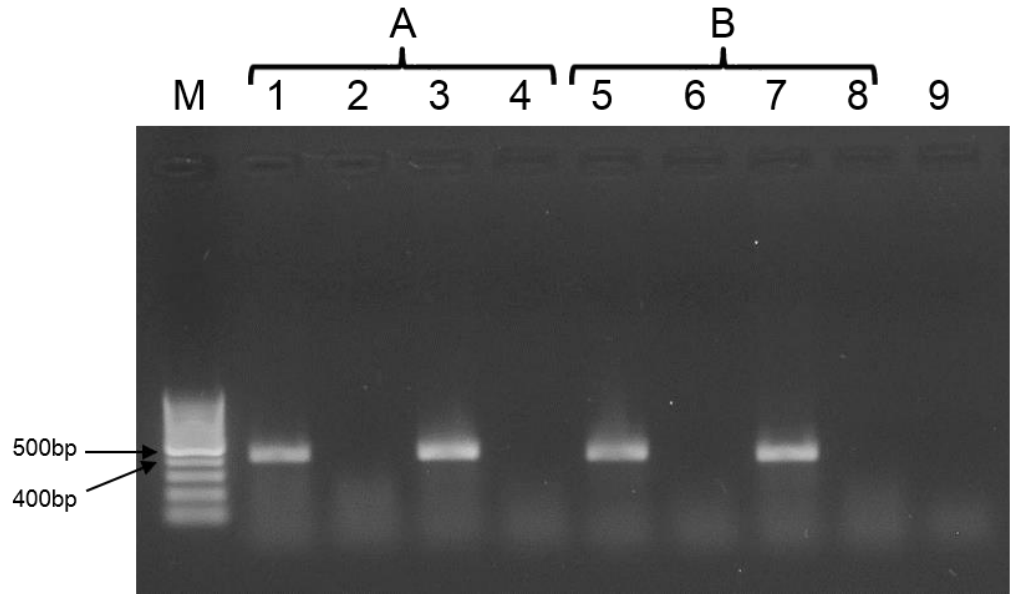

Figure 5.6. RT-PCR products from *CD2AP* mRNA expression in COS-7 (A) and BE-2 (B) cells run on a 2% agarose gel stained with EtBr. M shows the 100bp DNA ladder, while lane 9 is the non-template blank. Lanes 1 and 5 show the alternative allele RT(+) while lanes 2 and 6 show the alternative allele RT(-).for the respective cell lines. Lanes 3 and 7 show the reference allele RT(+) and lanes 4 and 8 show the reference allele RT(-). All RT(+) lanes show a band around 416bp (exon 7 and exon 8 (79bp and 95bp) and the two vector exons (180bp and 62bp)).This indicates that the vector introns have been spliced out and that there is no difference between alternative and reference products from either cell line.

5.4. Discussion

This project aimed to investigate two splice variants, 6:47544253A>G in intron 7 of *CD2AP* and rs6967117 in exon 17 of *EPHA1* using bioinformatic prediction programs and the minigene assay. Two cell lines were transfected with the reference and alternative minigene vectors and splicing products were analysed by Sanger sequencing. No difference was found between the reference and alternative products obtained from either variant from both cell lines.

Initial bioinformatics results for the two variants did not show very strong evidence to support either alternative allele as having an effect on splicing. However, given the importance of alternative splicing for transcript regulation in the brain (Lee and Irizarry 2003), the location of the variants and the unreliability of splicing prediction programs (Di Resta et al. 2014; Sharma et al. 2014), it was decided they should be put forward for experimental verification.

The *EPHA1* variant, rs6967117, did show a slight decrease in splice site acceptor score generated by BDGP and the creation of a new intronic donor site which was confirmed by HSF predictions suggesting that the variant activated a cryptic intronic donor site (see table 5.2 and also figure 5.1). If rs6967117 did activate a cryptic intronic donor site, a pseudoexon could be incorporated into the transcript (figure 5.1).

Both splice variants were predicted to influence the binding of splicing regulatory proteins (table 5.2). The *EPHA1* variant rs6967117 decreased the binding affinity of the splice site enhancer protein, SRSF1 and the *CD2AP* variant, 6:47544253A>G removed a SRSF5 binding site. If the variants influenced the binding of these regulatory proteins, this could disrupt the native splicing of these transcripts. The SRSF (serine/arginine-rich splicing factor) proteins are nuclear RNA-binding proteins integral to mRNA splicing and additional cellular processes involved in the regulation of gene expression (Manley and Krainer 2010). All contain one or two N-terminal RNA-binding domains and a RS domain containing arginine and serine dipeptides in the C-terminal. SRSF1, also known as SF2/ASF, is the most well studied SR protein and changes in the binding affinity of this protein have been linked to dysregulation of *MAPT* exon 10 splicing in the neurodegenerative disease, FTD (Hutton et al. 1998; Kondo et al. 2004; D'Souza and Schellenberg 2006). It has also been shown to be affected by miRNA binding (Wu et al. 2010) which is known to be dysregulated in LOAD (reviewed in (Femminella et al. 2015)). While SRSF5 (or SRp40) is less well studied, it is known to regulate the circadian rhythm (McGlincy et al. 2012) and was found to be differentially expressed in bipolar disorder (Akula et al. 2014). However, the SRSFs are a large family with similar domains (Manley and Krainer 2010), and there is some redundancy in their function which might allow for another SRSF protein to compensate for any disruption in binding of just one SRSF (Long and Caceres 2009).

The programs developed to predict the effects of mutations within the 3' and 5' consensus splice site regions are more robust than those predicting the effects of ESE binding proteins (Jian et al. 2014). This is mainly due to the increased knowledge of these consensus sites allowing better modelling of these regions and therefore better prediction programs to be written (Jian et al. 2014). Two prediction programs used in this study examine the effects of

the mutations on the 3' and 5' consensus regions, BDGP and HSF. Both these programs showed minimal evidence of the two variants affecting splicing, thus confirming the superior predictive annotations from these programs (table 5.2). The effect of mutations on RNA binding proteins should not be ignored, but it should be noted that further experimental information would be beneficial to improve the *in silico* prediction tools currently available (Di Resta et al. 2014; Sharma et al. 2014).

In silico tools are often used as an initial screening method for putative disease-causing variants. However, the combination of several tools which assess different aspects of splicing should be used. For example the effect of the variant on splice site strength, prediction of splicing isoforms, prediction of splicing binding proteins and any effects on secondary mRNA structure should be used to test the effect of a variant (Baralle and Baralle 2005; Di Resta et al. 2014; Sharma et al. 2014). Even so, there are difficulties consolidating the predictions from different programs and *in silico* tools are not completely reliable at predicting functional results (Di Resta et al. 2014; Sharma et al. 2014).

Through sequencing patients from pool 1 for the *EPHA1* variant, it was determined that rs1804527 was always found with the splicing variant. These two variants are 44bp apart and were found to be in complete linkage disequilibrium ($r^2 = 1$ and D' = 1). The second variant also suggested to influence splicing as *in silico* analysis shows it introduces two new SRSF binding sites, for SRSF2 and SRSF6 (Chapter 4, table 4.3). Therefore the minigene assay for *EPHA1* was created to contain both variants as this would reflect the haplotype found in individuals in the population.

Neither the *EPHA1* nor the *CD2AP* variant showed any effect on splicing in the COS-7 cell line following visualisation on the agarose gel. This was confirmed by Sanger sequencing. However, while this cell line is a good model for *in vitro* splicing experiments, given its high transfection efficiency, splicing is known to be tissue-specific, with more transcripts being alternatively spliced in the brain (Yeo et al. 2004). Therefore two neuronal cell lines were also transfected, U-87MG for the *EPHA1* variant and BE-2 for the *CD2AP* variant although no effect was also found for either variant in either cell line (figure 5.5 and 5.6).

It has recently been argued that the splicing minigene assay is not a good way of testing the effect of splicing variants *in vitro* as it only uses a small portion of the gene of interest which is inserted adjacent to a portion of a completely different gene (in this project, eukaryotic phosphatase gene) triggered by a viral origin of replication. Arguably the best way of examining the effect of splicing variants is by looking at RNA directly from patient individuals who contain the splicing variants. Given the difficulties obtaining RNA from such individuals, we are left using *in vitro* methods. New *in vitro* methods have been developed with vectors which contain the entire gene being investigated (Sharma et al. 2014). However if the variant is largely disruptive to splicing, the effects will be seen in the splicing vector used in this project and even in the non-human COS-7 cell line (Cooper 2005). The minigene assay has been shown to be an effective *in vivo* model accurately assessing both aberrant and normal splicing (Cooper 2005).

Exon arrays and RNA-seq are two high throughput methods used to study alternative splicing which are proving useful in identifying new AS events in different brain tissue regions (e.g. Twine et al. 2011; Mills and Janitz 2012). Unfortunately two caveats of these methods are that they require good quality RNA which can often be problematic to obtain from brain tissue (see Chapter 6, section 6.3.1 and 6.4.1 for issues with RNA preparation from brain tissue in this project) and that they will only show a snapshot of the transcriptome at one particular time.

Although no change is occurring in the two variants investigated in *CD2AP* and *EPHA1,* there is evidence to support a role of dysfunctional splicing in LOAD. Other studies performing transcriptome profiling of frontal, temporal and parietal lobes of LOAD patients compared with controls revealed distinctive AS and promoter use in LOAD, with the parietal lobes showing different patterns to the rest of the brain (Twine et al. 2011; Mills et al. 2013). Recently, the spliceosome component U1 small nuclear ribonucleoprotein (snRNP) (Bai et al. 2013), and the splicing factors from the A and B families of the heterogeneous nuclear ribonucleoproteins (hnRNP) (Berson et al. 2012) have been found to be involved in LOAD. Therefore it is possible that changes in splicing in other genes are contributing to LOAD disease pathology.

5.4.1. Conclusions

This project found no differences in the splicing products produced from the reference or the alternative allele of two potential splice variants, 6:47544253A>G in intron 7 of *CD2AP* and rs6967117 in exon 17 of *EPHA1.*

Better prediction programs which incorporate additional experimental databases and use different prediction algorithms are needed to improve the predictions given by current splicing programs.

While the two variants investigated in this chapter do not impact splicing, it is important to note that aberrant splicing may still have a role to play in the disease pathogenesis of LOAD associated GWAS genes. Future studies examining RNA-seq data may be able to provide additional information on this.

6. Assessing predicted functional variants in Alzheimer's disease associated genes, *CD2AP, EPHA1* and *CD33* using allelic expression imbalance

6.1. Introduction

Following the discovery of the genetic regions associated with complex disease-risk through genome wide association studies (GWAS), it has become an important next step to document all variations present in these gene regions in order to identify functional disease-causing variants (Chapter 3 and 4). However, once these variants have been identified and functionality has been assessed with bioinformatics, laboratory-based validation of these prioritized variants is essential.

Allelic expression imbalance (AEI) is a powerful tool to detect a range of mechanisms that may affect gene expression at the transcript level (Smith et al. 2013). Most genes will express transcript equally from both alleles of a variant, however 20-30% of the genes expressed in humans show AEI (Serre et al. 2008; Ge et al. 2009; Vidal et al. 2011). Allelic imbalance occurs when a particular gene expresses two alleles of a heterozygous SNP at different ratios. Through comparing the expression of two alleles in the same individual, *cis*-acting variants which affect phenotype can be identified (see figure 6.1). It is also possible to detect the effect of potential noncoding variants on transcript expression if the noncoding variant is in complete linkage disequilibrium with a coding variant which can be detected in the transcript. Many of the putative functional variants for GWAS are thought to be noncoding and affect gene regulation rather than directly modify the protein (Maurano et al. 2012). Therefore AEI will be useful for investigating these variants as it can easily identify mutations in *cis*-acting regulatory elements which may affect transcription (Serre et al. 2008; Ge et al. 2009; Adoue et al. 2014). For example, any variants affecting RNA stability, methylation status, nonsense mediated decay or random monoallelic expression will show AEI (Nothnagel et al. 2011; Eckersley-Maslin and Spector 2014).

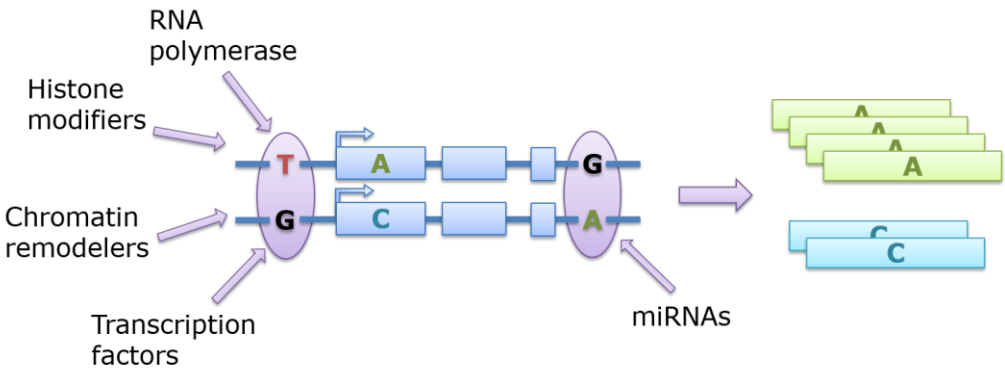

Figure 6.1. Detection of allelic expression imbalance. The blue boxes represent exons and the horizontal blue lines represent introns and flanking sequence for a gene. The light blue and green boxes represent transcribed and processed mRNA. Measuring the different levels of transcripts produced from each allele of a heterozygous variant can shed light on the effects of the coding variant (A>C change) on transcript production. Additionally, any effects of noncoding variants in complete linkage disequilibrium with the coding variants can also be tested. For example, the T>G change upstream of the gene could affect the binding of the indicated DNA and RNA-binding molecules while the G>A change downstream of the gene could affect the binding of microRNAs (miRNAs) and stabilising RNA-binding proteins.

In AEI the transcribed allele acts as an internal control as any trans-regulatory elements and environmental factors will affect the expression of both alleles. Therefore, it is a better method than expression quantitative trait location (eQTL) studies for investigating variants which affect gene expression. In eQTL studies, variation in gene expression is compared with individual genotypes. These studies are often underpowered due to differences in causative variations between individuals, modest effect sizes of many regulatory variants and correction for multiple testing (Cookson et al. 2009; Almlöf et al. 2012).

Normal cellular development, differentiation and function can rely on differential or even monoallelic expression of alleles through stochastic effects or through imprinting via parent-of-origin effects (e.g. Lin et al. 2012). Dysregulation of AEI can be pathogenic and may play a role in the phenotypic variation of complex diseases. For example, the autism spectrum disorders and schizophrenia both appear to be affected by differential expression of maternal and paternal alleles (Wilkins 2010). Discovering variants which exhibit AEI will allow disease-causing haplotypes to be identified (Bell and Beck 2009).

There are low and high throughput methods for identifying AEI. Early studies used single-base extension of a primer next to the SNP of interest which was

fairly low throughput, allowing several genes to be tested at a time. This was scaled up using microarray-based technologies to screen more genes at once. RNA-seq, a high throughput next generation sequencing technology, has allowed researchers to explore the transcriptome in an unbiased manner, revealing allelic bias in transcripts. However, careful data analysis is required and it is often recommended to verify results with an alternative sequencing method (Smith et al. 2013).

6.1.1. Aims

Potential functional variants identified in the 3' and 5' UTR regions of *CD2AP, EPHA1* and *CD33* in the NGS study (Chapter 4, table 4.9 and 4.10) will be experimentally validated using RNA allelic expression imbalance (AEI). The allelic RNA expression ratios of the potential functional variants will be measured in LOAD brain tissue from individuals known to be heterozygous for the SNP of interest. Given that previous laboratory experience has shown the RNA quality from the brain tissue samples is very variable (Kristelle Brown, pers. comm.), variants will also be investigated using Epstein-Barr transformed B-lymphoblastoid cell lines (LCLs) containing genetic material from two 1000 Genomes individuals identified as heterozygous for four of the variants being investigated in the brain tissue samples. Any variants that display differential allelic expression will be highlighted as the possible causative variant(s) for the GWAS signal and will be put forward for further investigation.

6.2. Methods

6.2.1. RNA extraction optimisation

Previous RNA extractions using the Nottingham DNA bank brain tissues indicated that RNA quality varied significantly depending on the extraction method used (K. Brown, pers. comm.). Therefore the initial work for the allelic expression imbalance involved optimising the RNA extraction protocol. RNA was extracted from one brain tissue sample (M669) using three extraction methods: TRIzol reagent followed by ethanol precipitation, Direct-zol RNA MiniPrep (Zymo) and the RNeasy Mini Kit (Qiagen).Two different initial homogenisation methods were also tested: either transitioning frozen tissue from -80°C to -20°C using RNA*later*-ICE (Ambion) and then homogenizing the sample in the extraction buffer using a hand-held pestle homogenizer or directly homogenising frozen tissue under liquid nitrogen with a pestle and

mortar. RNA integrity for all extraction methods was measured using the RIN (RNA integrity number) by running samples on the 2100 Bioanalyzer (Agilent) using the Agilent RNA 6000 kit and following manufacturer's instructions (Mueller et al. 2004).

RNA can be rapidly degraded by ubiquitous RNases into small fragments which can affect downstream analyses. Traditionally, electrophoresis in an agarose gel stained with ethidium bromide (EtBr) was used to determine the size (and integrity) of RNA molecules (Sambrook and W Russell 2001). High quality RNA can be seen in the gel as two bright bands, one at 28s which should be twice the intensity of the other band at the 18s from ribosomal RNA (rRNA) with other smaller bands (sometimes visible as a smear) indicating smaller RNA molecules (see figure 6.2). As the ratio of the 28s:18s is known to be around 2, this can be used to estimate the quality of the RNA with high quality RNA showing a 28s:18s ratio of 2 or more. However this was a subjective measurement and cannot be easily quantified or reliably reproduced across different laboratories.

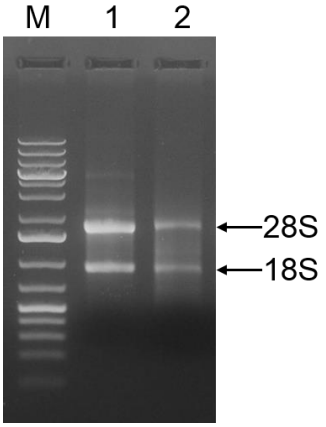

Figure 6.2. RNA samples run on 2% agarose gel stained with Ethidium bromide. RNA was loaded with denaturing buffer (formamide, formaldehyde, glycerol and bromothymol blue, adapted from (Sambrook and W Russell 2001)). M is a 100bp DNA ladder for comparison. Lane 1 is RNA extracted from cell lines, while lane 2 is RNA extracted from brain tissue showing degradation with a fainter band for 28s (~3500bp as marked by the DNA ladder). Both RNA extractions were performed using the RNeasy mini kit.

The Agilent 2100 Bioanalyzer uses automated microfluidics to separate DNA, RNA and protein on molecular weight (Schroeder et al. 2006). The molecules are separated using electrophoresis and detected by fluorescence initiated by laser. The resulting electropherogram shows the amount of RNA (or other molecule) at a particular size as the amount of fluorescence detected at that position (see figure 6.3). As the peak heights of the electropherogram are quantifiable the RIN value that is automatically calculated for each sample is reliable and reproducible (Imbeaud et al. 2005). Therefore this was the method used to verify RNA integrity for the different extractions.

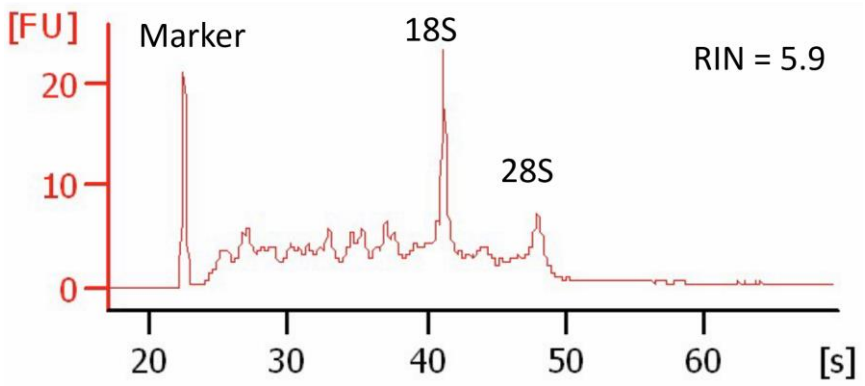

Figure 6.3 Agilent 2100 Bioanalyzer trace for a sample with a RIN of 5.9. Fluorescence [FU] is shown on the y-axis and time in seconds [s] on the x-axis. The fluorescent peaks for the marker, 18sRNA and 28s RNA are shown.

The RNeasy mini kit (Qiagen) using pestle and mortar homogenization under liquid nitrogen and manufacturer's protocol with the additional on-column DNase treatment was selected as the best method for both RNA quality and processing time for brain tissue samples.

6.2.2. Laboratory investigation of AEI variants

All genes were checked for RNA and protein expression in cortical brain tissue and LCLs using the Human Protein Atlas (http://www.proteinatlas.org/ accessed October 2013 (Uhlén et al. 2005; Uhlen et al. 2010)). All genes showed expression in brain tissue, however data was not available for gene expression in LCLs. Therefore the publicly available RNA-seq data from the cell lines were accessed to ascertain gene expression via the ENSEMBL genome browser (http://www.ensembl.org/Homo_sapiens/Location/View?r=1:14367-24367;contigviewbottom=url:http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/HG00137.6.M_120217_1.bam;format=Bam; accessed October 2013 (Lappalainen et al. 2013)). All genes had between 100 and 6000 reads aligned to their positions.

Initial investigation into the allelic expression differences in the list of prioritised 5' and 3' UTR variants was undertaken by an MSc student, Lena Karrar. Under my supervision she performed the RNA extractions and cDNA synthesis for the brain tissue. All LCL work, primer design, PCR amplification, Sanger sequencing and subsequent analysis was performed on my own.

To investigate allelic expression imbalance in brain tissue and in LCLs, genomic DNA (gDNA) and cDNA were amplified by PCR and Sanger sequenced in parallel following the protocol outlined in Chapter 2, section 2.1.4 and 2.1.5. The heterozygous gDNA samples are expected to amplify with an allele ratio of 50:50 and serve as a control for comparing the allele ratios obtained from cDNA samples.

