Investigating genetic variation in Alzheimer's disease

Tulsi Patel, BSc, MRes

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

June 2018

Abstract

Alzheimer's disease (AD) is the most common form of dementia, now the leading cause of death in the UK, which affects more than 35 million people worldwide. Genome-wide association studies identified 20 genetic loci associated with disease susceptibility, however these only exerted small effects on risk. Next-generation sequencing is now being employed to identify more of the missing heritability. This project utilised whole-exome sequencing to explore genetic variation using the Brains for Dementia Research (BDR) resource, a well-characterised cohort of neuropathologically confirmed samples. Exome-wide and candidate gene approaches were employed to assess coding variants for association with AD, using single-variant and burden tests. Coding variants in other neurodegenerative disease genes were also analysed as potential susceptibility factors for AD. Furthermore, polygenic risk scores (PRS) were generated to explore the ability to classify case and control individuals based on their genetic profiles. A synonymous variant in *PILRA* (rs2405442) was nominally associated with 3-fold increased risk of AD, also contributing strongly to *PILRA* burden. It was previously linked to AD through risk gene *ZCWPW1*; however, it has not been directly associated until now. Additional variants in GWAS gene *ABCA7* (rs3764645, rs3752234, rs3752237, rs4147915) and rare variants in *CLU* were also implicated, further supporting their roles in AD susceptibility. A variant in PD gene *LRRK2* (rs35303786) inferred protection against AD, implicating potential pleiotropy across the two diseases. PRS could distinguish AD cases from controls with 85.3% accuracy and also identified controls with high PRS but no cognitive impairments. This could be useful for identifying individuals at risk of developing AD in the future. We have uncovered tentative associations both in established and newly identified loci; highlighting several interesting candidates for further investigation. Although there remains a large amount of missing heritability, we hope that as the BDR resource grows, we will achieve increased power to detect significant associations with AD.

Publications

Patel T*, Brookes KJ, Turton J, Chaudhury S, Guetta-Baranes T, Guerreiro R, Bras J, Hernandez D, Singleton A, Hardy J, Francis PT and Morgan K (2018) Whole-exome sequencing of the BDR cohort: Evidence to support the role of the *PILRA* gene in Alzheimer's disease. *Neuropathology and Applied Neurobiology*.

Chaudhury S, **Patel T**, Guetta-Baranes T, Brookes KJ, Chappell S, Guerreiro R, Bras J, Hernandez D, Singleton A, Hardy J, Mann D, ARUK Consortium and Morgan K (2018) Polygenic risk score in post-mortem diagnosed sporadic early-onset Alzheimer's disease. *Neurobiology of Aging*. 62:244e1-8.

Boden K, Barber I, Clement N, **Patel T**, Guetta-Baranes T, Brookes KJ, Chappell S, Chapman N, Morgan K, Seymour G and Bottley A (2017) Methylation profiling of key genes associated with sporadic early-onset Alzheimer's disease. *Journal of Alzheimer's Disease Reports*. 1:97-108.

Barber I, Braae A, Clement N, **Patel T**, Guetta-Baranes T, Brookes KJ, Medway C, Chappell S, Guerreiro R, Bras J, Hernandez D, Singleton A, Hardy J, Mann D, ARUK Consortium and Morgan K (2017) Mutation analysis of sporadic early-onset Alzheimer's disease using the NeuroX array. *Neurobiology of