6.2.2.1. Brain tissue samples

Exome sequencing data for 326 cerebral cortical brain tissue samples from the Nottingham Brain Bank was kindly provided by Rita Guerreiro as part of a collaboration with John Hardy's research group at UCL (data published in Guerreiro et al. 2013). This allowed me to identify brain tissue samples that were heterozygous for the variant of interest. Consent was previously obtained for all samples and a local ethics committee approved the study. Samples were pathology-confirmed as LOAD except for M660 which changed diagnosis to mild cerebral amyloid angiopathy (CAA). Sample information for each of the SNPs are presented in table 6.1.

Table 6.1. Cerebral cortical brain tissue information for the SNPs investigated in the allelic expression imbalance study. The base change for the SNP relative to the forward reference sequence is presented in brackets in the first column. Missing information is denoted by "Missing". *APOE* genotype for the three epsilon alleles are denoted by 2, 3 or 4 is shown in the last column.

Gene and SNP ID	**Sample ID**	**Gender**	**Age-at-death**	**Diagnosis**	***APOE***
CD2AP rs151064033 (A to T)	M546	F	65	Confirmed AD	33
CD2AP rs141029774 (G to A)	M596	M	77	Confirmed AD	34
	M602	F	91	Confirmed AD	33
	M660	F	85	Mild CAA	34
EPHA1 rs1131883 (C to T)	M602	F	91	Confirmed AD	33
	M605	M	50	Confirmed AD	23
	M647	F	Missing	Confirmed AD	34
ZYX rs11552744 (G to A)	M636	M	81	Confirmed AD	33
	M647	F	Missing	Confirmed AD	34
	M648	F	Missing	Confirmed AD	34
	M650	M	69	Confirmed AD	33
CD33 rs201074739 (CCGG del)	M649	F	93	Confirmed AD	34
	M651	F	89	Confirmed AD	33
CD33 rs1803254 (G to C)	M547	F	67	Confirmed AD	34
	M638	F	84	Confirmed AD	34

RNA was extracted using the RNeasy mini kit as mentioned in section 6.2.1. Following extraction, cDNA was synthesised from the RNA using the AffinityScript Multiple Temperature cDNA synthesis kit (Agilent) as described in Chapter 2, section 2.1.4. Matched genomic DNA (gDNA) extracted using the phenol chloroform method (see section 2.1.1, Chapter 2) was also obtained for each AEI sample.

6.2.2.2. 1000 Genomes LCLs

Individuals from the 1000 Genomes project (1KG) who were heterozygous for four of the variants of interest were identified using the 1000 Genomes browser (http://browser.1000genomes.org/ accessed April 2014). Namely, HG00137 heterozygous for variants rs1131883 (*EPHA1*) and rs201074739 (*CD33*) and HG00255 heterozygous for rs11552744 (*ZYX*) and rs1803254 (*CD33*). LCLs for HG00137 and HG00255 were obtained from Coriell Cell Repositories from the NHGRI Sample Repository for Human Genetic Research.

Cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 media with 2mM L-glutamine, 1% Fungizone and 1% Penicillin/Streptomycin, supplemented with 15% FBS in 20ml total media in a T25 flask. All cell culture reagents were purchased from Sigma-Aldrich. Cells were maintained at 37°C in a humidified atmosphere with 5% CO_2. Cells received fresh media every 2-4 days depending on cell growth to keep the cell density between 200 000 and 500 000 cells/ml.

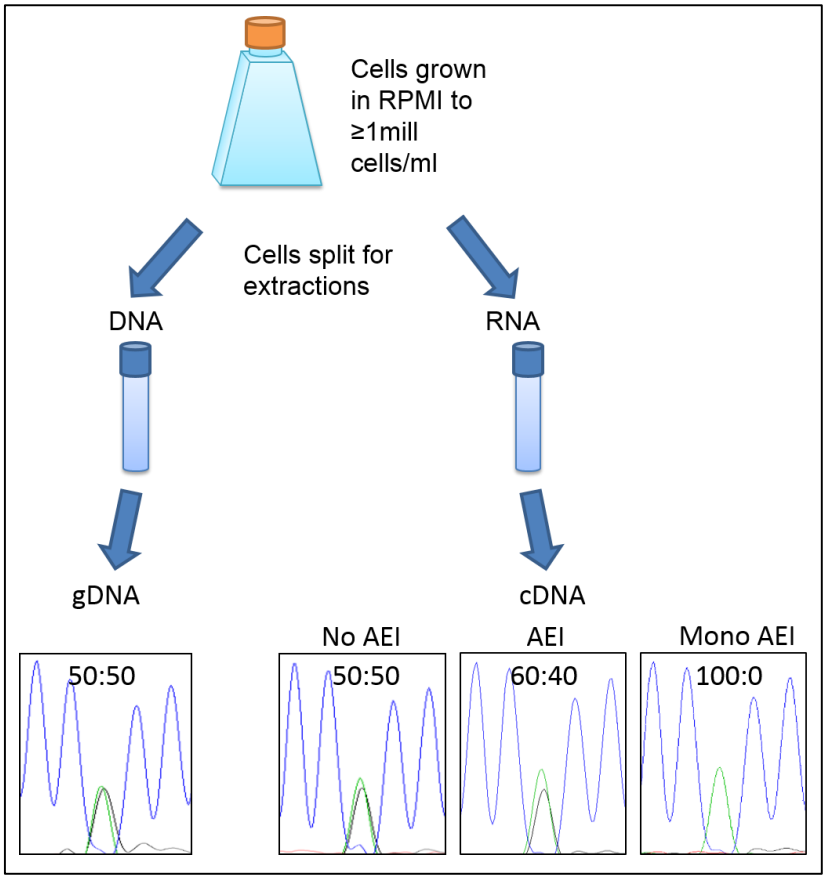

Figure 6.4 Outline of the experimental procedure for extracting RNA and DNA from the LCLs. Cells grown from one flask will be split for DNA and RNA extractions to create matched DNA and RNA pairs for each cell line. This was repeated in triplicate (not shown). RNA will be converted to cDNA, amplified by PCR and sequenced as will the gDNA. Allelic expression imbalance (AEI) will be seen in the cDNA as deviation from a 50:50 ratio in the heterozygous peak normalised to the gDNA peak sequence.

For each cell line three flasks were grown to at least one million cells/ml and split into two 15ml Falcon tubes for pelleting to allow a DNA and RNA extraction to be performed on cells from the same flask (see figure 6.4). Although the cell lines originally come from a single individual, DNA and RNA extractions from the same flask will be treated as matched pairs for downstream statistical analysis. gDNA was extracted from the cell lines using the QIAamp DNA Blood mini kit (Qiagen) according to manufacturer's instructions and stored at -20°C. RNA was extracted from the cell lines using RNeasy mini kit (Qiagen) following manufacturer's protocol and stored at -80°C. RNA was converted to complementary DNA (cDNA) (see Chapter 2, section 2.1.4).

If any allelic imbalance is detected for the coding variant predicted to create a premature stop codon (rs201074739), cell line HG00137 will be grown as

before, but split into six flasks with three of the flasks being treated with 200µg/ml puromycin (Sigma-Aldrich) for 6 hours prior to the extractions. Puromycin is a protein translation inhibitor that inactivates the 60S ribosomal subunit and destabilizes the polysome. The polysome is a structure that contains numerous ribosomes, including 60S, that read the same messenger RNA (mRNA) transcript simultaneously synthesizing protein (Qian et al. 1993; Carter et al. 1995). Transcripts are targeted for nonsense-mediated decay (NMD) following scanning with the 60S ribosome, therefore treatment with puromycin interrupts this process and prevents the transcript from being targeted for NMD. If the allelic imbalance is removed by the puromycin treatment then the transcript is very likely being targeted for NMD.

6.2.2.3. Primer design, PCR and sequencing

Primers were designed as described in Chapter 2, section 2.1.3. Where possible, cDNA primers were designed to span introns or exon/exon boundaries to ensure the amplification of cDNA rather than any gDNA which may be present (table 6.2).

Primers for the housekeeping gene *HPRT1* (hypoxanthine phosphoribosyltransferase 1) were obtained from Darryl Jackson (Research Technician, Faculty of Medicine and Health Sciences, University of Nottingham) (table 6.2). These primers were used to check the cDNA synthesis for any sample which did not amplify using the variant specific primers in table 6.2.

All PCRs were carried out using 10-100ng template DNA in total volumes of 30µl containing 1X PCR Buffer (Roche), 0.2mM of each dNTP, 1 U Taq DNA polymerase (Roche), 1pmol/µl of each primer. PCR reactions were performed in a Veriti 96-Well Fast Thermal Cycler (Applied Biosystems). PCR conditions for all reactions were as follows: an initial denaturing step of 94°C for 2 min, followed by 30 cycles of 94°C for 30 sec, optimized annealing temperature (Ta, see table 5.2) for 30 sec, 72°C for 1 min. This was followed by a final extension step of 72°C for 7 min. PCR products were visualized on 1% agarose gel stained with ethidium bromide (see Chapter 2, section 2.1.4).

PCR products were cleaned up, sequenced and evaluated as described in section 2.1.5

Table 6.2. Primer sequences for the allelic expression imbalance study. Template DNA (Templ. DNA) indicates which DNA template will be amplified from the primer pair. PCR product size is indicated in base pairs (bp), and annealing temperature (Ta) is indicated in **°C**.

Gene and SNP ID	**Templ. DNA**	**Sense primer (location)**	**Antisense primer (location)**	**Product size (bp)**	**Ta (°C)**
CD2AP rs151064033	gDNA/ cDNA	CD2AP_3UTRS1 (3' UTR) AAACAAACCCAGGCCTGATG	CD2AP_3UTRAS1 (3'UTR) TCTTCTGTTACACAAGGCCTG	389	58
CD2AP rs141029774	gDNA/ cDNA	CD2AP_3UTRS2 (3'UTR) TGCAGTTGATGTTGTAACCT	CD2AP_3UTRAS2 (3'UTR) ACTCTGAAATACACAAATGCTAAAGT	217	57
EPHA1 rs1131883	gDNA	EPHA1_I17S (intron 17) GAGCCTCCGGACTCATGC	EPHA1_3UTRAS (3'UTR) AGCACCGATAGCCTAGTCTG	235	59
	cDNA	EPHA1_E16S (exon 16) CTCATGAAGAACTGCTGGGC	As above, EPHA1_3UTRAS	476	59
ZYX rs11552744	gDNA	ZYX_I9S (intron 9) CCCTCACCCTTCCTTCTTCC	ZYX_3UTRAS (3'UTR) CCGCAAGCAGAGTACAAAGG	463	59
	cDNA	ZYX_E8/9S (spans exon 8 and 9) TACAAGTGTGAGGACTGCGG	As above, ZYX_3UTRAS	482	59
CD33 rs201074739	gDNA	CD33_E2bS (exon 2b) ACAGGCCCAAAATCCTCATC	CD33_I2bS (intron 2b) ACTAAATGTCCCCAGCACCA	347	58
	cDNA	As above, CD33_E2bS	CD33_E4AS (exon 3/4) CTGTAACACCAGCTCCTCCA	423	59
CD33 rs1803254	gDNA	CD33_I6S (intron 6) GACCCTCTTTGCCTTCTCCT	CD33_3UTRAS (3'UTR) GGTGTTTTCCATGTGTACCTTT	569	59
	cDNA	CD33_E6S (exon 6) GAATGACACCCACCCTACCA	As above, CD33_3UTRAS	535	59
HPRT1	cDNA	HPRT1_S AAATTCTTTGCTGACCTGCTG	HPRT1_AS TCCCCTGTTGACTGGTCATT	122	59

6.2.2.4. Allelic Expression analysis

Allelic expression imbalance (AEI) was assessed by using the custom built program, PeakPicker (v. 4) provided by Bing Ge from McGill University (Ge et al. 2005). The electropherograms for three gDNA replicates and the three cDNA replicates were loaded into the program and a gDNA sequence was selected as the reference sequence. The peaks of the heterozygous variant were selected in the reference sequence and the program normalised these values based on the peak heights of the other nucleotides in the sequence. These values were then used to normalise the rest of the loaded sequences and the ratios for the identified variant were calculated automatically by the software. AEI was determined as the ratio between the normalised alternative allele and the reference allele. A ratio difference of above 1.2 or below 0.8, which equates to an allelic ratio of 55:45 or 45:55, was used as a conservative cut off for a variant showing AEI. This ratio was shown by Ge et al. (Ge et al. 2005) to fall outside 95% confidence intervals for equal expression for sequencing data.

If there are any differences in AEI seen between brain tissue samples with the same variant of interest, VCFtools will be used as described in the General Methods (Chapter 2) to determine all exonic variants in linkage disequilibrium with the variant of interest using the exome sequencing data provided by R. Guerreiro (UCL). This will determine if there are any sequence differences between the samples which may explain the differences in AEI.

To calculate the significance between the cDNA and gDNA ratios in the brain tissue and for the LCL samples, a matched t-test was performed in SPSS (version 22, IBM Corp. Released 2011) where the sample size was greater than three.

6.3. Results

6.3.1. RNA extraction optimisation

The RNeasy Mini Kit (Qiagen) produced the best quality RNA from sample M669 with a RIN of 6 (table 6.3). Therefore this method was used to extract RNA from brain tissue for the AEI experiment.

Table 6.3. RNA quality assessed by RIN score for each different method and initial homogenisation. The two homogenisation methods used were transitioning frozen tissue from -80°C to -20°C overnight in RNA*later*-ICE (Ambion) prior to homogenisation with a pestle or homogenising tissue from -80°C with a pestle and mortar under liquid nitrogen. The homogenisation method using RNA*later*-ICE was not attempted for the RNeasy extraction due to poor RIN values obtained for the other extraction methods.

Method	**RIN scores for two methods of homogenisation**	
	RNA*later*-ICE	**Liquid nitrogen**
TRIzol and ethanol precipitation	<1	5.5
Direct-zol RNA MiniPrep (Zymo)	3.2	5.5
RNeasy Mini Kit (Qiagen)	-	6

Using the optimised RNA extraction method, RNeasy Mini Kit, the effect of the length of time a sample had been in storage (measured in years) and the RNA quality (measured as RIN) was assessed using a Spearman correlation coefficient (SPSS v22) which is more robust to outliers. Data was normally distributed for both age of sample and RIN (Shapiro-Wilk, $p = 0.132$ and $p = 0.849$; N=11), however one potential outlier was noticed in the scatter plot (see figure 6.5, A). There was a significant strong negative correlation between the length of time a sample had been in storage and the RIN value of the sample (Spearman's rho = -0.784, $p = 0.004$, N = 11) that improved with the removal of the outlier (Spearman's rho = -0.88, $p = 0.001$, N = 10). The longer a brain tissue sample had been stored, the lower the RNA quality obtained from that sample. To determine if RNA quality could be reliably predicted from the length of time a sample had been stored, a linear regression was performed (figure 6.5). Assuming the data point 27 years and 2.8 RIN is an outlier, the variation in the age of a sample can explain 69.9% of the variation in the RNA quality.

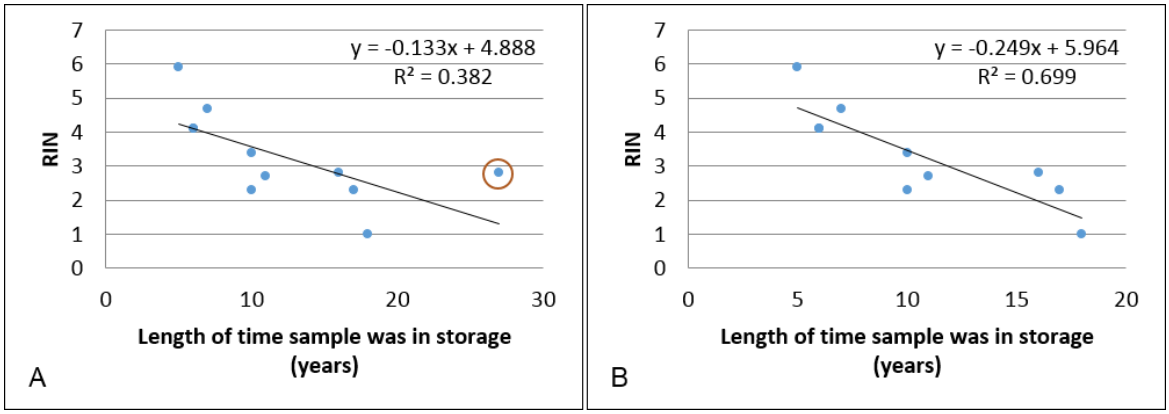

Figure 6.5 Scatter plot showing the relationship between the RNA quality (RIN) on the y-axis and the number of years the sample was in storage on the x axis for the data with (A) and without (B) the suspected outlier (27 years, 2.8 RIN) circled in red on (A). The equation of the regression line and the adjusted R^2 values are shown.

6.3.2. Allelic expression imbalance analysis

6.3.2.1. Brain tissue samples

PCR amplification of cDNA was only attempted for cerebral cortical brain tissue samples with a RIN greater than 3.0 as these samples consistently amplified the house-keeping gene *HPRT1* (table 6.4). The *ZYX* variant, rs11552744, was the only variant for which three brain samples provided sufficient quality RNA (RIN >3) for analysis. Therefore this was the only variant which could be statistically tested using matched pairs t-test in SPSS (v22). This variant showed no significant difference between the cDNA ratios and the gDNA ratios (mean difference 0.04 (95% CI, -0.11 to 0.19); t(2) = 1.21, p-value =0.35). Due to the small sample size, a nonparametric test was considered, however a similar result was obtained for the nonparametric Wilcoxon Signed Rank Test. Therefore the matched pairs t-test was selected as this reports 95% confidence intervals (CI) of the interval of the difference.

The *EPHA1* and *CD2AP* variants had one or two samples which provided sufficient quality RNA and these were examined for possible AEI without statistical testing (table 6.4, figure 6.6). The *CD2AP* variant rs151064033 was only present in one brain tissue sample which did not produce sufficient quality RNA to allow AEI to be examined (table 6.4). The *EPHA1* variant rs1131883 was only examined in one sample as this was the only sample to produce RNA with a RIN greater than 3. Only sample M596 for the *CD2AP* variant rs141029774 had an average normalised cDNA ratio greater than the conservative cut off ratio of 1.2.

Although *in silico* evidence suggested that all genes are expressed in cerebral cortical tissue, both *CD33* variants failed to amplify from any brain tissue samples. However PCR amplification was achieved for these samples with the housekeeping primers, *HPRT1* indicating successful RNA extraction and cDNA synthesis. The gDNA ratio was not calculated for these variants as no comparison can be made.

Apart from sample M596, no allelic expression imbalance was detected using the conservative cut-off of below 0.8 or above 1.2 (Ge et al. 2005), see table 6.4 and figure 6.6 A-C. As samples M596 and M660 produced differing AEI ratios for the *CD2AP* variant rs141029774, the exome variants for *CD2AP* were compared for these samples. No sequence differences were identified between samples M596 and M660 in the exome sequencing data provided by R. Guerreiro (UCL).

Table 6.4. RNA quality and normalised allele ratios for the SNPs as amplified from brain tissue. If the sample had a RIN less than 3 and failed to amplify house-keeping primers it was excluded from the study. CD33 did not amplify from the brain tissue despite multiple attempts. The gDNA and cDNA allele ratios shown are the average normalised allele ratios with the standard deviation in brackets.

Gene and SNP ID	**Location**	**Ref/Alt**	**MAF**	**Brain sample**	**RIN**	**gDNA (Alt/Ref allele)**	**cDNA (Alt/Ref allele)**
CD2AP rs151064033	6:47594204	A/T	0.019	M546	1	-	-
CD2AP rs141029774	6:47594318	G/A	0.011	M596	3.0	0.96 (± 0.01)	1.28 (± 0.12)
				M602	2.8	-	-
				M660	4.7	0.95 (± 0.02)	0.94 (± 0.05)
EPHA1 rs1131883	7:143088531	C/T	0.030	M602	2.8	-	-
				M605	2.3	-	-
				M647	4.1	1.00 (± 0.03)	1.13 (± 0.09)
ZYX rs11552744	7:143087956	G/A	0.033	M636	2.3	-	-
				M647	4.1	1.00 (± 0.06)	1.11 (± 0.12)
				M648	4.1	1.00 (± 0.07)	1.04 (± 0.07)
				M650	5.9	1.00 (± 0.09)	0.98 (± 0.05)
CD33 rs201074739	19:51729104-51729107	CCGG/-	0.024	M649	4.0	-	-
				M651	4.1	-	-
CD33 rs1803254	19:51743144	G/C	0.039	M547	2.8	-	-
				M638	3.4	-	-

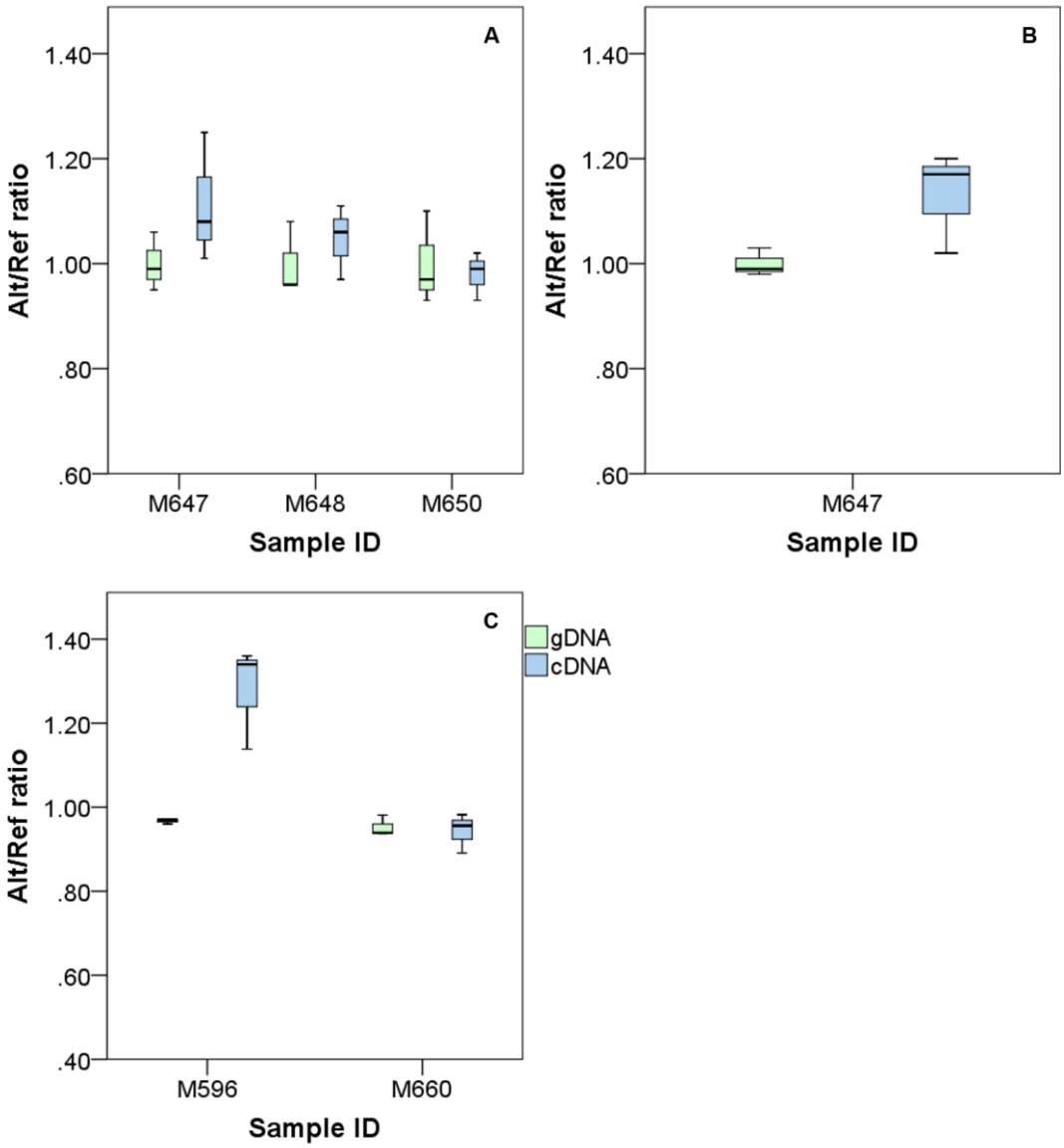

Figure 6.6. Average normalised alternative/reference allele ratios for gDNA (green) and cDNA (blue) from brain tissue samples as calculated in PeakPicker for (A) rs11552744 (ZYX), (B) rs1131883 (EPHA1) and (C) rs141029774 (CD2AP).

6.3.2.2. 1000 Genomes LCLs

The LCLs from 1000 Genomes each contained two of the variants from above therefore four variants could theoretically be tested (table 5.5). RIN values for cell lines were reproducibly above 9 (data not shown). However, *EPHA1* did not amplify from the cell line despite multiple attempts. The housekeeping gene, *HPRT1* did successfully amplify from the HG00137 cell line samples as did the *CD33* variant, rs201074739. Therefore it appears that *EPHA1* is not

expressed in the LCLs. A matched pair t test was performed in SPSS (v22) for the RNA and DNA extractions from the three separate flasks (table 6.5).
The two variants in HG00255, rs11552744 and rs1803254, did not show a significant allele ratio difference (table 6.5) although for variant rs1803254 the cDNA did have an average allele ratio higher than 1.2 (table 6.5, figure 6.7).

Table 6.5. RNA quality and normalised allele ratios for the SNPs as amplified from 1kG LCLs. A matched-pairs t-test was performed in SPSS (v22). The gDNA and cDNA ratios shown are the average normalised allele ratios (alternative/reference allele) obtained from PeakPicker.

Gene and SNP ID	**Cell line**	**gDNA ratio**	**cDNA ratio**	**p-value**
EPHA1 rs1131883	HG00137	-	-	-
ZYX rs11552744	HG00255	1.00 (± 0.09)	1.05 (± 0.01)	t(2) = -0.964; p = 0.437
CD33 rs1803254	HG00255	1.00 (± 0.04)	1.25 (± 0.28)	t(2) = -1.49; p = 0.275
CD33 rs201074739	HG00137	1.00 (± 0.05)	0.09 (± 0.02)	t(2) = 24.42; p = 0.002

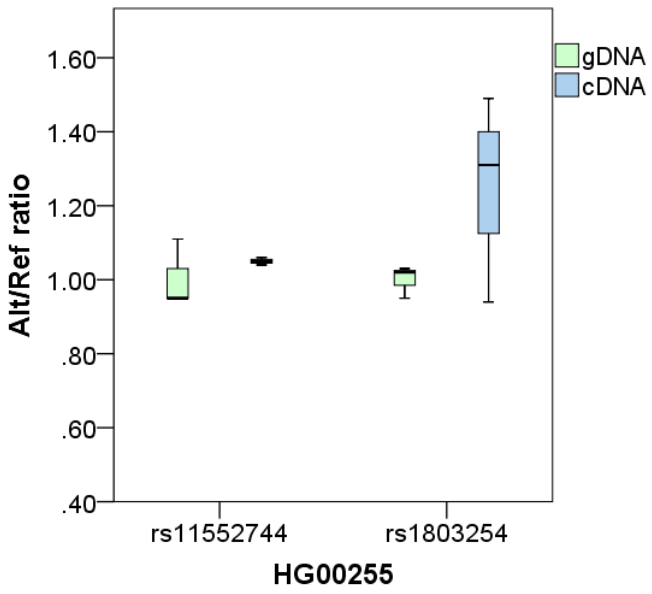

Figure 6.7. Average normalized allele ratios (alternative/reference allele) for the gDNA (green) and the cDNA (blue) for rs11552744 (ZYX) and rs1803254 (CD2AP) from 1000 Genomes cell line HG00255.

The frameshift variant in *CD33*, rs201074739 showed significant allele ratio difference (table 6.5, figure 6.8). Therefore the cell line was treated with 200µg/ml puromycin (Sigma-Aldrich) and the RNA and DNA was extracted from treated and untreated flasks as described in the methods. The puromycin treatment partially reduced the allele ratio difference in the cDNA from 0.09 (± 0.02) to 0.2 (± 0.02) (figure 6.8 and 6.9). The allele ratio difference for the puromycin treated cDNA was still significantly different to the gDNA ratio (gDNA ratio 1.00 (± 0.04); t(2) = 27.65; p = 0.001).

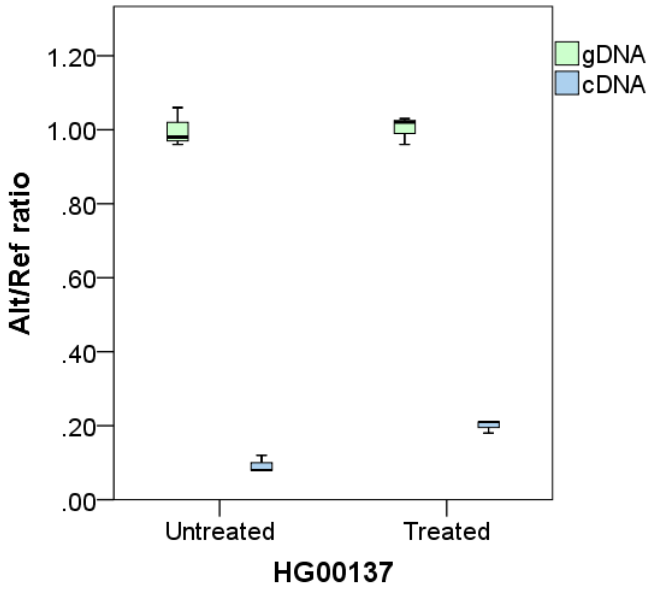

Figure 6.8. Average normalized allele ratios (alternative/reference allele) for gDNA (green) and cDNA (blue) for rs201074739 (*CD33*) from 1000 Genomes cell line HG00137, showing the untreated cell line ratios on the left and the ratios for the cell lines treated with 200µg/ml puromycin on the right.

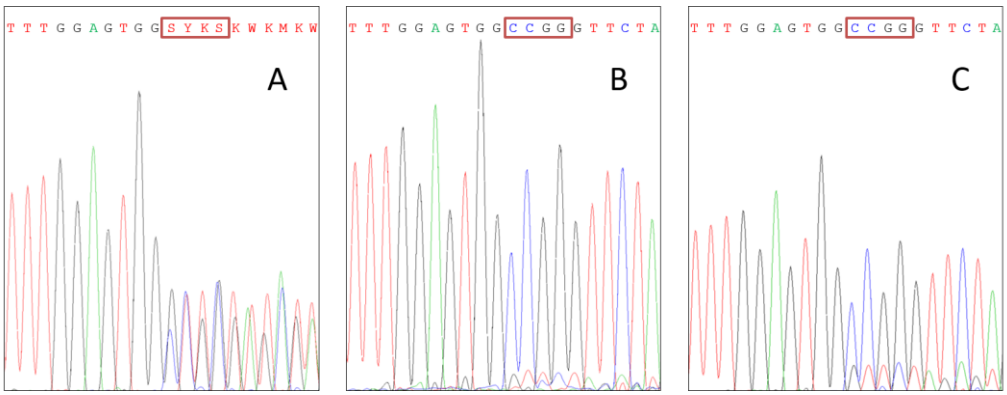

Figure 6.9. Electropherograms for the *CD33* frameshift variant, rs201074739 showing the gDNA sequence (A) with the deletion clearly visible. While the cDNA sequence faintly shows the deletion (B) which shows a slightly stronger signal after treatment with 200µg/ml puromycin (C).

6.4. Discussion

The main aims of this chapter were to use allelic expression imbalance (AEI) to validate variants identified in the next generation sequencing chapter (Chapter 3) which were predicted to affect the regulation of the Alzheimer's disease associated genes, *CD2AP*, *EPHA1* (including *ZYX*) and *CD33*. Cerebral cortical tissue and Epstein-Barr transformed human B-lymphoblastoid cell lines (LCLs) from 1000 Genomes were investigated for AEI in the prioritised variants. One brain tissue sample, M596, showed a cDNA allele ratio slightly greater than the cut off ratio of 1.2 for the *CD2AP* variant, rs141029774, however due to small sample size, the significance of this difference could not be tested for deviation from the genomic DNA ratio (table 6.4, figure 6.6). In the 1000 Genomes LCLs, two variants in *CD33* showed potential AEI. The variant, rs1803254 in the HG00255 cell line had a cDNA ratio greater than 1.2, however this was not significantly different from the gDNA ratio due to large standard deviation in the ratios (figure 6.7). The other variant, rs201074739, showed consistent and clear AEI in the HG00137 LCLs (figure 6.8), with little to no RNA being produced from the deletion allele (figure 6.9).

6.4.1. RNA quality

Transitioning the frozen brain tissue from -80°C to -20°C in RNA*later*-ICE did not increase the RNA quality for the test sample, M669. However RNA*later* has been found to be beneficial for other studies (Mutter et al. 2004). Many of the brain tissue samples in the Nottingham brain bank have been stored well over 5 years (figure 6.5) therefore it is possible that this longer storage time at -80°C affects the stability of the RNA in these particular tissues when transitioning with RNA*later*-ICE.

The length of time a brain tissue sample had been stored at -80°C accounted for 69.9% of the variation seen in the RNA quality of that brain tissue sample (figure 6.5). It should be noted however that it is often not sample storage which has the largest effect on RNA quality but the method of sample collection. In addition to post-mortem interval (PMI), the pH of a sample and how quickly the individual died (agonal state) have been found to have the largest effect on RNA quality from brain tissue (Chevyreva et al. 2008). The length of the agonal state also affects the pH of the tissue, with a fast cause of death leaving tissues with a high pH and subsequently increased RNA quality

(Chevyreva et al. 2008). Unfortunately this information is missing for these samples. It would be interesting to include PMI, pH of the brain tissue sample and the length of agonal state to see if these factors could account for the missing variation in RNA quality and produce a model which could better predict the RNA quality of a sample.

RIN values should be selected according to the type of analysis that will be done. Experiments can be successfully performed on samples with RIN as low as 1.4 (noted as “biologically relevant” (Ribeiro-Silva et al. 2007)). Given that this study examines the allelic ratios of RNA (converted to cDNA) it is important that the RNA is sufficiently intact to allow confidence that the allelic ratios obtained from the sample are representative of the ratios present in the brain tissue prior to death. However, it has been found that inter-sample variation actually accounts for more variation between experiments than RNA degradation as measured by RIN (Opitz et al. 2010). Therefore a RIN of 3 was selected as a cut off as samples with this RIN repeatedly amplified using the house-keeping primers for *HPRT1*.

6.4.2. Allelic expression imbalance investigation

6.4.2.1. AEI in brain tissue

Six samples out of fifteen did not produce RNA of sufficient quality for further analysis (table 5.4). *CD33* did not amplify at all from the RNA extracted in this project. While the Human Protein Atlas (http://www.proteinatlas.org/ Uhlén et al. 2005; Uhlen et al. 2010) appears to indicate CD33 expression in the cerebral cortex, CD33 protein is predominantly expressed in the microglia of the brain (Griciuc et al. 2013). Therefore it is possible that the RNA quantified in the Human Protein Atlas actually originates from microglia which may not be present at high numbers in the tissue sampled for this experiment which could explain why this gene did not amplify in these samples.

One sample (M596) for the *CD2AP* variant, rs141029774 did have an average ratio greater than 1.2 however the other sample tested for this variant (M660) did not. M596 had a RIN of 3.0 which is just on the quality score used for this experiment. The apparent AEI could be originating from unequal degradation of the two alleles rather than reflecting regulatory events. Examining the whole exome sequencing data (from R. Guerreiro, UCL) for this sample revealed no

additional variants compared with M660 indicating that it is not a difference in another exonic variant which is responsible for the allelic expression difference. However without sequencing the intronic sequence, we cannot tell if it could be a noncoding variant which is responsible for the nonsignificant AEI differences between these two samples.

No variants assessed in the brain tissue consistently showed AEI ratio above 1.2 or below 0.8 (table 6.4, figure 6.6). Therefore it can be assumed that these variants are not affecting *cis*-regulation. Additionally, variants identified as being in near or complete LD with these variants can also be assumed to be having minimal or no effect on gene regulation in *cis*.

6.4.2.2. AEI in 1000 Genomes LCLs

RNA-seq data indicated that *EPHA1* should be transcribed in the cell lines however in this project I was unable to amplify the transcript from this gene. The primers previously worked for the brain tissue samples and the cell line cDNA successfully amplified for the other gene investigated, *CD33*. It is possible that differences in the preparations of the cDNA for the two experiments mean that the cDNA for the RNA-seq ensured that the *EPHA1* transcript was better captured or possibly that the 3'UTR was not captured in the cDNA preparation for the RT-PCR analysis. It is also possible that the slight differences in laboratory tissue culture growth conditions would mean that the cells grown here did not express *EPHA1*.

While EBV transformed LCLs are used as a renewable source of genetic material for a large number of repositories such as HapMap and 1000 Genomes, the reliability of LCL as a model organism has not been widely studied. Repeated freeze-thaw cycles of LCLs have been found to cause the cell culture to mature into a self-renewing LCL culture with homogeneous gene expression very different from newly established LCLs which maintain individual specific variation in gene expression and qualities of primary B cell biology (Çalışkan et al. 2014). Additionally, other studies have shown that EBV can change the gene expression of the transformed LCLs depending on the number of copies of the virus which has inserted into the cell during the transformation (Caliskan et al. 2011; Houldcroft et al. 2014). Therefore it is possible that the gene expression of the cell line obtained from Coriell Institute for this analysis is exhibiting a different gene expression profile to the cell line

which was used for the RNA-seq analysis, explaining the lack of *EPHA1* expression seen in the cell lines in this project.

The *ZYX* variant, rs11552744 did not show any AEI (figure 6.7). This was consistent with the results from the brain tissue samples (figure 6.6). This variant is in the 3'UTR of the gene and was predicted to modify an enhancer region. Enhancer regions in the 3'UTR have been shown to fold forward and interact with the promoter region of a gene (Mao et al. 2010; Jash et al. 2012), although in this instance the variant is having no effect on gene expression measured using AEI.

The two *CD33* variants, rs1803254 and rs201074739 both showed AEI that deviated outside the cutoff used in this study of above 1.2 or below 0.8. However this was only significant in the frameshift deletion variant, rs201074739 (figure 6.8 and 6.9). The large standard deviation for the nonsignificant variant, rs1803254 could indicate that in come cell cultures the variant is exhibiting AEI, while in others it is not. This variant falls in the 3'UTR of two of the most common isoforms for *CD33* but is not present in the third (See figure 4.1, Chapter 4). It could be that the AEI seen is due to a change in the ratios of the three isoforms being produced in the cells. To investigate this, the levels of the three different transcript isoforms for *CD33* should be measured to see if there is a correlation with the AEI seen for this variant.

The variant rs201074739 is a 4bp deletion in exon 2b of *CD33* and is present in all three RefSeq transcripts. If RNA were transcribed from the deletion allele, this would create a truncated protein lacking the two main functional domains, the signaling and the immunological domains (as predicted in Chapter 4, figure 4.2). However following treatment with puromycin the AEI is partially recovered, indicating that this NMD inhibitor can alleviate the allelic imbalance observed. This lends support to the theory that transcript produced from the deletion allele is targeted for NMD. NMD is a mechanism which degrades mRNA containing premature termination codons and ensures that truncated proteins are not created (Chang et al. 2007). Therefore it is not surprising that transcript from the deletion allele, rs201074739, is not maintained as mRNA but rather quickly targeted for degradation. It could be possible that an alternative variant in LD with the frameshift variant is responsible for the AEI seen and for the reason that the NMD does not

completely rescue the phenotype, however examining the 1000 Genomes phase 3 data for the variant shows no other variants in strong LD which could be causing the AEI. For a more thorough investigation into the effects of puromycin inhibition of the NMD, a time delay series should be performed using a range of concentrations of puromycin and times of treatment. The treatment was selected from the literature as effective concentrations and times of treatment in LCLs ranged from 100µg/ml to 500µg/ml for 2 to 14 hours (Andreutti-Zaugg et al. 1997; Lamba et al. 2003; Chen et al. 2006; Castellsagué et al. 2010; Nguyen-Dumont et al. 2011) and as a treatment of 100µg/ml for 2 hours had been reported to inhibit more than 90% of protein synthesis in LCLs (Lamba et al. 2003).

6.4.3. Conclusions

As a method for quickly screening putative *cis*-regulatory variants, allelic expression imbalance is a quick and efficient way to identify variants worthy of putting forward for more elaborate time-consuming functional studies.

Only one of the variants investigated showed significant AEI, however this variant, rs201074739 was shown to be likely targeted for NMD. Therefore the functional variants responsible for the GWAS signals in *CD2AP, EPHA1* and *CD33* remain to be found.

7. General Discussion

The aim of this project was to identify and assess rare variants at the LOAD-risk associated gene loci, *CD2AP, EPHA1* and *CD33.* The three loci were sequenced using targeted whole genome sequencing which identified over a thousand variants across the three genes (Chapter 3). In Chapter 4, variants were prioritized for causality using both functional annotation and potential LOAD association. For functional annotation, numerous *in silico* annotation tools were used to provide functional support for causative variants. Variant association with LOAD was investigated using an imputed GWAS dataset. Prioritised variants were genotyped in an independent dataset to confirm potential LOAD association (Chapter 4). Finally, two laboratory methods were employed to validate functional predictions. The splicing variants were assessed with the minigene assay (Chapter 5) and allelic expression imbalance was used to investigate potential regulatory effects of variants in Chapter 6. Detailed discussions on the immediate findings of each chapter have already been presented. Therefore this chapter aims to consolidate the findings of this thesis as they relate to the wider field and also to address specific questions which arose through the duration of this PhD.

7.1. Deep sequencing GWAS loci as a method to identify disease causing variants

An emerging issue with resequencing studies is the sheer number of variants generated. This was an issue encountered in this thesis which uncovered over a thousand variants in the three GWAS genes, *CD2AP*, *EPHA1* and *CD33* (Chapter 3, table 3.8). In order to increase confidence in the variants called, three variant calling algorithms were tested (Chapter 3, table 3.7). The variants were also filtered on quality even before annotation occurred (Chapter 3, section 3.2.3.5.1), effectively reducing the number of variants from 1273 to 831. Support for this filtering was shown in the improvement in the TsTv ratios following filtering, which better approximated the 1000 Genomes dataset (Chapter 3, table 3.9). Additionally, variants were validated using exome sequencing as an alternative sequencing method.

One of the most well-known deep resequencing studies of GWAS loci was for inflammatory bowel disease (Rivas et al. 2011). This study used pooled sequencing of 56 genes in 350 cases and 350 controls with follow up

genotyping of 70 rare and low-frequency protein changing variants to identify new rare and functional variants for further study. Following this, several other resequencing studies for GWAS loci have been published (examples from LOAD GWAS loci include *CLU* (Bettens et al. 2012), *CLU*, *PICALM* and *CR1* (Lord et al. 2012) and *ABCA7* (Cuyvers et al. 2015)). However, ultimately assigning causality to variants identified from these studies is difficult (MacArthur et al. 2014).

In 2015 a targeted whole exome resequencing study for the LOAD GWAS genes *ABCA7*, *BIN1*, *CD2AP*, *CLU*, *CR1*, *EPHA1* and *MS4A4*/*MS4A6A* (excluding *CD33*) was published. Vardarajan et al. (2015) discovered rare deleterious coding mutations which recurred among unrelated patients and showed some evidence of segregation within families. They did not examine *CD33*, however they found a deleterious variant in *EPHA1* (p.P460L, rs202178565) which was significantly associated with disease in a Caribbean Hispanic family. This variant was also called in the NGS dataset from this study (Chapter 4, Figure 4.1 (b), table 4.3) although with a minor allele frequency of 0.007 it was not put forward for genotyping as the available dataset would not have sufficient power to detect any association (See Chapter 2, section 2.2.1.2 and figure 2.1). No missense variants were found in *CD2AP* in this dataset, however Vardarajan et al. identified a missense variant in *CD2AP* (p.K633R, rs116754410) which was found to be significantly associated with disease in the Caucasian dataset. The conclusions of the paper were that multiple rare coding mutations are found in LOAD-associated GWAS loci which could combine with common noncoding variants as being independently associated with LOAD. Certainly further investigation is required as to the potential functional mechanisms for the involvement of these rare coding variants in LOAD progression. Particularly given the finding from the 1000 Genomes project which discovered several deleterious rare coding variants with apparently no effect in their healthy control samples (Xue et al. 2012).

NGS has proven indispensable for uncovering rare variants and has discovered disease risk alleles in many complex diseases, including Alzheimer's disease, however there are a number of cases where it is unable to detect causal variants due to the experimental process involved (reviewed in Koboldt et al. 2013). This could include incorrect annotation of synonymous

and noncoding variants which may have a regulatory or splicing function. The development of functional annotation databases such as ENCODE (Consortium 2011) and GTEx (Ardlie et al. 2015) are beginning to address this issue. However, the causal variant could still be missed by NGS if it is not called in the sequencing data due to insufficient read depth or if the variant is a structural variant, such as an inversion or duplication (Koboldt et al. 2013). The sequencing method in this study may have missed a structural variant, however the read depth was high for this study making it unlikely that any variants were missed due to this (average depth of coverage per chromosome ranged from 290 for *EPHA1* to 434 for *CD2AP,* Chapter 3, table 3.8).

Further complicating causal variant identification from GWAS studies is the possibility that it may be multiple variants within the LD block rather than a single variant which may be the functional unit. This has been shown for noncoding variants in regulatory regions affecting regulatory element activity (Corradin et al. 2014) with support from allelic effect on gene expression (Guo et al. 2015). Following these studies, it is apparent that association testing single variants in isolation may not be the best approach and a more systematic strategy to test variants function and haplotypes within a GWAS locus is needed (Lowe and Reddy 2015).

The specific next generation sequencing (NGS) approach used in this project was targeted genome sequencing using pooled individuals. NGS is a fast evolving area with intense competition between life technology companies to produce ever longer reads at lower prices. However, pooled sequencing (or pool-seq) remains an attractive option for researchers (Schlötterer et al. 2014). Given the reduction in sequencing costs it is now feasible to whole genome sequence pools of individuals without the need of targeting specific loci. However, if pool-seq is used without tags for identifying sequences from individual samples to separate individuals during analysis, the phasing of haplotypes is impossible. Additionally, follow up functional work requires another sequencing step to identify the individual sample which contains the mutation. This was done before the splicing experiments in Chapter 5 using Sanger sequencing and Exome sequencing data (provided by R. Guerreiro, UCL) was used to identify individuals containing the variant of interest for the allelic expression imbalance investigation in Chapter 6.

With the publication of the final phase of the 1000 Genomes project (Auton et al. 2015) and initiatives such as the UK10K (Walter et al. 2015), it is likely that all low frequency (MAF <5%) and the majority of rare (MAF <1%) variants for many populations will soon be documented in databases. Any remaining lower frequency variants to be discovered presumably exist as personal de novo mutations or as extremely regional or population-specific variants. This increased knowledge of genetic architecture will allow the reliable imputation of rare variants in existing GWAS databases by providing a more complete reference genotype panel for imputation. Both biallelic and multiallelic rare variants (as low as 0.2% MAF) have already been successfully imputed using the reference panel from the final Phase 3 release of the 1000 Genomes project (Auton et al. 2015) and the UK10 haplotype reference panel (Huang et al. 2015). Therefore it is likely that NGS studies to uncover variants within the linkage disequilibrium regions of GWAS genes will not be necessary. What will become important will be ways of prioritising likely causative variants from the thousands of variants found within a GWAS LD block. NGS would still have a place in identifying novel rare variants which associate with disease and in genetic studies of small regional populations or families.

7.2. Are rare variants causing the disease association of GWAS risk loci for common complex diseases?

The genetic architecture underlying common complex diseases is not well defined. The initial common disease common variant (CDCV) hypothesis (Gibson 2011) predicted that disease risk would be accounted for by common variants present in all populations, and that these variants would have a combined or additive effect on disease risk. Given the lack of complete accountability for the heritability of complex diseases with the common disease risk variants currently identified (the “missing heritability” problem (Manolio et al. 2009)), several alternative hypotheses have been proposed. Large numbers of small-effect common variants, small numbers of large-effect rare variants, large numbers of large-effect rare variants (the common disease rare variant (CDRV) hypothesis) or a combination of genotypic, environmental and epigenetic interactions have all been proposed (Gibson 2011). It is worth noting that the definition of what constitutes a rare variant differs seemingly on author preference (Saint Pierre and Génin 2014). This PhD project defined rare as a variant with a MAF <5%, however these variants are termed “low-frequency” in other studies, with rare variants referring to those with MAF

<1%. Following evidence of complex disease association with common, low frequency and rare variants (Fu et al. 2013; Panoutsopoulou et al. 2013; Ratnapriya et al. 2014) it seems plausible that a combination of these effects may be responsible. The only model which has been clearly refuted following current resequencing studies was that of a small number of rare variants with large-effects (Saint Pierre and Génin 2014).

Understanding human demography and evolutionary process will provide further insight into the possible genetic architecture underlying common complex diseases. Considering protein coding variants identified to date in the UK10 dataset (Walter et al. 2015), functional rare (MAF <5%) variants are found five times more than functional common variants, supporting the theory that purifying selection is acting on these coding variants (Tennessen et al. 2012). It is now known that large numbers of rare variants are likely found in the human genome due to a rapid population expansion over the last 5000 years (Simons et al. 2014; Auton et al. 2015). Population genetics models comparing African and European populations show that recent demography has not changed the number of deleterious mutations present in the population and that rare mutations do not necessarily contribute to complex disease genetic risk (Simons et al. 2014).

Potentially deleterious functional rare variants uncovered at some of the LOAD GWAS loci through resequencing studies have already been discussed above. Rare variants in the GWAS locus *SORL1* were identified associating with an autosomal dominant early onset form of Alzheimer's disease (Pottier et al. 2012) prior to the gene being identified in the 2013 IGAP meta-analysis (Lambert et al. 2013). *SORL1* had been previously identified as associating with AD in candidate gene studies as SORL1 is a neuronal APOE receptor (Rogaeva et al. 2007). Two recent resequencing studies focusing on *ABCA7* have generated differing conclusions for this locus. Rare loss-of-function variants in *ABCA7* were found to increase the risk of LOAD in Icelandic, European and American populations (Steinberg et al. 2015). However, the loss-of-function variants were not occurring "on the background of the rs4147929[A] common variant" associated with AD (Steinberg et al. 2015). Therefore these rare variants appear to be associated with LOAD independently from the common GWAS variant. A Belgian resequencing study on *ABCA7* using targeted WGS found an intronic low-frequency variant

(rs78117248) which was suggested to be responsible for the *ABCA7* GWAS association (Cuyvers et al. 2015). It appears that even resequencing studies targeting a single locus have not clarified the debate.

To fully understand what role rare variants play in the genetics of complex diseases, different methods will be needed (Saint Pierre and Génin 2014). To uncover disease associations for rare variants, samples of over 10 000 cases and controls would be needed (Zuk et al. 2014). Therefore different approaches should be tried such as spatial genetic epidemiology where populations at fine geographic scale are genotyped to detect low-frequency genetic variants concentrated in small geographic regions (Saint Pierre and Génin 2014).

7.2.1. Rare coding variants in novel genes associated with late onset Alzheimer's disease

Whether rare variants are responsible for the association detected through GWAS or not, rare coding variants in novel genes have been identified as being associated with increased LOAD risk in European and other populations. Rare coding variants with large effects clearly implicate a gene with a disease and allow for easier translation to functional investigation. The most well-known rare coding variant for LOAD is in *TREM2* (rs75932628, R47H). This variant was concurrently identified by Jonsson et al. 2013 in an Icelandic cohort and by Guerreiro et al. 2013 in European samples. The 2013 IGAP meta-analysis identified the locus as being associated with LOAD, with the tag variant rs9381040 ($p = 6.3 \times 10^{-7}$), located roughly 5.5kb away from 3' end of *TREML2* and 24kb from 5'end of *TREM2* (Lambert et al. 2013). A further study investigating the association signal at the area determined that rare variants in *TREML2* also appear to be influencing Alzheimer's disease risk (Benitez et al. 2014). Although no follow-up studies on *TREML2* appear to have been published yet. *TREM2* is expressed on myeloid cells including microglia, monocyte-derived dendritic cells, osteoclasts and bone-marrow derived macrophages. Therefore the gene clearly contributes to the immune pathways already implicated in LOAD through other genes identified through GWAS (Lambert et al. 2013). The association of *TREM2* with Alzheimer's disease has been well replicated in several different cohorts (reviewed in Del-Aguila et al. 2015).

The second novel gene and rare variant identified through WES and WGS techniques is *PLD3* (rs145999145, V232M) (Cruchaga et al. 2013). A replication study in a German population was able to replicate the gene but not the variant association with AD (Schulte et al. 2015). A Belgian replication study did find association for rs145999145, however they also found that frequency of the variant varied considerably across populations (van der Lee et al. 2015). This emphasises the difficulties replicating rare variant associations which often need careful matching of cases and controls for ethnic backgrounds. Population stratification is an important issue to consider in association studies, particularly when examining rare variants. If data from ancestry informative markers are available, population stratification can be corrected for but these data are often missing from NGS studies. Two other large cohort studies, Heilmann et al. (2015) and Lambert et al. (2015) were unable to replicate the association of *PLD3* with LOAD. This has led to the actual association of *PLD3* with Alzheimer's disease being called into question through a series of brief communications published in Nature in April 2015 (Cruchaga and Goate 2015; Heilmann et al. 2015; Hooli et al. 2015; Lambert et al. 2015). The role of *PLD3* in Alzheimer's disease remains to be resolved.

Other novel coding variants have been found in *UNC5C* and *AKAP9*. The variant in *UNC5C*, rs137875858, T835M was found through the combination of WES and WGS and a linkage analysis in a large European cohort of LOAD. The variant is associated with increased disease risk (Wetzel-Smith et al. 2014). The finding was replicated across 4 cohorts. Two rare variants in *AKAP9*, rs144662445 and rs149979685 (MAF 0.43 and 0.36% respectively) associate with LOAD in an African American cohort (Logue et al. 2014). These variants are not found in European populations so appear to be unique to African populations (Logue et al. 2014).

The only novel coding variant which has consistently been replicated as being associated with AD is *TREM2*. Replicating rare variant associations are difficult as they may reflect population-specific mutations. Large datasets which are carefully matched for population ancestry could be used to replicate these novel gene associations with AD (Del-Aguila et al. 2015). *TREM2* with its role in immune response and inflammation has confirmed the involvement of these pathways in Alzheimer's disease, as immunity was flagged as

potentially important following genes identified in GWAS, *CR1*, *EPHA1*, *CLU*, *MS4A6A*, *CD33*, *HLA* and *INPP5D*.

7.3. Prioritising rare variants using functionality and imputation of existing GWAS datasets

The discovery of disease-associated loci through the unbiased approach of GWAS studies revolutionised common, complex disease genetics. However, subsequent deep resequencing studies aiming to document variants in GWAS loci, such as this PhD project, require prior functional knowledge in order to prioritise likely causative variants. Therefore they are directly dependent on available functional databases and prediction programs.

At the beginning of the PhD project, NGS was starting to become widely used by research groups to investigate common complex diseases. The issue of handling the huge numbers of variants generated through this technique was just being recognised. Prioritising these variants is a problem of multiple hypothesis testing that requires filtering the false positives (neutral variants) from the true positives (causal variants). It is likely that only one or a small number of variants will be truly pathogenic for a given disease, however at the outset all are just as likely (a priori) to be causal. Reliable methods of filtering are important to uncover true causative variants. For deep resequencing studies a quick way of reducing variants is by removing common variants. This is only appropriate, however, if the hypothesis is that rare variants are more plausible disease candidates.

Another common filtering method is on functionality which requires annotation through *in silico* predictions or through position comparisons with functional databases. However, annotation software available at the time of this PhD (see Chapter 4) had yet to completely catch up with the demand and were limited by the availability of prediction software and databases. Methods for predicting functionality of coding variants are the most well understood. Therefore it is perhaps unsurprising that many resequencing studies focused efforts on coding variants (e.g. Vardarajan et al. 2015). Using protein prediction algorithms such as SIFT and PolyPhen (as used in this thesis, Chapter 4, table 4.3) is an easy way for determining potential functional missense coding variants.

Another way of predicting possible functionality is through comparative genomics. Methods have been developed which can predict nucleotide-level constraint such as PhastCons and PhyloP (both used in this thesis, Chapter 4, table 4.3. and table 4.4). These use multiple species sequence alignments to estimate rates of evolutionary change compared with expected rates for neutral positions (which are not undergoing selection). PhyloP (Pollard et al. 2010) considers each nucleotide separately, while PhastCons (Siepel et al. 2005) uses a hidden Markov model to predict conserved elements, meaning that the score for a particular nucleotide will depend on the score of the surrounding sequence (Cooper and Shendure 2011). Neither PhastCons nor PhyloP include functional information and they are not allele specific. Regulatory elements evolve faster than coding regions (Schmidt et al. 2010), making conservation based methods less appropriate for annotating noncoding regions.

The filtering framework used in this thesis combined functional annotation with LOAD disease association to prioritise variants. Functional annotation was primarily provided by VEP, Ensembl's variant effect predictor (Chapter 4, section 4.2.1) (McLaren et al. 2010). Other programs available at the time included ANNOVAR (Wang et al. 2010b) and SNPNexus (Chelala et al. 2009; Dayem Ullah et al. 2012; Dayem Ullah et al. 2013). VEP was selected as access to the source code enabled the default annotation to be supplemented with ENCODE data including methylation and acetylation, DNase I hypersensitivity and transcription factor binding sites. This provided additional annotation for the noncoding variants (Chapter 4, table 4.5).

A potential drawback to VEP is the identification of splicing variants, which are only identified if they fall within 1-3bp of an exon or 3-8bp of an intron (McLaren et al. 2010). To remedy this, three freely available programs, ESEfinder (Chapter 4 table 4.3), BDGP Splice Site Prediction and HSF (Chapter 5, section 5.2.1 and section 5.3.1) were used in this thesis to prioritise potential splicing variants for further study. A new model from Xiong et al. 2015 uses machine-learning to examine DNA sequence variation and RNA splicing patterns from the same individuals and includes additional information such as intron and exon lengths, regulatory motifs, RNA secondary structure, RNA binding protein sites and trans-acting factors. The model predicts the amount of each exon included in different tissues,

effectively predicting alternative splicing. Using this program to annotate NGS variants in future would allow all variants, including those in noncoding or intronic regions to be assessed for potential functional activity. It is known that many splicing regulatory variants are missed as they are incorrectly annotated, and so may never be investigated for their effects on splicing (Singh and Cooper 2012).

Any potential functional variant cannot be implicated as being involved in LOAD if it is not also associated with the disease. Therefore as an additional filtering method in this thesis, an independent GWAS dataset was imputed to allow rare variants to be tested for association with LOAD (Chapter 4). Unfortunately given the lack of power of the imputed dataset the association of the GWAS tag variants in *CD2AP*, *EPHA1* or *CD33* could not be replicated (Chapter 4, table 4.6). Neither did any potentially associated variant identified from the imputed dataset show association in the independent Nottingham genotyping dataset (Chapter 4, table 4.12 and table 4.13). This is either due to a true lack of association which was incorrectly assumed in the imputed dataset (a false positive) or an incorrect assumption of the odds ratio (OR) of the variant.

The imputation performed in this thesis used the 1000 Genomes phase 1 data as a reference panel. This allowed rare variants which were not originally genotyped on the GWAS chip to be included for association testing. Additionally the control samples were supplemented with samples from the WTCCC2 dataset. These samples were genotyped on a different chip to the GWAS samples. Differential genotyping error can occur when imputing from different genotyping chips which could bias the association (Sinnott and Kraft 2012; Johnson et al. 2013). Additional biases can also be found across the same chip as association can be biased towards better genotyped SNPs. Additionally through re-using GWAS data, association can be biased towards the GWAS tag SNP and not towards the causal variant (Faye et al. 2013). However there is not yet a universal method suggested to control for these biases. It is possible that the accuracy of the imputed genotypes in the GWAS dataset could be improved through the use of the final 1000 Genomes Project genotyping panel or the UK10 consortium as the reference panel. Both have been found to substantially improve the prediction of rare variants (Auton et al. 2015; Walter et al. 2015).

During this PhD project, several functional databases have become available such as the GTEx project (Genotype-Tissue Expression www.gtexportal.org) and the human brain eQTL Almanac (Braineac www.braineac.org). Additionally, RNA-seq datasets are becoming more widely available for humans (e.g. Illumina Human Body Map) and there are RNA-seq data on cerebral cortical cells from mice including glia, neurons and vascular cells (http://web.stanford.edu/group/barres_lab/brain_rnaseq.html). All of these resources will be useful for future researchers wishing to prioritise potential deleterious variants.

As an example, data from GTEx (eQTL) and GEUVADIS (RNA-seq data) have been used to create a new predictive model for nonsense-mediated decay (NMD) of mRNA transcripts (Rivas et al. 2015). Using the frameshift variant in CD33 (rs201074739) as an example and searching for the variant in the GTEx database reveals that it is listed as a protein truncating variant on that website. This supports the findings from the AEI in Chapter 5 (figure 5.7 and 5.8). Protein-truncating variants (PTV) are usually assumed to have a huge effect on gene function. However, some PTV may be protective (Cohen et al. 2006), or even neutral (MacArthur et al. 2012). If PTV results in NMD they can be recessive and may protect against a detrimental phenotype or may cause disease via haploinsufficiency (Rivas et al. 2015). The ability to check *in silico* databases for functional evidence on these rare and low frequency variants will help steer future laboratory experimental design.

7.3.1. Difficulties annotating noncoding variants

Noncoding variants are likely to play important role in complex disease genetics. GWAS studies have revealed over 90% of the variants are noncoding, located outside of protein coding genes, with only around 10-15% of those in linkage disequilibrium (LD) with a protein coding variant (Maurano et al. 2012; Schaub et al. 2012). Functional noncoding variants, falling in transcription start sites, DNase I hypersensitivity regions and UTRs were found for both rare and common alleles in the UK10 dataset, confirming the importance of noncoding variants for determining phenotypic changes (Walter et al. 2015). It is difficult to annotate noncoding variants so many of the resequencing studies have focused on the coding variants. Nevertheless, noncoding variants can increase complex disease risk through influencing regulatory elements such as enhancers. Over 70% of GWAS variants have

been found to localise to areas of open chromatin thought to contain enhancers (as identified by DNase I hypersensitivity profiling (Maurano et al. 2012)). This has been confirmed by studies using histone marks associated with enhancer regions (H3K4me3 and H3K27ac) (Ernst et al. 2011; Akhtar-Zaidi et al. 2012; Trynka et al. 2013). Unfortunately, it is difficult to assess the impact of variants falling in enhancers as the target gene is often not known and a linear relationship between enhancer and nearest gene cannot be inferred (Dekker et al. 2013). It is also probable that gene expression is regulated by many regulatory elements which raises the possibility that multiple variants in different regulatory elements may function together to affect gene expression. This “multiple enhancer variant” hypothesis has been shown to occur in six common autoimmune diseases (Corradin et al. 2014).

In recent years several programs have been developed which are able to annotate noncoding variants. GWAS3D, one such program, was developed in 2013 (Li et al. 2013). Using chromatin state, functional genomics, sequence motifs and conservation, GWAS3D assesses the probability that a genetic variant will affect regulatory pathways and the underlying disease/trait. Changing the local chromosome conformation could block or generate looping interactions between distal elements and promoter regions influencing gene regulation (Sakabe et al. 2012). GWAVA (Genome Wide Annotation of Variants) is another program developed to annotate noncoding variants (Ritchie et al. 2014). This program uses ENCODE annotations and has a plugin for Ensembl VEP or it can be run through a webpage (https://www.sanger.ac.uk/resources/software/gwava/). GWAVA specifically predicts functionality of noncoding variants using several annotations for noncoding elements. In addition to the ENCODE annotations, data from the US National Institutes of Health Roadmap Epigenomics project and genome-wide properties like evolutionary conservation (genomic evolutionary rate profiling (GERP)), GC-content from Genome Reference Consortium and CpG context (Ritchie et al. 2014). The whole package provides a similar annotation to the method which was used in this thesis, combining the annotation provided by VEP with annotation provided by ENCODE (Chapter 4, section 4.2.1).

ANNOVAR has been updated to include the improved splicing predictions through the use of Xiong et al. 2015 machine learning splicing prediction

algorithm. Including this algorithm means that potential intronic and noncoding splicing variants can be annotated for splicing functionality. This would have been useful for prioritising potential splicing variants from the NGS data for Chapter 5. ANNOVAR has also been updated to include CADD (Combined Annotation Dependent Depletion) scores (Kircher et al. 2014) which were developed to annotate and prioritise noncoding regions for functionality. CADD scores any single-nucleotide variant or small indel through the integration of many annotations (including ENCODE) and outputs a single measure. Making it easier for researchers to prioritise variants (Kircher et al. 2014).

Using several high throughput techniques the ENCODE project has provided biochemical functionality for about 80% of the genome (Consortium 2011), identifying regions of transcription, transcription factor association, chromatin structure and histone modification. HaploReg (Ward and Kellis 2012) and RegulomeDB (Boyle et al. 2012) have used this information to annotate genetic variants and offer useful resources for annotating regulatory variation. However tissue or cell specific effects of the regulatory elements also need to be considered and these are not included in these databases. The integration of additional functional annotations for noncoding variants into annotation programs as these are developed will allow increased functional annotation for noncoding variants.

7.4. High throughput laboratory screening method for putative deleterious variants.

Even after filtering on variant call quality metrics, functional annotation and likely association, this thesis still had nine noncoding variants and 15 coding variants with MAF less than 5% and more than 5 or 6 lines of functional evidence (Chapter 4, table 4.9 and table 4.10). If this study was extended to encompass all GWAS genes currently identified as being associated with LOAD, and it is hypothesized that all genes would carry a similar number of potential prioritized variants, around 200 variants would be expected to look ‘promising’ for all 21 LOAD GWAS hits. Therefore a high throughput laboratory screening method would be needed to assess these variants.

This thesis used allelic expression imbalance (AEI) as a method for screening several putative disease-causing variants (Chapter 6). One of the main

reasons for selecting this method was that it could also be extended to the noncoding variants predicted to affect regulatory regions. AEI can also be made high throughput using RNA-seq instead of RT-PCR and sequencing. However, given the associated costs of RNA-seq and the high quality RNA required, this was not attempted with the samples available for this study (See Chapter 6, section 6.4.1 for RNA quality from the Nottingham brain tissue samples).

Another way of approaching the problem of disease variant association is through associating genetic variants with gene expression. There are several databases of expression quantitative trait loci (eQTLs), the most recent being the Genotype-Tissue Expression (GTEx) consortium (Ardlie et al. 2015). The combination of GWAS and eQTL analyses may shed light on which gene is reflecting the disease association. This is particularly useful when the gene tagged by GWAS is not in LD with a phenotype-associated variant or if there isn't clear biologic rationale for the association. Tissue specific eQTLs will be particularly important in neurological conditions as the brain has a high degree of tissue specific gene regulation.

Given the large role which regulatory variants appear to play in complex diseases, high throughput laboratory methods for investigating these variants definitely need to be considered. Using methods which investigate regulatory elements such as ChIPseq, DNase I hypersensitivity regions, methylation and acetylation can reveal allele specific differences in gene regulation and will be useful for the high throughput functional analysis of GWAS loci for complex diseases (Lowe and Reddy 2015).

7.5. Future of LOAD genetics

Investigation into the genetics of LOAD is far from over. The most recent GWAS has certainly uncovered many more genetic risk loci for the disease, however the roles each of the 21 loci play in the disease progression still need to be elucidated. Additional heritability for LOAD still remains to be discovered, with the 21 currently identified GWAS only accounting for about 60% of the population attributable risk of the disease (Medway and Morgan 2014).

Further work will likely focus on pulling apart the molecular mechanisms involved in the disease. With the majority of GWAS variants identified in

noncoding regions (Lambert et al. 2013), future research could focus on untangling the chromatin organisation around these areas. Chromatin is arranged in three dimensions (3D), with regulatory regions and genomic loci interacting within specific topologically associated domains (TADs). To fully understand the relationship between regulatory elements and the genes they regulate, these 3D interactions need to be mapped. Chromatin conformation capture (3C) based methods would investigate these 3D interactions. These methods use formaldehyde to crosslink chromatin, covalently bonding chromatin segments and joining DNA sequences which are in close proximity. The DNA is sheared with enzymes or sonication. The crosslinked DNA fragments are then ligated, purified and sequenced.

Through 3C based methods it has been found that regulatory elements cannot be assumed to interact with the closest gene (Dekker et al. 2013). The main mechanism that brings distal regulatory elements in close proximity to their target genes is chromatin looping (Sanyal et al. 2012). This results in regulatory elements interacting with genes several thousands of base pairs apart. Using an unbiased 3C method such as capture Hi-C would generate a genome-wide interaction map of all crosslinked loci (Jäger et al. 2015). Capture Hi-C increases the resolution of the original Hi-C method, and would allow for targeted sequence enrichment of GWAS loci to identify chromatin interactions across the genome (e.g. Jäger et al. 2015; Mifsud et al. 2015). A potential drawback to this method is the high starting quantity of DNA required for the 3C technique. However using this method could identify enhancer elements interacting with the LOAD genes, which would provide additional areas for resequencing.

Generating induced pluripotent stem cells (iPSCs) from patient samples has been proposed as an alternative *in vitro* model to current systems (Ross and Akimov 2014; Goldstein et al. 2015). This would enable functional experimental investigations of the genetic effects of identified potential causative variants to be undertaken in human cells induced to develop into neuronal, glial or other brain-derived cells. Thus facilitating experimentation on cells most likely being targeted or affected by the disease pathogenesis. Initial experiments with iPSCs focused on familial mutations in *APP*, *PSEN1* and *PSEN2* and on tests of Aβ toxicity on human neuronal cell types (Goldstein et al. 2015). iPSC work on sporadic late onset AD is less well developed.

Reprogramming patient iPSCs for familial and sporadic AD was first attempted in 2012 using primary fibroblasts (Israel et al. 2012). Hossini et al. 2015 successfully induced neuronal cells from iPSCs derived from dermal fibroblasts from one LOAD patient. It is likely that future iPSC work will enable the molecular mechanisms behind the disease pathogenesis to be better explored. The method will also allow for the generation of cell lines from LOAD patients containing specific disease causing haplotypes which will provide better models for the testing of emerging therapeutics.

Another possible direction for future research could investigate the involvement of somatic mutations in Alzheimer's disease. NGS investigation into somatic and germline mutation rates has revealed that different populations of cells within one individual can harbor different mutations and these genetically different populations of cells can contribute to disease (Lodato et al. 2015). For example a somatic mutation introduced into progenitor cells during the development of an individual's brain can cause neurological diseases and brain malformations (Poduri et al. 2013). Using single cell sequencing techniques, Lodato et al 2015 showed that an individual's brain appears to contain populations of cells with a range of different genotypes (Lodato et al. 2015). Frigerio et al 2015 used deep sequencing to successfully detect single-nucleotide mosaic variants in the entorhinal cortex of sporadic AD patients, suggesting that these variants may play a role in LOAD (Sala Frigerio et al. 2015). It is worth noting that many of the common complex diseases occur in later life and the variants implicated in these diseases have modest pathogenic effects. Therefore genetic causes could arise through somatic mutations. It is possible that somatic mutation might play a role in diseases other than cancer (Poduri et al. 2013; Shendure and Akey 2015). Further investigation as to the likely involvement of these mutations in LOAD will be required.

7.6. Conclusions

This thesis shows that pooled next generation sequencing of target enriched GWAS loci, *CD2AP*, *EPHA1* and *CD33* was successful in identifying potential functional rare variants in the three GWAS genes. Although none of the prioritised variants identified in this study were both functional and associated, it is clear that rare variants have a role to play in the common, complex diseases, including late onset Alzheimer's disease. Additionally, there is

undoubtedly room for improving methods for prioritising NGS variants, both coding and noncoding. As functional databases and imputation reference panels improve, so will the combined prioritisation strategy employed in future studies conducted along similar lines to those described in this thesis. Given the potential regulatory role for many GWAS variants, it is likely that future work using chromatin conformation capture based methods will further elucidate the role of this class of variants in LOAD. Insight provided by genetic studies on late onset Alzheimer's disease have already provided new avenues for drug development, disease treatment, diagnosis and screening. Future genetic work will be important to further elucidate the molecular mechanisms behind the progression of this devastating disease.

References

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467: 1061–73.

Abraham R, Moskvina V, Sims R, Hollingworth P, Morgan A, Georgieva L, Dowzell K, Cichon S, Hillmer AM, O'Donovan MC, Williams J, Owen MJ, et al. 2008. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. BMC Med. Genomics 1: 44.

Adoue V, Schiavi A, Light N, Almlöf JC, Lundmark P, Ge B, Kwan T, Caron M, Rönnblom L, Wang C, Chen S-H, Goodall AH, et al. 2014. Allelic expression mapping across cellular lineages to establish impact of non-coding SNPs. Mol. Syst. Biol. 10: 754.

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12: R18.

Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, Saiakhova A, Bartels CF, Balasubramanian D, Myeroff L, Lutterbaugh J, Jarrar A, Kalady MF, Willis J, Moore JH, et al. 2012. Epigenomic enhancer profiling defines a signature of colon cancer. Science 336: 736–9.

Akula N, Barb J, Jiang X, Wendland JR, Choi KH, Sen SK, Hou L, Chen DTW, Laje G, Johnson K, Lipska BK, Kleinman JE, et al. 2014. RNA-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and GTPase binding in bipolar disorder. Mol. Psychiatry 19: 1179–85.

Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, et al. 2011. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers. Dement. 7: 270–9.

Alladi S, Xuereb J, Bak T, Nestor P, Knibb J, Patterson K, Hodges JR. 2007. Focal cortical presentations of Alzheimer's disease. Brain 130: 2636–45.

Almlöf JC, Lundmark P, Lundmark A, Ge B, Maouche S, Göring HHH, Liljedahl U, Enström C, Brocheton J, Proust C, Godefroy T, Sambrook JG, et al. 2012. Powerful identification of cis-regulatory SNPs in human primary monocytes using allele-specific gene expression. PLoS One 7: e52260.

Altmann A, Weber P, Bader D, Preuss M, Binder EB, Müller-Myhsok B. 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum. Genet. 131: 1541–54.

Altmann A, Weber P, Quast C, Rex-Haffner M, Binder EB, Müller-Myhsok B. 2011. vipR: variant identification in pooled DNA using R. Bioinformatics 27: i77–84.

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, Bakker PIW de, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, et al. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–8.

Alzheimer A. 1907. Uber eine eigenartige Erkrankung der Hirnrinde. Allg. Zeitschrife Psychiatr. 64: 146–148.

Anderson CA, Soranzo N, Zeggini E, Barrett JC. 2011. Synthetic Associations Are Unlikely to Account for Many Common Disease Genome-Wide Association Signals. PLoS Biol. 9: 5.

Andreutti-Zaugg C, Scott RJ, Iggo R. 1997. Inhibition of nonsense-mediated messenger RNA decay in clinical samples facilitates detection of human MSH2

mutations with an in vivo fusion protein assay and conventional techniques. Cancer Res. 57: 3288–93.

Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-.). 348: 648–660.

Arendt T. 2009. Synaptic degeneration in Alzheimer's disease. Acta Neuropathol. 118: 167–79.

Arisi I, D'Onofrio M, Brandi R, Felsani A, Capsoni S, Drovandi G, Felici G, Weitschek E, Bertolazzi P, Cattaneo A. 2011. Gene expression biomarkers in the brain of a mouse model for Alzheimer's disease: mining of microarray data by logic classification and feature selection. J. Alzheimers. Dis. 24: 721–38.

Asimit JL, Zeggini E. 2012. Imputation of rare variants in next-generation association studies. Hum. Hered. 74: 196–204.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, et al. 2015. A global reference for human genetic variation. Nature 526: 68–74.

Bachmeier C, Paris D, Beaulieu-Abdelahad D, Mouzon B, Mullan M, Crawford F. 2013. A multifaceted role for apoE in the clearance of beta-amyloid across the blood-brain barrier. Neurodegener. Dis. 11: 13–21.

Bai B, Hales CM, Chen P-C, Gozal Y, Dammer EB, Fritz JJ, Wang X, Xia Q, Duong DM, Street C, Cantero G, Cheng D, et al. 2013. U1 small nuclear ribonucleoprotein complex and RNA splicing alterations in Alzheimer's disease. Proc. Natl. Acad. Sci. U. S. A. 110: 16562–7.

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet. 11: 773–85.

Bao R, Huang L, Andrade J, Tan W, Kibbe WA, Jiang H, Feng G. 2014. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. Cancer Inform. 13: 67–82.

Baralle D, Baralle M. 2005. Splicing in action: assessing disease causing sequence changes. J. Med. Genet. 42: 737–48.

Barnes WM. 1994. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proc. Natl. Acad. Sci. U. S. A. 91: 2216–20.

Basak JM, Verghese PB, Yoon H, Kim J, Holtzman DM. 2012. Low-density lipoprotein receptor represents an apolipoprotein E-independent pathway of Aβ uptake and degradation by astrocytes. J. Biol. Chem. 287: 13959–71.

Beecham GW, Martin ER, Li Y-J, Slifer MA, Gilbert JR, Haines JL, Pericak-Vance MA. 2009. Genome-wide association study implicates a chromosome 12 risk locus for late-onset Alzheimer disease. Am. J. Hum. Genet. 84: 35–43.

Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova J-L, Abel L. 2015. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc. Natl. Acad. Sci. U. S. A. 112: 5473–8.

Bell CG, Beck S. 2009. Advances in the identification and analysis of allele-specific expression. Genome Med. 1: 56.

Benitez BA, Jin SC, Guerreiro R, Graham R, Lord J, Harold D, Sims R, Lambert J-C, Gibbs JR, Bras J, Sassi C, Harari O, et al. 2014. Missense variant in TREML2 protects against Alzheimer's disease. Neurobiol. Aging 35: 1510.e19–26.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in

high-throughput sequencing. Nucleic Acids Res. 40: e72.

Berson A, Barbash S, Shaltiel G, Goll Y, Hanin G, Greenberg DS, Ketzef M, Becker AJ, Friedman A, Soreq H. 2012. Cholinergic-associated loss of hnRNP-A/B in Alzheimer's disease impairs cortical splicing and cognitive function in mice. EMBO Mol. Med. 4: 730–42.

Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, Hogan MF, Schjeide BMM, Hooli B, Divito J, Ionita I, Jiang H, Laird N, et al. 2008. Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. Am. J. Hum. Genet. 83: 623–32.

Betel D, Wilson M, Gabow A, Marks DS, Sander C. 2008. The microRNA.org resource: targets and expression. Nucleic Acids Res. 36: D149–53.

Bettens K, Brouwers N, Engelborghs S, Lambert J-C, Rogaeva E, Vandenberghe R, Bastard N Le, Pasquier F, Vermeulen S, Dongen J Van, Mattheijssens M, Peeters K, et al. 2012. Both common variations and rare non-synonymous substitutions and small insertion/deletions in CLU are associated with increased Alzheimer risk. Mol. Neurodegener. 7: 3.

Blennow K, Hampel H, Weiner M, Zetterberg H. 2010. Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. Nat. Rev. Neurol. 6: 131–44.

Bodger MP, Hart DN. 1998. Molecular cloning and functional analysis of the CD33 promoter. Br. J. Haematol. 102: 986–95.

Borenstein AR, Wu Y, Mortimer JA, Schellenberg GD, McCormick WC, Bowen JD, McCurry S, Larson EB. 2005. Developmental and vascular risk factors for Alzheimer's disease. Neurobiol. Aging 26: 325–34.

Bowen DM, Smith CB, White P, Davison AN. 1976. Neurotransmitter-related enzymes and indices of hypoxia in senile dementia and other abiotrophies. Brain 99: 459–96.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. 2012. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22: 1790–7.

Braae A, Thompson CE, Morgan K. 2014. Comparison of custom designed KASP and TaqMan genotyping assays for a rare genetic variant identified through resequencing GWAS loci. LGC Appl. note GAPP–0003.

Braak H, Braak E. 1991. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 82: 239–59.

Bradshaw EM, Chibnik LB, Keenan BT, Ottoboni L, Raj T, Tang A, Rosenkrantz LL, Imboywa S, Lee M, Korff A Von, Morris MC, Evans DA, et al. 2013. CD33 Alzheimer's disease locus: altered monocyte function and amyloid biology. Nat. Neurosci. 16: 848–50.

Bu G. 2009. Apolipoprotein E and its receptors in Alzheimer's disease: pathways, pathogenesis and therapy. Nat. Rev. Neurosci. 10: 333–44.

Caliskan M, Cusanovich DA, Ober C, Gilad Y. 2011. The effects of EBV transformation on gene expression levels and methylation profiles. Hum. Mol. Genet. 20: 1643–52.

Çalışkan M, Pritchard JK, Ober C, Gilad Y. 2014. The effect of freeze-thaw cycles on gene expression levels in lymphoblastoid cell lines. PLoS One 9: e107166.

Cao H, Bono B de, Belov K, Wong ES, Trowsdale J, Barrow AD. 2009. Comparative genomics indicates the mammalian CD33rSiglec locus evolved by an ancient large-scale inverse duplication and suggests all Siglecs share a common ancestral region. Immunogenetics 61: 401–17.

Cao H, Crocker PR. 2011. Evolution of CD33-related siglecs: regulating host immune functions and escaping pathogen exploitation? Immunology 132: 18–26.

Carrasquillo MM, Belbin O, Hunter TA, Ma L, Bisceglio GD, Zou F, Crook JE, Pankratz VS, Sando SB, Aasly JO, Barcikowska M, Wszolek ZK, et al. 2011. Replication of EPHA1 and CD33 associations with late-onset Alzheimer's disease: a multi-centre case-control study. Mol. Neurodegener. 6: 54.

Carrasquillo MM, Zou F, Pankratz VS, Wilcox SL, Ma L, Walker LP, Younkin SG, Younkin CS, Younkin LH, Bisceglio GD, Ertekin-Taner N, Crook JE, et al. 2009. Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nat. Genet. 41: 192–8.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat. Rev. Genet. 3: 285–98.

Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. 2003. ESEfinder: A web resource to identify exonic splicing enhancers. Nucleic Acids Res. 31: 3568–71.

Carter MS, Doskow J, Morris P, Li S, Nhim RP, Sandstedt S, Wilkinson MF. 1995. A Regulatory Mechanism That Detects Premature Nonsense Codons in T-cell Receptor Transcripts in Vivo Is Reversed by Protein Synthesis Inhibitors in Vitro. J. Biol. Chem. 270: 28995–29003.

Castellano JM, Kim J, Stewart FR, Jiang H, DeMattos RB, Patterson BW, Fagan AM, Morris JC, Mawuenyega KG, Cruchaga C, Goate AM, Bales KR, et al. 2011. Human apoE isoforms differentially regulate brain amyloid-β peptide clearance. Sci. Transl. Med. 3: 89ra57.

Castellsagué E, González S, Guinó E, Stevens KN, Borràs E, Raymond VM, Lázaro C, Blanco I, Gruber SB, Capellá G. 2010. Allele-specific expression of APC in adenomatous polyposis families. Gastroenterology 139: 439–47, 447.e1.

Chang Y-F, Imam JS, Wilkinson MF. 2007. The nonsense-mediated decay RNA surveillance pathway. Annu. Rev. Biochem. 76: 51–74.

Chartier-Harlin MC, Parfitt M, Legrain S, Pérez-Tur J, Brousseau T, Evans A, Berr C, Vidal O, Roques P, Gourlet V. 1994. Apolipoprotein E, epsilon 4 allele as a major risk factor for sporadic early and late-onset forms of Alzheimer's disease: analysis of the 19q13.2 chromosomal region. Hum. Mol. Genet. 3: 569–74.

Chelala C, Khan A, Lemoine NR. 2009. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics 25: 655–61.

Chen X, Truong T-TN, Weaver J, Bove BA, Cattie K, Armstrong BA, Daly MB, Godwin AK. 2006. Intronic alterations in BRCA1 and BRCA2: effect on mRNA splicing fidelity and expression. Hum. Mutat. 27: 427–35.

Chen Y, Fu AKY, Ip NY. 2012. Eph receptors at synapses: implications in neurodegenerative diseases. Cell. Signal. 24: 606–11.

Chevyreva I, Faull RLM, Green CR, Nicholson LFB. 2008. Assessing RNA quality in postmortem human brain tissue. Exp. Mol. Pathol. 84: 71–7.

Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N, Theologis A, Yang WH, et al. 1999. Genome-wide mapping with biallelic markers in Arabidopsis thaliana. Nat. Genet. 23: 203–7.

Christensen DZ, Schneider-Axmann T, Lucassen PJ, Bayer TA, Wirths O. 2010. Accumulation of intraneuronal Abeta correlates with ApoE4 genotype. Acta Neuropathol. 119: 555–66.

Chung SJ, Lee J-H, Kim SY, You S, Kim MJ, Lee J-Y, Koh J. 2012. Association of GWAS Top Hits With Late-onset Alzheimer Disease in Korean Population. Alzheimer Dis. Assoc. Disord.

Citron M, Oltersdorf T, Haass C, McConlogue L, Hung AY, Seubert P, Vigo-Pelfrey C,

Lieberburg I, Selkoe DJ. 1992. Mutation of the beta-amyloid precursor protein in familial Alzheimer's disease increases beta-protein production. Nature 360: 672–4.

Clark CM. 2011. Use of Florbetapir-PET for Imaging β-Amyloid Pathology. JAMA 305: 275.

Cochran JN, Rush T, Buckingham SC, Roberson ED. 2015. The Alzheimer's disease risk factor CD2AP maintains blood-brain barrier integrity. Hum. Mol. Genet.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38: 1767–71.

Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH. 2006. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. N. Engl. J. Med. 354: 1264–1272.

Consortium TEP. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9: e1001046.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. 2009. Mapping complex disease traits with global gene expression. Nat. Rev. Genet. 10: 184–94.

Coon KD, Myers AJ, Craig DW, Webster JA, Pearson J V, Lince DH, Zismann VL, Beach TG, Leung D, Bryden L, Halperin RF, Marlowe L, et al. 2007. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. J. Clin. Psychiatry 68: 613–8.

Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. 12: 628–40.

Cooper TA. 2005. Use of minigene systems to dissect alternative splicing elements. Methods 37: 331–40.

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261: 921–3.

Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal lari R, Lupien M, Markowitz S, Scacheri PC. 2014. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 24: 1–13.

Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS. 2006. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat. Methods 3: 503–9.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. U. S. A. 107: 21931–6.

Crocker PR, Paulson JC, Varki A. 2007. Siglecs and their roles in the immune system. Nat. Rev. Immunol. 7: 255–66.

Cruchaga C, Goate AM. 2015. Cruchaga & Goate reply. Nature 520: E5–6.

Cruchaga C, Haller G, Chakraverty S, Mayo K, Vallania FLM, Mitra RD, Faber K, Williamson J, Bird T, Diaz-Arrastia R, Foroud TM, Boeve BF, et al. 2012. Rare variants in APP, PSEN1 and PSEN2 increase risk for AD in late-onset Alzheimer's disease families. PLoS One 7: e31039.

Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, Harari O, Norton J, Budde J, Bertelsen S, Jeng AT, Cooper B, et al. 2013. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature 505: 550–554.

Cruts M, Broeckhoven C Van. 1998. Presenilin mutations in Alzheimer's disease.

Hum. Mutat. 11: 183–190.

Cuyvers E, Roeck A De, Bossche T Van den, Cauwenberghe C Van, Bettens K, Vermeulen S, Mattheijssens M, Peeters K, Engelborghs S, Vandenbulcke M, Vandenberghe R, Deyn PP De, et al. 2015. Mutations in ABCA7 in a Belgian cohort of Alzheimer's disease patients: a targeted resequencing study. Lancet Neurol. 14: 814–822.

D'Souza I, Schellenberg GD. 2006. Arginine/serine-rich protein interaction domain-dependent modulation of a tau exon 10 splicing enhancer: altered interactions and mechanisms for functionally antagonistic FTDP-17 mutations Delta280K AND N279K. J. Biol. Chem. 281: 2460–9.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. 2011. The variant call format and VCFtools. Bioinformatics 27: 2156–8.

Davies P, Maloney AJ. 1976. Selective loss of central cholinergic neurons in Alzheimer's disease. Lancet (London, England) 2: 1403.

Dayem Ullah AZ, Lemoine NR, Chelala C. 2012. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). Nucleic Acids Res. 40: W65–70.

Dayem Ullah AZ, Lemoine NR, Chelala C. 2013. A practical guide for the functional annotation of genetic variations using SNPnexus. Brief. Bioinform. 14: 437–47.

DeConti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip. Rev. RNA 4: 49–60.

Dekker J, Marti-Renom MA, Mirny LA. 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. 14: 390–403.

Del-Aguila JL, Koboldt DC, Black K, Chasse R, Norton J, Wilson RK, Cruchaga C. 2015. Alzheimer's disease: rare variants with large effect sizes. Curr. Opin. Genet. Dev. 33: 49–55.

Deng Y-L, Liu L-H, Wang Y, Tang H-D, Ren R-J, Xu W, Ma J-F, Wang L-L, Zhuang J-P, Wang G, Chen S-D. 2012. The prevalence of CD33 and MS4A6A variant in Chinese Han population with Alzheimer's disease. Hum. Genet. 131: 1245–9.

Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 37: e67.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. PLoS Biol. 8: 12.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36: e105.

Donahue CP, Muratore C, Wu JY, Kosik KS, Wolfe MS. 2006. Stabilization of the tau exon 10 stem loop alters pre-mRNA splicing. J. Biol. Chem. 281: 23302–6.

Dong Y, Wang J, Sheng Z, Li G, Ma H, Wang X, Zhang R, Lu G, Hu Q, Sugimura H, Zhou X. 2009. Downregulation of EphA1 in colorectal carcinomas correlates with invasion and metastasis. Mod. Pathol. 22: 151–60.

Dredge BK, Polydorides AD, Darnell RB. 2001. The splice of life: alternative splicing and neurological disease. Nat. Rev. Neurosci. 2: 43–50.

Du H, Rosbash M. 2002. The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing. Nature 419: 86–90.

Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P,

Delacourte A, Frisoni G, Fox NC, Galasko D, Gauthier S, Hampel H, et al. 2010. Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol. 9: 1118–27.

Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'brien J, et al. 2007. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol. 6: 734–46.

Dubois B, Feldman HH, Jacova C, Hampel H, Molinuevo JL, Blennow K, DeKosky ST, Gauthier S, Selkoe D, Bateman R, Cappa S, Crutch S, et al. 2014. Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. Lancet Neurol. 13: 614–29.

Dustin ML, Olszowy MW, Holdorf AD, Li J, Bromley S, Desai N, Widder P, Rosenberger F, Merwe PA van der, Allen PM, Shaw AS. 1998. A novel adaptor protein orchestrates receptor patterning and cytoskeletal polarity in T-cell contacts. Cell 94: 667–77.

Eckersley-Maslin MA, Spector DL. 2014. Random monoallelic expression: regulating gene expression one allele at a time. Trends Genet.

Eddy SR. 2004. What is Bayesian statistics? Nat. Biotechnol. 22: 1177–1178.

Eikelenboom P, Veerhuis R, Scheper W, Rozemuller AJM, Gool WA van, Hoozemans JJM. 2006. The significance of neuroinflammation in understanding Alzheimer's disease. J. Neural Transm. 113: 1685–95.

Eph Nomenclature Committee. 1997. Unified nomenclature for Eph family receptors and their ligands, the ephrins. Eph Nomenclature Committee. Cell 90: 403–4.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43–9.

Ertekin-Taner N. 2007. Genetics of Alzheimer's disease: a centennial review. Neurol. Clin. 25: 611–67, v.

Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, Finch CE, St Laurent G, Kenny PJ, Wahlestedt C. 2008. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat. Med. 14: 723–30.

Fairbrother WG, Yeh R-F, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. Science 297: 1007–13.

Faye LL, Machiela MJ, Kraft P, Bull SB, Sun L. 2013. Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. PLoS Genet. 9: e1003609.

Femminella GD, Ferrara N, Rengo G. 2015. The emerging role of microRNAs in Alzheimer's disease. Front. Physiol. 6: 40.

Fleminger S, Oliver DL, 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: 857–62.

Fu W, O'Connor TD, Akey JM. 2013. Genetic architecture of quantitative traits and complex diseases. Curr. Opin. Genet. Dev. 23: 678–83.

Ge B, Gurd S, Gaudin T, Dore C, Lepage P, Harmsen E, Hudson TJ, Pastinen T. 2005. Survey of allelic expression using EST mining. Genome Res. 15: 1584–91.

Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KCL, Gagné V, Dias J, Hoberman R, et al. 2009. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nat. Genet. 41: 1216–22.

Giacobini E, Gold G. 2013. Alzheimer disease therapy--moving from amyloid-β to tau. Nat. Rev. Neurol. 9: 677–86.

Gibson G. 2011. Rare and common variants: twenty arguments. Nat. Rev. Genet. 13: 135–45.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. 11: 759–69.

Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, Haynes A, Irving N, James L. 1991. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature 349: 704–6.

Goldstein LSB, Reyna S, Woodruff G. 2015. Probing the secrets of Alzheimer's disease using human-induced pluripotent stem cell technology. Neurotherapeutics 12: 121–5.

Gómez-Isla T, Price JL, McKeel DW, Morris JC, Growdon JH, Hyman BT. 1996. Profound loss of layer II entorhinal cortex neurons occurs in very mild Alzheimer's disease. J. Neurosci. 16: 4491–500.

Grabowski P. 2011. Alternative splicing takes shape during neuronal development. Curr. Opin. Genet. Dev. 21: 388–94.

Griciuc A, Serrano-Pozo A, Parrado AR, Lesinski AN, Asselin CN, Mullin K, Hooli B, Choi SH, Hyman BT, Tanzi RE. 2013. Alzheimer's disease risk gene CD33 inhibits microglial uptake of amyloid beta. Neuron 78: 631–643.

Griffin JD, Linch D, Sabbath K, Larcom P, Schlossman SF. 1984. A monoclonal antibody reactive with normal and leukemic human myeloid progenitor cells. Leuk. Res. 8: 521–34.

Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, Morgan A, Jehu L, Segurado R, Stone D, Schadt E, Karnoub M, Nowotny P, et al. 2007. Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum. Mol. Genet. 16: 865–73.

Guerreiro R, Brás J, Hardy J, Singleton A. 2014. Next generation sequencing techniques in neurological diseases: redefining clinical and molecular associations. Hum. Mol. Genet. 23: R47–53.

Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, Cruchaga C, Sassi C, Kauwe JSK, Younkin S, Hazrati L, Collinge J, et al. 2013. TREM2 variants in Alzheimer's disease. N. Engl. J. Med. 368: 117–27.

Guo L, Du Y, Qu S, Wang J. 2015. rVarBase: an updated database for regulatory features of human variants. Nucleic Acids Res.

Hafner C, Schmitz G, Meyer S, Bataille F, Hau P, Langmann T, Dietmaier W, Landthaler M, Vogt T. 2004. Differential gene expression of Eph receptors and ephrins in benign human tissues and cancers. Clin. Chem. 50: 490–9.

Halliday GM, Double KL, Macdonald V, Kril JJ. 2003. Identifying severely atrophic cortical subregions in Alzheimer's disease. Neurobiol. Aging 24: 797–806.

Hardy J. 2009. The amyloid hypothesis for Alzheimer's disease: a critical reappraisal. J. Neurochem. 110: 1129–34.

Hardy J, Allsop D. 1991. Amyloid deposition as the central event in the aetiology of Alzheimer's disease. Trends Pharmacol. Sci. 12: 383–8.

Hardy J, Selkoe DJ. 2002. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science 297: 353–6.

Hardy JA, Higgins GA. 1992. Alzheimer's disease: the amyloid cascade hypothesis. Science 256: 184–5.

Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, Pahwa JS,

Moskvina V, Dowzell K, Williams A, Jones N, Thomas C, et al. 2009. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat. Genet. 41: 1088–93.

Harris FM, Brecht WJ, Xu Q, Tesseur I, Kekonius L, Wyss-Coray T, Fish JD, Masliah E, Hopkins PC, Scearce-Levie K, Weisgraber KH, Mucke L, et al. 2003. Carboxyl-terminal-truncated apolipoprotein E4 causes Alzheimer's disease-like neurodegeneration and behavioral deficits in transgenic mice. Proc. Natl. Acad. Sci. U. S. A. 100: 10966–71.

Harris R. 1982. Genetics of Alzheimer's disease. Br. Med. J. (Clin. Res. Ed). 284: 1065–6.

Hashimoto T, Serrano-Pozo A, Hori Y, Adams KW, Takeda S, Banerji AO, Mitani A, Joyner D, Thyssen DH, Bacskai BJ, Frosch MP, Spires-Jones TL, et al. 2012. Apolipoprotein E, Especially Apolipoprotein E4, Increases the Oligomerization of Amyloid Peptide. J. Neurosci. 32: 15181–15192.

Hawkes CA, Sullivan PM, Hands S, Weller RO, Nicoll JAR, Carare RO. 2012. Disruption of arterial perivascular drainage of amyloid-β from the brains of mice expressing the human APOE ε4 allele. PLoS One 7: e41636.

He F-F, Zhang C, Chen S, Deng B-Q, Wang H, Shao N, Tian X-J, Fang Z, Sun X-F, Liu J-S, Zhu Z-H, Meng X-F. 2011. Role of CD2-associated protein in albumin overload-induced apoptosis in podocytes. Cell Biol. Int. 35: 827–34.

Heilmann S, Drichel D, Clarimon J, Fernández V, Lacour A, Wagner H, Thelen M, Hernández I, Fortea J, Alegret M, Blesa R, Mauleón A, et al. 2015. PLD3 in non-familial Alzheimer's disease. Nature 520: E3–E5.

Heneka MT, Carson MJ, Khoury J El, Landreth GE, Brosseron F, Feinstein DL, Jacobs AH, Wyss-Coray T, Vitorica J, Ransohoff RM, Herrup K, Frautschy SA, et al. 2015. Neuroinflammation in Alzheimer's disease. Lancet Neurol. 14: 388–405.

Herath NI, Doecke J, Spanevello MD, Leggett BA, Boyd AW. 2009. Epigenetic silencing of EphA1 expression in colorectal cancer is correlated with poor survival. Br. J. Cancer 100: 1095–102.

Herholz K, Ebmeier K. 2011. Clinical amyloid imaging in Alzheimer's disease. Lancet. Neurol. 10: 667–70.

Herrup K. 2015. The case for rejecting the amyloid cascade hypothesis. Nat. Neurosci. 18: 794–799.

Hirai H, Maru Y, Hagiwara K, Nishida J, Takaku F. 1987. A novel putative tyrosine kinase receptor encoded by the eph gene. Science 238: 1717–20.

Hirata H, Tatsumi H, Sokabe M. 2008. Zyxin emerges as a key player in the mechanotransduction at cell adhesive structures. Commun. Integr. Biol. 1: 192–5.

Hollingworth P, Harold D, Sims R, Gerrish A, Lambert J-C, Carrasquillo MM, Abraham R, Hamshere ML, Pahwa JS, Moskvina V, Dowzell K, Jones N, et al. 2011. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat. Genet. 43: 429–35.

Holtzman DM, Herz J, Bu G. 2012. Apolipoprotein E and apolipoprotein E receptors: normal biology and roles in Alzheimer disease. Cold Spring Harb. Perspect. Med. 2: a006312.

Holtzman DM, Morris JC, Goate AM. 2011. Alzheimer's disease: the challenge of the second century. Sci. Transl. Med. 3: 77sr1.

Homer N, Merriman B, Nelson SF. 2009. BFAST: an alignment tool for large scale genome resequencing. PLoS One 4: e7767.

Hooli B V, Lill CM, Mullin K, Qiao D, Lange C, Bertram L, Tanzi RE. 2015. PLD3 gene variants and Alzheimer's disease. Nature 520: E7–8.

Hossini AM, Megges M, Prigione A, Lichtner B, Toliat MR, Wruck W, Schröter F, Nuernberg P, Kroll H, Makrantonaki E, Zouboulis CC, Zouboulis CC, et al. 2015. Induced pluripotent stem cell-derived neuronal cells from a sporadic Alzheimer's disease donor as a model for investigating AD-associated gene regulatory networks. BMC Genomics 16: 84.

Houldcroft CJ, Petrova V, Liu JZ, Frampton D, Anderson CA, Gall A, Kellam P. 2014. Host genetic variants and gene expression patterns associated with Epstein-Barr virus copy number in lymphoblastoid cell lines. PLoS One 9: e108384.

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44: 955–9.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5: e1000529.

Hruska M, Dalva MB. 2012. Ephrin regulation of synapse formation, function and plasticity. Mol. Cell. Neurosci. 50: 35–44.

Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, Danecek P, Malerba G, Trabetti E, Zheng H-F, Turki S Al, Amuzu A, et al. 2015. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat. Commun. 6: 8111.

Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P. 2009. Genotype-imputation accuracy across worldwide human populations. Am. J. Hum. Genet. 84: 235–50.

Hutton M, Lendon CL, Rizzu P, Baker M, Froelich S, Houlden H, Pickering-Brown S, Chakraverty S, Isaacs A, Grover A, Hackett J, Adamson J, et al. 1998. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. Nature 393: 702–5.

Hyman BT, Phelps CH, Beach TG, Bigio EH, Cairns NJ, Carrillo MC, Dickson DW, Duyckaerts C, Frosch MP, Masliah E, Mirra SS, Nelson PT, et al. 2012. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. Alzheimers. Dement. 8: 1–13.

Imbeaud S, Graudens E, Boulanger V, Barlet X, Zaborski P, Eveno E, Mueller O, Schroeder A, Auffray C. 2005. Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. Nucleic Acids Res. 33: e56.

International HapMap Consortium. 2003. The International HapMap Project. Nature 426: 789–96.

International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437: 1299–320.

Israel MA, Yuan SH, Bardy C, Reyna SM, Mu Y, Herrera C, Hefferan MP, Gorp S Van, Nazor KL, Boscolo FS, Carson CT, Laurent LC, et al. 2012. Probing sporadic and familial Alzheimer's disease using induced pluripotent stem cells. Nature 482: 216–20.

Jack CR, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, Thies B, Phelps CH. 2011. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers. Dement. 7: 257–62.

Jäger R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, Heindl A, Whiffin N, Carnicer MJ, Broome L, Dryden N, Nagano T, Schoenfelder S, et al. 2015. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nat. Commun. 6: 6178.

Jarrett JT, Berger EP, Lansbury PT. 1993. The carboxy terminus of the beta amyloid protein is critical for the seeding of amyloid formation: Implications for the pathogenesis of Alzheimer's disease. Biochemistry 32: 4693–4697.

Jash A, Yun K, Sahoo A, So J-S, Im S-H. 2012. Looping mediated interaction between the promoter and 3' UTR regulates type II collagen expression in chondrocytes. PLoS One 7: e40828.

Jellinger KA. 2006. Clinicopathological analysis of dementia disorders in the elderly - An update. J. Alzheimer's Dis. 9: 61–70.

Jian X, Boerwinkle E, Liu X. 2014. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. Genet. Med. 16: 497–503.

Jiang Q, Lee CYD, Mandrekar S, Wilkinson B, Cramer P, Zelcer N, Mann K, Lamb B, Willson TM, Collins JL, Richardson JC, Smith JD, et al. 2008. ApoE promotes the proteolytic degradation of Abeta. Neuron 58: 681–93.

Jin SC, Pastor P, Cooper B, Cervantes S, Benitez BA, Razquin C, Goate A, Cruchaga C. 2012. Pooled-DNA sequencing identifies novel causative variants in PSEN1, GRN and MAPT in a clinical early-onset and familial Alzheimer's disease Ibero-American cohort. Alzheimers. Res. Ther. 4: 34.

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, Bakker PIW de. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24: 2938–9.

Johnson EO, Hancock DB, Levy JL, Gaddis NC, Saccone NL, Bierut LJ, Page GP. 2013. Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. Hum. Genet. 132: 509–22.

Johnson JW, Kotermanski SE. 2006. Mechanism of action of memantine. Curr. Opin. Pharmacol. 6: 61–7.

Johnson MB, Kawasawa YI, Mason CE, Krsnik Z, Coppola G, Bogdanović D, Geschwind DH, Mane SM, State MW, Sestan N. 2009. Functional and evolutionary insights into human brain development through global transcriptome analysis. Neuron 62: 494–509.

Jonghe C De, Rogaeva EA, Tysoe C, Singleton A, Vanderstichele H, Meschino W, Dermaut B, Vanderhoeven I, Backhovens H, Vanmechelen E, Morris CM, Hardy J, et al. 1999. Aberrant Splicing in the Presenilin-1 Intron 4 Mutation Causes Presenile Alzheimer's Disease by Increased A 42 Secretion. Hum. Mol. Genet. 8: 1529–1540.

Jonsson T, Stefansson H, Steinberg S, Jonsdottir I, Jonsson P V., Snaedal J, Bjornsson S, Huttenlocher J, Levey AI, Lah JJ, Rujescu D, Hampel H, et al. 2013. Variant of TREM2 Associated with the Risk of Alzheimer's Disease. N. Engl. J. Med. 368: 107–116.

Jun G, Naj AC, Beecham GW, Wang L-S, Buros J, Gallins PJ, Buxbaum JD, Ertekin-Taner N, Fallin MD, Friedland R, Inzelberg R, Kramer P, et al. 2010. Meta-analysis confirms CR1, CLU, and PICALM as alzheimer disease risk loci and reveals interactions with APOE genotypes. Arch. Neurol. 67: 1473–84.

Kalaria RN, Maestre GE, Arizaga R, Friedland RP, Galasko D, Hall K, Luchsinger JA, Ogunniyi A, Perry EK, Potocnik F, Prince M, Stewart R, et al. 2008. Alzheimer's disease and vascular dementia in developing countries: prevalence, management, and risk factors. Lancet. Neurol. 7: 812–26.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AMM, Pletikos M, Meyer KA, Sedmak G, Guennel T, Shin Y, et al. 2011. Spatio-temporal transcriptome of the human brain. Nature 478: 483–9.

Kang J, Lemaire HG, Unterbeck A, Salbaum JM, Masters CL, Grzeschik KH, Multhaup G, Beyreuther K, Müller-Hill B. The precursor of Alzheimer's disease amyloid A4 protein resembles a cell-surface receptor. Nature 325: 733–6.

Karch CM, Jeng AT, Nowotny P, Cady J, Cruchaga C, Goate AM. 2012. Expression of novel Alzheimer's disease risk genes in control and Alzheimer's disease brains. PLoS One 7: e50976.

Kilpinen H, Barrett JC. 2013. How next-generation sequencing is transforming complex disease genetics. Trends Genet. 29: 23–30.

Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35: 125–31.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46: 310–5.

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. 2013. The next-generation sequencing revolution and its impact on genomics. Cell 155: 27–38.

Koffie RM, Hashimoto T, Tai H-C, Kay KR, Serrano-Pozo A, Joyner D, Hou S, Kopeikina KJ, Frosch MP, Lee VM, Holtzman DM, Hyman BT, et al. 2012. Apolipoprotein E4 effects in Alzheimer's disease are mediated by synaptotoxic oligomeric amyloid- . Brain 135: 2155–2168.

Kok E, Haikonen S, Luoto T, Huhtala H, Goebeler S, Haapasalo H, Karhunen PJ. 2009. Apolipoprotein E-dependent accumulation of Alzheimer disease-related lesions begins in middle age. Ann. Neurol. 65: 650–657.

Kondo S, Yamamoto N, Murakami T, Okumura M, Mayeda A, Imaizumi K. 2004. Tra2β, SF2/ASF and SRp30c modulate the function of an exonic splicing enhancer in exon 10 of tau pre-mRNA. Genes to Cells 9: 121–130.

Koudinov AR, Koudinova N V. 1997. Alzheimer's soluble amyloid beta protein is secreted by HepG2 cells as an apolipoprotein. Cell Biol. Int. 21: 265–71.

Kullander K, Klein R. 2002. Mechanisms and functions of Eph and ephrin signalling. Nat. Rev. Mol. Cell Biol. 3: 475–86.

LaDu MJ, Gilligan SM, Lukens JR, Cabana VG, Reardon CA, Eldik LJ Van, Holtzman DM. 1998. Nascent astrocyte particles differ from lipoproteins in CSF. J. Neurochem. 70: 2070–2081.

Lamba JK, Adachi M, Sun D, Tammur J, Schuetz EG, Allikmets R, Schuetz JD. 2003. Nonsense mediated decay downregulates conserved alternatively spliced ABCC4 transcripts bearing nonsense codons. Hum. Mol. Genet. 12: 99–109.

Lambert J-C, Grenier-Boley B, Bellenguez C, Pasquier F, Campion D, Dartigues J-F, Berr C, Tzourio C, Amouyel P. 2015. PLD3 and sporadic Alzheimer's disease risk. Nature 520: E1.

Lambert J-C, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, Combarros O, Zelenika D, Bullido MJ, Tavernier B, Letenneur L, Bettens K, et al. 2009. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat. Genet. 41: 1094–9.

Lambert J-C, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, Jun G, DeStefano AL, Bis JC, Beecham GW, Grenier-Boley B, Russo G, et al. 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. 45: 1452–1458.

Lander ES. 1996. The New Genomics: Global Views of Biology. Science (80-.). 274: 536–539.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, et al.

2013. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506–11.

Lassmann T, Hayashizaki Y, Daub CO. 2011. SAMStat: monitoring biases in next generation sequencing data. Bioinformatics 27: 130–1.

Le SQ, Durbin R. 2011. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res. 21: 952–60.

Lee CJ, Irizarry K. 2003. Alternative splicing in the nervous system: an emerging source of diversity and regulation. Biol. Psychiatry 54: 771–776.

Lee CYD, Tse W, Smith JD, Landreth GE. 2012. Apolipoprotein E promotes β-amyloid trafficking and degradation by modulating microglial cholesterol levels. J. Biol. Chem. 287: 2032–44.

Lee SJ van der, Holstege H, Wong TH, Jakobsdottir J, Bis JC, Chouraki V, Rooij JGJ van, Grove ML, Smith A V, Amin N, Choi S-H, Beiser AS, et al. 2015. PLD3 variants in population studies. Nature 520: E2–3.

Lehtonen S, Tienari J, Londesborough A, Pirvola U, Ora A, Reima I, Lehtonen E. 2008. CD2-associated protein is widely expressed and differentially regulated during embryonic development. Differentiation. 76: 506–17.

Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15–20.

Li H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27: 718–9.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–9.

Li H, Homer N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. 11: 473–83.

Li H, Wetten S, Li L, St Jean PL, Upmanyu R, Surh L, Hosford D, Barnes MR, Briley JD, Borrie M, Coletta N, Delisle R, et al. 2008. Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. Arch. Neurol. 65: 45–53.

Li J, Kanekiyo T, Shinohara M, Zhang Y, LaDu MJ, Xu H, Bu G. 2012. Differential regulation of amyloid-β endocytic trafficking and lysosomal degradation by apolipoprotein E isoforms. J. Biol. Chem. 287: 44593–601.

Li MJ, Wang LY, Xia Z, Sham PC, Wang J. 2013. GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. Nucleic Acids Res. 41: W150–8.

Li Q, Lee J-A, Black DL. 2007. Neuronal regulation of alternative pre-mRNA splicing. Nat. Rev. Neurosci. 8: 819–31.

Licatalosi DD, Darnell RB. 2006. Splicing regulation in neurologic disease. Neuron 52: 93–101.

Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc. Natl. Acad. Sci. U. S. A. 108: 11093–8.

Lin M, Hrabovsky A, Pedrosa E, Wang T, Zheng D, Lachman HM. 2012. Allele-biased expression in differentiating human neurons: implications for neuropsychiatric disorders. PLoS One 7: e44017.

Linenberger ML. 2005. CD33-directed therapy with gemtuzumab ozogamicin in acute myeloid leukemia: progress in understanding cytotoxicity and potential mechanisms of

drug resistance. Leukemia 19: 176–82.

Liu Y, Julkunen V, Paajanen T, Westman E, Wahlund L-O, Aitken A, Sobow T, Mecocci P, Tsolaki M, Vellas B, Muehlboeck S, Spenger C, et al. 2012. Education increases reserve against Alzheimer's disease--evidence from structural MRI analysis. Neuroradiology 54: 929–38.

Liu-Seifert H, Siemers E, Holdridge KC, Andersen SW, Lipkovich I, Carlson C, Sethuraman G, Hoog S, Hayduk R, Doody R, Aisen P. 2015. Delayed-start analysis: Mild Alzheimer's disease patients in solanezumab trials, 3.5 years. Alzheimer's Dement. Transl. Res. Clin. Interv. 1: 111–121.

Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, Chittenden TW, D'Gama AM, Cai X, Luquette LJ, Lee E, Park PJ, et al. 2015. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science (80-.). 350: 94–98.

Logue MW, Schu M, Vardarajan BN, Buros J, Green RC, Go RCP, Griffith P, Obisesan TO, Shatz R, Borenstein A, Cupples LA, Lunetta KL, et al. 2011. A comprehensive genetic association study of Alzheimer disease in African Americans. Arch. Neurol. 68: 1569–79.

Logue MW, Schu M, Vardarajan BN, Farrell J, Bennett DA, Buxbaum JD, Byrd GS, Ertekin-Taner N, Evans D, Foroud T, Goate A, Graff-Radford NR, et al. 2014. Two rare AKAP9 variants are associated with Alzheimer's disease in African Americans. Alzheimers. Dement. 10: 609–618.e11.

Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. Biochem. J. 417: 15.

Lord J, Turton J, Medway C, Shi H, Brown K, Lowe J, Mann D, Pickering-Brown S, Kalsheker N, Passmore P, Morgan K. 2012. Next generation sequencing of CLU, PICALM and CR1: pitfalls and potential solutions. Int. J. Mol. Epidemiol. Genet. 3: 262–75.

Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, Sawyer SL. 2013. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc. Natl. Acad. Sci. U. S. A. 110: 19872–7.

Lowe WL, Reddy TE. 2015. Genomic approaches for understanding the genetics of complex disease. Genome Res. 25: 1432–1441.

Lu C, Ren W, Su X-M, Chen J-Q, Wu S-H, Guo X-R, Huang S-M, Chen L-H, Zhou G-P. 2008. CREB and Sp1 regulate the human CD2AP gene promoter activity in renal tubular epithelial cells. Arch. Biochem. Biophys. 474: 143–9.

Ma Y, Yang H, Qi J, Liu D, Xiong P, Xu Y, Feng W, Zheng G, Li P, Fang M, Tan Z, Zheng F, et al. 2010. CD2AP is indispensable to multistep cytotoxic process by NK cells. Mol. Immunol. 47: 1074–82.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335: 823–8.

MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, Barrett JC, Biesecker LG, et al. 2014. Guidelines for investigating causality of sequence variants in human disease. Nature 508: 469–76.

Makalowska I, Lin C-F, Makalowski W. 2005. Overlapping genes in vertebrate genomes. Comput. Biol. Chem. 29: 1–12.

Malik M, Chiles J, Xi HS, Medway C, Simpson J, Potluri S, Howard D, Liang Y, Paumi CM, Mukherjee S, Crane P, Younkin S, et al. 2015. Genetics of CD33 in Alzheimer's disease and acute myeloid leukemia. Hum. Mol. Genet. 24: 3557–70.

Malik M, Simpson JF, Parikh I, Wilfred BR, Fardo DW, Nelson PT, Estus S. 2013. CD33 Alzheimer's Risk-Altering Polymorphism, CD33 Expression, and Exon 2 Splicing. J. Neurosci. 33: 13320–5.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. Nat. Methods 7: 111–8.

Manley JL, Krainer AR. 2010. A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). Genes Dev. 24: 1073–4.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, et al. 2009. Finding the missing heritability of complex diseases. Nature 461: 747–53.

Mao J, Li C, Zhang Y, Li Y, Zhao Y. 2010. Human with-no-lysine kinase-4 3'-UTR acting as the enhancer and being targeted by miR-296. Int. J. Biochem. Cell Biol. 42: 1536–43.

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11: 499–511.

Mardis ER. 2011. A decade's perspective on DNA sequencing technology. Nature 470: 198–203.

Mardis ER. 2013. Next-generation sequencing platforms. Annu. Rev. Anal. Chem. (Palo Alto. Calif). 6: 287–303.

Matlin AJ, Clark F, Smith CWJ. 2005. Understanding alternative splicing: towards a cellular code. Nat. Rev. Mol. Cell Biol. 6: 386–98.

Matys V, Kel-Margoulis O V, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, et al. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34: D108–10.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. Science 337: 1190–5.

McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier J-B, Donnelly P. 2014. Choice of transcripts and software has a large effect on variant annotation. Genome Med. 6: 26.

McGlincy NJ, Valomon A, Chesham JE, Maywood ES, Hastings MH, Ule J. 2012. Regulation of alternative splicing by the circadian clock and food related cues. Genome Biol. 13: R54.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–303.

McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. 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: 939–939.

McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, et al. 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. Dement. 7: 263–9.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor.

Bioinformatics 26: 2069–70.

McNeill R, Sare GM, Manoharan M, Testa HJ, Mann DMA, Neary D, Snowden JS, Varma AR. 2007. Accuracy of single-photon emission computed tomography in differentiating frontotemporal dementia from Alzheimer's disease. J. Neurol. Neurosurg. Psychiatry 78: 350–5.

Medway C, Morgan K. 2013. Sialic Acid Binding Immunoglobulin-Like Lectin-3. In: Morgan K and Carrasquillo MM, editors. Genetic Variants in Alzheimer's Disease,.

Medway C, Morgan K. 2014. Review: The genetics of Alzheimer's disease; putting flesh on the bones. Neuropathol. Appl. Neurobiol. 40: 97–105.

Meng X, D'Arcy C. 2012. Education and dementia in the context of the cognitive reserve hypothesis: a systematic review with meta-analyses and qualitative analyses. PLoS One 7: e38268.

Mertes F, Elsharawy A, Sauer S, Helvoort JMLM van, Zaag PJ van der, Franke A, Nilsson M, Lehrach H, Brookes AJ. 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief. Funct. Genomics 10: 374–86.

Metzker ML. 2010. Sequencing technologies - the next generation. Nat. Rev. Genet. 11: 31–46.

Miao H, Wang B. 2012. EphA receptor signaling--complexity and emerging themes. Semin. Cell Dev. Biol. 23: 16–25.

Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, Herman B, Happe S, et al. 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. 47: 598–606.

Mills JD, Janitz M. 2012. Alternative splicing of mRNA in the molecular pathology of neurodegenerative diseases. Neurobiol. Aging 33: 1012.e11–1012.e24.

Mills JD, Nalpathamkalam T, Jacobs HIL, Janitz C, Merico D, Hu P, Janitz M. 2013. RNA-Seq analysis of the parietal cortex in Alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism. Neurosci. Lett. 536: 90–95.

Miner JH, Morello R, Andrews KL, Li C, Antignac C, Shaw AS, Lee B. 2002. Transcriptional induction of slit diaphragm genes by Lmx1b is required in podocyte differentiation. J. Clin. Invest. 109: 1065–72.

Mitchell AJ. 2009. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. J. Psychiatr. Res. 43: 411–31.

Montagne A, Barnes SR, Sweeney MD, Halliday MR, Sagare AP, Zhao Z, Toga AW, Jacobs RE, Liu CY, Amezcua L, Harrington MG, Chui HC, et al. 2015. Blood-Brain Barrier Breakdown in the Aging Human Hippocampus. Neuron 85: 296–302.

Morgan K. 2011. The three new pathways leading to Alzheimer's disease. Neuropathol. Appl. Neurobiol. 37: 353–7.

Morgulis A, Gertz EM, Schäffer AA, Agarwala R. 2006. WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22: 134–41.

Morris JC, Blennow K, Froelich L, Nordberg A, Soininen H, Waldemar G, Wahlund L-O, Dubois B. 2014. Harmonized diagnostic criteria for Alzheimer's disease: recommendations. J. Intern. Med. 275: 204–13.

Mu XJ, Lu ZJ, Kong Y, Lam HYK, Gerstein MB. 2011. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res. 39: 7058–76.

Mueller O, Lightfoot S, Schroeder A. 2004. RNA integrity number (RIN)–standardization of RNA quality control. Agil. Technol. Appl. Note, Tech. Rep. 5989–

1165E.

Mutter GL, Zahrieh D, Liu C, Neuberg D, Finkelstein D, Baker HE, Warrington JA. 2004. Comparison of frozen and RNALater solid tissue storage methods for use in RNA expression microarrays. BMC Genomics 5: 88.

Naj AC, Jun G, Beecham GW, Wang L-S, Vardarajan BN, Buros J, Gallins PJ, Buxbaum JD, Jarvik GP, Crane PK, Larson EB, Bird TD, et al. 2011. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat. Genet. 43: 436–41.

Neuropathology Group. Medical Research Council Cognitive Function and Aging Study. 2001. Pathological correlates of late-onset dementia in a multicentre, community-based population in England and Wales. Lancet 357: 169–175.

Nguyen-Dumont T, Jordheim LP, Michelon J, Forey N, McKay-Chopin S, Sinilnikova O, Calvez-Kelm F Le, Southey MC, Tavtigian S V, Lesueur F. 2011. Detecting differential allelic expression using high-resolution melting curve analysis: application to the breast cancer susceptibility gene CHEK2. BMC Med. Genomics 4: 39.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature 463: 457–63.

Nothnagel M, Wolf A, Herrmann A, Szafranski K, Vater I, Brosch M, Huse K, Siebert R, Platzer M, Hampe J, Krawczak M. 2011. Statistical inference of allelic imbalance from transcriptome data. Hum. Mutat. 32: 98–106.

O'Brien RJ, Wong PC. 2011. Amyloid precursor protein processing and Alzheimer's disease. Annu. Rev. Neurosci. 34: 185–204.

Opitz L, Salinas-Riester G, Grade M, Jung K, Jo P, Emons G, Ghadimi BM, Beissbarth T, Gaedcke J. 2010. Impact of RNA degradation on gene expression profiling. BMC Med. Genomics 3: 36.

Ovcharenko I, Nobrega MA, Loots GG, Stubbs L. 2004. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. Nucleic Acids Res. 32: W280–6.

Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B, Kwiatkowski DP, Swerdlow HP, Quail MA. 2012. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. BMC Genomics 13: 1.

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. Brief. Bioinform. 15: 256–78.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 40: 1413–5.

Panoutsopoulou K, Tachmazidou I, Zeggini E. 2013. In search of low-frequency and rare variants affecting complex traits. Hum. Mol. Genet. 22: R16–21.

Pei Y-F, Li J, Zhang L, Papasian CJ, Deng H-W. 2008. Analyses and comparison of accuracy of different genotype imputation methods. PLoS One 3: e3551.

Peiper SC, Ashmun RA, Look AT. 1988. Molecular cloning, expression, and chromosomal localization of a human gene encoding the CD33 myeloid differentiation antigen. Blood 72: 314–21.

Pericak-Vance MA, Bebout JL, Gaskell PC, Yamaoka LH, Hung WY, Alberts MJ, Walker AP, Bartlett RJ, Haynes CA, Welsh KA. 1991. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. Am. J. Hum. Genet. 48: 1034–50.

Perl DP. 2010. Neuropathology of Alzheimer's disease. Mt. Sinai J. Med. 77: 32–42.

Pierre A Saint, Génin E. 2014. How important are rare variants in common disease? Brief. Funct. Genomics 13: 353–61.

Poduri A, Evrony GD, Cai X, Walsh CA. 2013. Somatic Mutation, Genomic Variation, and Neurological Disease. Science (80-.). 341: 1237758–1237758.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20: 110–21.

Pottier C, Hannequin D, Coutant S, Rovelet-Lecrux A, Wallon D, Rousseau S, Legallic S, Paquet C, Bombois S, Pariente J, Thomas-Anterion C, Michon A, et al. 2012. High frequency of potentially pathogenic SORL1 mutations in autosomal dominant early-onset Alzheimer disease. Mol. Psychiatry 17: 875–9.

Prince M, Knapp M, M G, McCrone P, Prina M, Comas-Herrera A, Wittenberg R, Adelaja B, Hu B, King D, Rehill A, Salimkumar D. 2014. Dementia UK: Update.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, Bakker PIW de, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–75.

Qian L, Vu MN, Carter MS, Doskow J, Wilkinson MF. 1993. T cell receptor-beta mRNA splicing during thymic maturation in vivo and in an inducible T cell clone in vitro. J. Immunol. 151: 6801–14.

Qureshi IA, Mehler MF. 2012. Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. Nat. Rev. Neurosci. 13: 528–41.

Raj T, Ryan KJ, Replogle JM, Chibnik LB, Rosenkrantz L, Tang A, Rothamel K, Stranger BE, Bennett DA, Evans DA, Jager PL De, Bradshaw EM. 2014. CD33: increased inclusion of exon 2 implicates the Ig V-set domain in Alzheimer's disease susceptibility. Hum. Mol. Genet. 23: 2729–36.

Ramsköld D, Wang ET, Burge CB, Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. 5: e1000598.

Ratnapriya R, Zhan X, Fariss RN, Branham KE, Zipprer D, Chakarova CF, Sergeev Y V, Campos MM, Othman M, Friedman JS, Maminishkis A, Waseem NH, et al. 2014. Rare and common variants in extracellular matrix gene Fibrillin 2 (FBN2) are associated with macular degeneration. Hum. Mol. Genet. ddu276–.

Reese MG, Eeckman FH, Kulp D, Haussler D. 1997. Improved Splice Site Detection in Genie. J. Comput. Biol. 4: 311–323.

Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. Trends Genet. 17: 502–510.

Reiman EM, Webster JA, Myers AJ, Hardy J, Dunckley T, Zismann VL, Joshipura KD, Pearson J V, Hu-Lince D, Huentelman MJ, Craig DW, Coon KD, et al. 2007. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. Neuron 54: 713–20.

Reitz C, Mayeux R. 2014. Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers. Biochem. Pharmacol. 88: 640–51.

Resta C Di, Manzoni M, Berisso MZ, Siciliano G, Benedetti S, Ferrari M. 2014. Evaluation of damaging effects of splicing mutations: validation of an in vitro method for diagnostic laboratories. Clin. Chim. Acta. 436: 276–82.

Ribeiro-Silva A, Zhang H, Jeffrey SS. 2007. RNA extraction from ten year old formalin-fixed paraffin-embedded breast cancer samples: a comparison of column purification and magnetic bead-based technologies. BMC Mol. Biol. 8: 118.

Ritchie GRS, Dunham I, Zeggini E, Flicek P. 2014. Functional annotation of noncoding sequence variants. Nat. Methods 11: 294–6.

Rivas M a, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, Fennell T, Kirby A, et al. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. 43: 1066–73.

Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, Maller JB, Kukurba KR, DeLuca DS, Fromer M, Ferreira PG, Smith KS, et al. 2015. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science (80-.). 348: 666–669.

Robertson MJ, Soiffer RJ, Freedman AS, Rabinowe SL, Anderson KC, Ervin TJ, Murray C, Dear K, Griffin JD, Nadler LM. 1992. Human bone marrow depleted of CD33-positive cells mediates delayed but durable reconstitution of hematopoiesis: clinical trial of MY9 monoclonal antibody-purged autografts for the treatment of acute myeloid leukemia. Blood 79: 2229–36.

Rogaeva E, Meng Y, Lee JH, Gu Y, Kawarai T, Zou F, Katayama T, Baldwin CT, Cheng R, Hasegawa H, Chen F, Shibata N, et al. 2007. The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. Nat. Genet. 39: 168–77.

Ross CA, Akimov SS. 2014. Human-induced pluripotent stem cells: potential for neurodegenerative diseases. Hum. Mol. Genet. 23: R17–26.

Rovelet-Lecrux A, Hannequin D, Raux G, Meur N Le, Laquerrière A, Vital A, Dumanchin C, Feuillette S, Brice A, Vercelletto M, Dubas F, Frebourg T, et al. 2006. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat. Genet. 38: 24–6.

Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J. 2013a. The RNAsnp web server: predicting SNP effects on local RNA secondary structure. Nucleic Acids Res. 41: W475–W479.

Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J. 2013b. RNAsnp: Efficient Detection of Local RNA Secondary Structure Changes Induced by SNPs. Hum. Mutat. 34: 546–56.

Sakabe NJ, Savic D, Nobrega MA. 2012. Transcriptional enhancers in development and disease. Genome Biol. 13: 238.

Sala Frigerio C, Lau P, Troakes C, Deramecourt V, Gele P, Loo P Van, Voet T, Strooper B De. 2015. On the identification of low allele frequency mosaic mutations in the brains of Alzheimer's disease patients. Alzheimers. Dement.

Salminen A, Kaarniranta K. 2009. Siglec receptors and hiding plaques in Alzheimer's disease. J. Mol. Med. (Berl). 87: 697–701.

Sambrook J, W Russell D. 2001. Molecular Cloning: A Laboratory Manual. Cold Spring Harb. Lab. Press. Cold Spring Harb. NY 999.

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. Nature 489: 109–13.

Sato N, Hori O, Yamaguchi A, Lambert J-C, Chartier-Harlin M-C, Robinson PA, Delacourte A, Schmidt AM, Furuyama T, Imaizumi K, Tohyama M, Takagi T. 2002. A Novel Presenilin-2 Splice Variant in Human Alzheimer's Disease Brain Tissue. J. Neurochem. 72: 2498–2505.

Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. 2012. Linking disease associations with regulatory information in the human genome. Genome Res. 22: 1748–59.

Schellenberg GD, Montine TJ. 2012. The genetics and neuropathology of Alzheimer's disease. Acta Neuropathol. 124: 305–323.

Schlötterer C, Tobler R, Kofler R, Nolte V. 2014. Sequencing pools of individuals —

mining genome-wide polymorphism data without big funding. Nat. Rev. Genet. 15: 749–763.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328: 1036–40.

Schneider LS, Mangialasche F, Andreasen N, Feldman H, Giacobini E, Jones R, Mantua V, Mecocci P, Pani L, Winblad B, Kivipelto M. 2014. Clinical trials and late-stage drug development for Alzheimer's disease: an appraisal from 1984 to 2014. J. Intern. Med. 275: 251–83.

Schnetz-Boutaud NC, Hoffman J, Coe JE, Murdock DG, Pericak-Vance MA, Haines JL. 2012. Identification and confirmation of an exonic splicing enhancer variation in exon 5 of the Alzheimer disease associated PICALM gene. Ann. Hum. Genet. 76: 448–53.

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T. 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol. Biol. 7: 3.

Schulte EC, Kurz A, Alexopoulos P, Hampel H, Peters A, Gieger C, Rujescu D, Diehl-Schmid J, Winkelmann J. 2015. Excess of rare coding variants in PLD3 in late- but not early-onset Alzheimer's disease. Hum. Genome Var. 2: 14028.

Selkoe DJ. 1991. The molecular pathology of Alzheimer's disease. Neuron 6: 487–98.

Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, Fan J-B, Hudson TJ. 2008. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. PLoS Genet. 4: e1000006.

Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, Bis JC, Smith A V, Carassquillo MM, Lambert JC, Harold D, Schrijvers EMC, et al. 2010. Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA 303: 1832–40.

Sharma N, Sosnay PR, Ramalho AS, Douville C, Franca A, Gottschalk LB, Park J, Lee M, Vecchio-Pagan B, Raraigh KS, Amaral MD, Karchin R, et al. 2014. Experimental assessment of splicing variants using expression minigenes and comparison with in silico predictions. Hum. Mutat. 35: 1249–59.

Shen H, Li J, Zhang J, Xu C, Jiang Y, Wu Z, Zhao F, Liao L, Chen J, Lin Y, Tian Q, Papasian CJ, et al. 2013. Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. PLoS One 8: e59494.

Shendure J, Akey JM. 2015. The origins, determinants, and consequences of human mutations. Science (80-.). 349: 1478–1483.

Shih NY, Li J, Cotran R, Mundel P, Miner JH, Shaw AS. 2001. CD2AP localizes to the slit diaphragm and binds to nephrin via a novel C-terminal domain. Am. J. Pathol. 159: 2303–8.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15: 1034–50.

Simons YB, Turchin MC, Pritchard JK, Sella G. 2014. The deleterious mutation load is insensitive to recent population history. Nat. Genet. 46: 220–4.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15: 121–32.

Singh RK, Cooper TA. 2012. Pre-mRNA splicing in disease and therapeutics. Trends Mol. Med. 18: 472–82.

Sinnott JA, Kraft P. 2012. Artifact due to differential error when cases and controls are imputed from different platforms. Hum. Genet. 131: 111–9.

Smith RM, Webb A, Papp AC, Newman LC, Handelman SK, Suhy A, Mascarenhas R, Oberdick J, Sadee W. 2013. Whole transcriptome RNA-Seq allelic expression in human brain. BMC Genomics 14: 571.

So H-C, Gui AHS, Cherny SS, Sham PC. 2011. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet. Epidemiol. 35: 310–7.

Sorg C, Grothe MJ. 2015.

Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR, Kaye J, Montine TJ, Park DC, Reiman EM, et al. 2011. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers. Dement. 7: 280–92.

Steinberg S, Stefansson H, Jonsson T, Johannsdottir H, Ingason A, Helgason H, Sulem P, Magnusson OT, Gudjonsson SA, Unnsteinsdottir U, Kong A, Helisalmi S, et al. 2015. Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. Nat. Genet. 47: 445–447.

Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a common mechanism of human inherited disease. Genome Res. 21: 1563–71.

Strittmatter WJ, Weisgraber KH, Goedert M, Saunders AM, Huang D, Corder EH, Dong L-M, Jakes R, Alberts MJ, Gilbert JR, Han S-H, Hulette C, et al. 1994. Hypothesis: Microtubule Instability and Paired Helical Filament Formation in the Alzheimer Disease Brain Are Related to Apolipoprotein E Genotype. Exp. Neurol. 125: 163–171.

Strittmatter WJ, Weisgraber KH, Huang DY, Dong LM, Salvesen GS, Pericak-Vance M, Schmechel D, Saunders AM, Goldgaber D, Roses AD. 1993. Binding of human apolipoprotein E to synthetic amyloid beta peptide: isoform-specific effects and implications for late-onset Alzheimer disease. Proc. Natl. Acad. Sci. U. S. A. 90: 8098–102.

Sung YJ, Wang L, Rankinen T, Bouchard C, Rao DC. 2012. Performance of genotype imputations using data from the 1000 Genomes Project. Hum. Hered. 73: 18–25.

Tabet N. 2006. Acetylcholinesterase inhibitors for Alzheimer's disease: anti-inflammatories in acetylcholine clothing! Age Ageing 35: 336–8.

Take H, Watanabe S, Takeda K, Yu ZX, Iwata N, Kajigaya S. 2000. Cloning and characterization of a novel adaptor protein, CIN85, that interacts with c-Cbl. Biochem. Biophys. Res. Commun. 268: 321–8.

Tan L, Yu J-T, Zhang W, Wu Z-C, Zhang Q, Liu Q-Y, Wang W, Wang H-F, Ma X-Y, Cui W-Z. 2012. Association of GWAS-linked loci with late-onset Alzheimer's disease in a northern Han Chinese population. Alzheimers. Dement.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337: 64–9.

The UniProt Consortium. 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 40: D71–5.

Tossidou I, Teng B, Drobot L, Meyer-Schwesinger C, Worthmann K, Haller H, Schiffer M. 2010. CIN85/RukL is a novel binding partner of nephrin and podocin and mediates

slit diaphragm turnover in podocytes. J. Biol. Chem. 285: 25285–95.

Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, Raychaudhuri S. 2013. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. 45: 124–30.

Twine NA, Janitz K, Wilkins MR, Janitz M. 2011. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. PLoS One 6: e16266.

Tysoe C, Whittaker J, Xuereb J, Cairns NJ, Cruts M, Broeckhoven C Van, Wilcock G, Rubinsztein DC. 1998. A Presenilin-1 Truncating Mutation Is Present in Two Cases with Autopsy-Confirmed Early-Onset Alzheimer Disease. Am. J. Hum. Genet. 62: 70–76.

Uhlén M, Björling E, Agaton C, Szigyarto CA-K, Amini B, Andersen E, Andersson A-C, Angelidou P, Asplund A, Asplund C, Berglund L, Bergström K, et al. 2005. A human protein atlas for normal and cancer tissues based on antibody proteomics. Mol. Cell. Proteomics 4: 1920–32.

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, et al. 2015. Tissue-based map of the human proteome. Science (80-.). 347: 1260419–1260419.

Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Björling L, et al. 2010. Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol. 28: 1248–50.

Varani L, Hasegawa M, Spillantini MG, Smith MJ, Murrell JR, Ghetti B, Klug A, Goedert M, Varani G. 1999. Structure of tau exon 10 splicing regulatory element RNA and destabilization by mutations of frontotemporal dementia and parkinsonism linked to chromosome 17. Proc. Natl. Acad. Sci. U. S. A. 96: 8229–34.

Vardarajan BN, Ghani M, Kahn A, Sheikh S, Sato C, Barral S, Lee JH, Cheng R, Reitz C, Lantigua R, Reyes-Dumeyer D, Medrano M, et al. 2015. Rare coding mutations identified by sequencing of Alzheimer disease genome-wide association studies loci. Ann. Neurol. 78: 487–498.

Vidal DO, Souza JES de, Pires LC, Masotti C, Salim ACM, Costa MCF, Galante PAF, Souza SJ de, Camargo AA. 2011. Analysis of allelic differential expression in the human genome using allele-specific serial analysis of gene expression tags. Genome 54: 120–7.

Wakabayashi T, Strooper B De. 2008. Presenilins: members of the gamma-secretase quartets, but part-time soloists too. Physiology (Bethesda). 23: 194–204.

Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JRB, Xu C, Futema M, Lawson D, Iotchkova V, Schiffels S, et al. 2015. The UK10K project identifies rare variants in health and disease. Nature 526: 82–90.

Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY. 2011. Understanding the transcriptome through RNA structure. Nat. Rev. Genet. 12: 641–55.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280: 1077–82.

Wang K, Dickson SP, Stolle CA, Krantz ID, Goldstein DB, Hakonarson H. 2010a. Interpretation of association signals and identification of causal variants from genome-wide association studies. Am. J. Hum. Genet. 86: 730–42.

Wang K, Li M, Hakonarson H. 2010b. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38: e164.

Wang Y, Neumann H. 2010. Alleviation of neurotoxicity by microglial human Siglec-

11. J. Neurosci. 30: 3482–8.

Ward LD, Kellis M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. 30: 1095–106.

Wetzel-Smith MK, Hunkapiller J, Bhangale TR, Srinivasan K, Maloney JA, Atwal JK, Sa SM, Yaylaoglu MB, Foreman O, Ortmann W, Rathore N, Hansen D V, et al. 2014. A rare mutation in UNC5C predisposes to late-onset Alzheimer's disease and increases neuronal cell death. Nat. Med. 20: 1452–7.

Wijsman EM, Pankratz ND, Choi Y, Rothstein JH, Faber KM, Cheng R, Lee JH, Bird TD, Bennett DA, Diaz-Arrastia R, Goate AM, Farlow M, et al. 2011. Genome-wide association of familial late-onset Alzheimer's disease replicates BIN1 and CLU and nominates CUGBP2 in interaction with APOE. PLoS Genet. 7: e1001308.

Wilkins JF. 2010. Antagonistic coevolution of two imprinted loci with pleiotropic effects. Evolution 64: 142–51.

Will CL, Lührmann R. 2011. Spliceosome structure and function. Cold Spring Harb. Perspect. Biol. 3: a003707–.

Wind AW, Schellevis FG, Staveren G Van, Scholten RP, Jonker C, Eijk JT Van. 1997. Limitations of the Mini-Mental State Examination in diagnosing dementia in general practice. Int. J. Geriatr. Psychiatry 12: 101–8.

Wisniewski HM, Kozlowski PB. 1982. Evidence for Blood-Brain Barrier Changes in Senile Dementia of the Alzheimer Type (SDAT). Ann. N. Y. Acad. Sci. 396: 119–129.

Wisniewski KE, Wisniewski HM, Wen GY. 1985. Occurrence of neuropathological changes and dementia of Alzheimer's disease in Down's syndrome. Ann. Neurol. 17: 278–82.

Wortmann M. 2012. Dementia: a global health priority - highlights from an ADI and World Health Organization report. Alzheimers. Res. Ther. 4: 40.

Wray NR, Purcell SM, Visscher PM. 2011. Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol. 9: e1000579.

Wu H, Sun S, Tu K, Gao Y, Xie B, Krainer AR, Zhu J. 2010. A splicing-independent function of SF2/ASF in microRNA processing. Mol. Cell 38: 67–77.

Wu Y-T, Fratiglioni L, Matthews FE, Lobo A, Breteler MMB, Skoog I, Brayne C. 2015. Dementia in western Europe: epidemiological evidence and implications for policy making. Lancet Neurol.

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, et al. 2015. The human splicing code reveals new insights into the genetic determinants of disease. Science (80-.). 347: 1254806–1254806.

Xu H-G, Ren W, Zou L, Wang Y, Jin R, Zhou G-P. 2011. Transcriptional control of human CD2AP expression: the role of Sp1 and Sp3. Mol. Biol. Rep. 39: 1479–86.

Xu PT, Gilbert JR, Qiu HL, Ervin J, Rothrock-Christian TR, Hulette C, Schmechel DE. 1999. Specific regional transcription of apolipoprotein E in human brain neurons. Am. J. Pathol. 154: 601–11.

Xue Y, Chen Y, Ayub Q, Huang N, Ball E V, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN, Tyler-Smith C. 2012. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am. J. Hum. Genet. 91: 1022–32.

Yamazaki T, Masuda J, Omori T, Usui R, Akiyama H, Maru Y. 2009. EphA1 interacts with integrin-linked kinase and regulates cell morphology and motility. J. Cell Sci. 122: 243–55.

Yeo G, Holste D, Kreiman G, Burge CB. 2004. Variation in alternative splicing across

human tissues. Genome Biol. 5: R74.

Young-Pearse TL, Bai J, Chang R, Zheng JB, LoTurco JJ, Selkoe DJ. 2007. A critical function for beta-amyloid precursor protein in neuronal migration revealed by in utero RNA interference. J. Neurosci. 27: 14459–69.

Yu J-T, Tan L, Hardy J. 2014. Apolipoprotein E in Alzheimer's disease: an update. Annu. Rev. Neurosci. 37: 79–100.

Yuan Q, Chu C, Jia J. 2012. Association studies of 19 candidate SNPs with sporadic Alzheimer's disease in the North Chinese Han population. Neurol. Sci. 33: 1021–8.

Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, Phatnani HP, Guarnieri P, Caneda C, Ruderisch N, Deng S, Liddelow SA, et al. 2014. An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. J. Neurosci. 34: 11929–47.

Zheng H-F, Ladouceur M, Greenwood CMT, Richards JB. 2012. Effect of genome-wide genotyping and reference panels on rare variants imputation. J. Genet. Genomics 39: 545–50.

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. 2014. Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. U. S. A. 111: E455–64.

Appendix

Appendix 1 - Full sample list for pooled target enriched genome sequencing

Pool	Number	ID	Sample No.	Sex	*APOE*	Centre
1	1	AD219	Missing	F	33	Nottingham
	2	AD232	34/04	F	34	Nottingham
	3	M547	87/23	F	34	Manchester
	4	AD218	108/03	M	34	Nottingham
	5	AD236	30/07	F	22	Nottingham
	6	M659	09/01	M	Missing	Manchester
	7	AD221	77/04	F	33	Nottingham
	8	M523	99/19	M	33	Manchester
	9	M551	96/27	F	34	Manchester
	10	M641	06/01	M	33	Manchester
	11	M604	03/19	F	Missing	Manchester
	12	AD235	106/05	M	24	Nottingham
2	13	AD222	Missing	F	34	Nottingham
	14	AD224	58/06	M	33	Nottingham
	15	AD245	29/07	F	Missing	Nottingham
	16	AD233	45/04	F	34	Nottingham
	17	AD197	20/00	Missing	44	Nottingham
	18	M596	03/01	M	Missing	Manchester
	19	M565	97/08	F	33	Manchester
	20	AD220	32/04	M	Missing	Nottingham
	21	AD238	43/03	F	Missing	Nottingham
	22	M540	96/15	M	Missing	Manchester
	23	M546	87/44	F	33	Manchester
	24	AD257	09D21609	M	Missing	Nottingham
3	25	AD231	118/01	M	34	Nottingham
	26	M605	97/10	M	23	Manchester
	27	AD227	16/07	M	34	Nottingham
	28	AD251	73/08	M	Missing	Nottingham
	29	AD207	189/01	F	34	Nottingham
	30	AD246	73/07	M	Missing	Nottingham
	31	AD212	11/02	F	34	Nottingham
	32	M651	09/15	F	Missing	Manchester
	33	M530	98/15	F	34	Manchester
	34	AD253	08D21533	F	Missing	Nottingham
	35	AD203	107/01	Missing	33	Nottingham
	36	AD210	202/01	M	34	Nottingham

4	37	AD255	08D29579	M	Missing	Nottingham
	38	M571	98/05	F	34	Manchester
	39	M094	99/04	M	33	Manchester
	40	AD206	149/01	F	33	Nottingham
	41	AD249	27/08	F	Missing	Nottingham
	42	M589	96/23	M	34	Manchester
	43	M531	96/02	M	44	Manchester
	44	M593	03/07	M	Missing	Manchester
	45	AD223	51/05	Missing	24	Nottingham
	46	AD201	627/0	Missing	34	Nottingham
	47	AD216	95/03	M	34	Nottingham
	48	M576	98/08	F	33	Manchester
5	49	M643	06/05	F	34	Manchester
	50	AD243	36/06	M	Missing	Nottingham
	51	M579	94/15	M	33	Manchester
	52	M562	93/07	F	33	Manchester
	53	M543	36/90	F	44	Manchester
	54	M522	03/04	F	Missing	Manchester
	55	AD213	118/02	M	33	Nottingham
	56	AD244	19/07	F	Missing	Nottingham
	57	AD225	88/6	M	34	Nottingham
	58	AD199	153/00	Missing	44	Nottingham
	59	M524	99/14	M	23	Manchester
	60	AD242	18/06	F	Missing	Nottingham
6	61	AD248	22/08	M	Missing	Nottingham
	62	M573	96/24	M	23	Manchester
	63	M526	99/08	Missing	34	Manchester
	64	M647	08/03	F	Missing	Manchester
	65	M528	99/09	M	33	Manchester
	66	M646	08/01	F	Missing	Manchester
	67	AD117	Missing	F	33	Nottingham
	68	AD252	74/08	F	Missing	Nottingham
	69	M577	97/29	M	34	Manchester
	70	M599	98/17	F	33	Manchester
	71	M590	04/01	M	Missing	Manchester
	72	AD254	08D27037	F	Missing	Nottingham

7	73	M637	04/05	F	33	Manchester
	74	M645	07/21	M	Missing	Manchester
	75	AD211	06/02	M	34	Nottingham
	76	M648	08/30	F	Missing	Manchester
	77	M594	03/18	F	Missing	Manchester
	78	AD208	191/01	Missing	33	Nottingham
	79	M575	98/03	F	33	Manchester
	80	M639	05/01	M	24	Manchester
	81	M527	99/07	F	33	Manchester
	82	M536	04/04	F	Missing	Manchester
	83	M649	08/31	F	Missing	Manchester
	84	M644	07/08	M	Missing	Manchester
8	85	M529	99/11	M	44	Manchester
	86	AD115	Missing	F	24	Nottingham
	87	M638	04/07	F	Missing	Manchester
	88	M642	06/02	F	33	Manchester
	89	AD141	Missing	F	33	Nottingham
	90	AD146	Missing	F	33	Nottingham
	91	AD112	Missing	M	34	Nottingham
	92	AD107	Missing	M	33	Nottingham
	93	AD123	Missing	F	34	Nottingham
	94	AD125	Missing	M	34	Nottingham
	95	AD157	Missing	M	34	Nottingham
	96	AD109	Missing	M	23	Nottingham

Appendix 2 - Perl script for interleaving paired-end reads

The below Perl script (saved as Interleaving.pl) opens the forward reads file (FILEA) and interrogates the reverse reads file (FILEB) for the matching paired read as indicated by the header information (readnames1 and readnames2). The reverse read is then reverse complemented and the quality score is also reversed, before both reads are written to the OUTFILE interleaved as 1Read1, 1Read2, 2Read1, 2Read2. The OUTFILE was then aligned with BFAST.

```
#!/usr/bin/perl
#Interleaving.pl

$filenameA = $ARGV[0];              #forward reads
$filenameB = $ARGV[1];              #reverse reads
$filenameOut = $ARGV[2];
open $FILEA, "< $filenameA";
open $FILEB, "< $filenameB";
open $OUTFILE, "> $filenameOut";

while(<$FILEA>) {
      $readname1 = $_;              #$_ holds read name
      print $OUTFILE $readname1;

      $read1 = <$FILEA>;
      print $OUTFILE $read1;

      $superfluous_plus = <$FILEA>;
      print $OUTFILE $superfluous_plus;

      $qual1 = <$FILEA>;
      print $OUTFILE $qual1;

      $readname2 = <$FILEB>;
      die("Readnames differ: $readname1 != $readname2") if
($readname1 != $readname2);
      print $OUTFILE $readname2;

      $read2 = <$FILEB>;

      chop $read2;                  #Reverse complement read2
      $read2 = reverse($read2);
      $read2 =~ tr/ACGTacgt/TGCAtgca/;
      print $OUTFILE "$read2\n";

      $superfluous_plus = <$FILEB>;
      print $OUTFILE $superfluous_plus;

      $qual2 = <$FILEB>;
      chop $qual2;

      $qual2 = reverse($qual2);     #Reverse qual2
      print $OUTFILE "$qual2\n";
}
close($OUTFILE)
```

Appendix 3 - VEP and ENCODE annotation tool

This Perl script uses the Ensembl variant effect predictor (VEP) Perl API and includes UCSC track data as well as ENCODE data (files under `--custom` in the script below) to provide additional variant annotation.

```
#!/usr/bin/perl
#proxy_finder_main.pl

use strict;
use warnings;

my ($inputfile) = @ARGV; #input vcf file

#call in VEP Perl Script & custom files
system("perl/home/pmzcwm/Perl/Proxy_Finder/VEP/
      variant_effect_predictor.pl
      --i $inputfile
      --force_overwrite
      --o $inputfile\_OutputVEP
      --hgnc
      --check_existing
      --polyphen p
      --sift p
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/GERP.bed.gz,CONS
_ELEMENT,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/tfbsConsSites.be
d.gz,TRANSFAC,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/CPG.bed.gz,CpG,b
ed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/targetScan_mirna
.bed.gz,MIRNA,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/vistaEnhancers.b
ed.gz,VISTA,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/vistaEnhancers.b
ed.gz,VISTA,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/ENC_DNase.bed.gz
,ENC_DNASE,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/ENC_TFBS.bed.gz,
ENC_TFBS,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Methylation/SK_N
_SH/wgEncodeHaibMethyl450SknshSitesRep1.bed.gz,ENC_SK_N_SH_Meth
450,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Methylation/SK_N
_SH/wg
EncodeHaibMethylRrbsSknshHaibSitesRep1.bed.gz,ENC_SK_N_SH_MethR
RBS1,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Methylation/SK_N
_SH/wgEncodeHaibMethylRrbsSknshHaibSitesRep2.bed.gz,ENC_SK_N_SH
```

```
_MethRRBS2,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/H1_H
ESC/Peaks/wgEncodeBroadHistoneH1hescH3k4me1StdPk.bed.gz,H1HESC_
H3K4me1,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/H1_H
ESC/Peaks/wgEncodeBroadHistoneH1hescH3k4me2StdPk.bed.gz,H1HESC_
H3K4me2,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/H1_H
ESC/Peaks/wgEncodeBroadHistoneH1hescH3k4me3StdPk.bed.gz,H1HESC_
H3K4me3,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/H1_H
ESC/Peaks/wgEncodeBroadHistoneH1hescH3k27acStdPk.bed.gz,H1HESC_
H3K27ac,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/NH-
A/Peaks/wgEncodeBroadHistoneNhaH3k4me1StdPk.bed.gz,NHA_H3K4me1,
bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/NH-
A/Peaks/wgEncodeBroadHistoneNhaH3k4me2StdPk.bed.gz,NHA_H3K4me2,
bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/NH-
A/Peaks/wgEncodeBroadHistoneNhaH3k4me3StdPk.bed.gz,NHA_H3K4me3,
bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/NH-
A/Peaks/wgEncodeBroadHistoneNhaH3k27acStdPk.bed.gz,
NHA_H3K27ac,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/HepG
2/Peaks/wgEncodeBroadHistoneHepg2H3K4me1StdPk.bed.gz,HepG2_H3K4
me1,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/HepG
2/Peaks/wgEncodeBroadHistoneHepg2H3K4me2StdPk.bed.gz,HepG2_H3K4
me2,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/HepG
2/Peaks/wgEncodeBroadHistoneHepg2H3K4me3StdPk.bed.gz,HepG2_H3K4
me3,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/HepG
2/Peaks/wgEncodeBroadHistoneHepg2H3k27acStdPk.bed.gz,HepG2_H3K2
7ac,bed,overlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/Brai
n/ucsfChipSeqH3K4me3BrainCoverage.txt.bed.gz,UCSF_H3K4me3,bed,o
verlap
      --custom
/home/pmzcwm/Perl/Proxy_Finder/VEP/SuperTracks/Histone_Mod/Brai
n/uMassBrainHistonePeaksNeuron.bed.gz,UMASS_H3K4me3,bed,overlap
      --regulatory
      --cache
      --write_cache");
close;
```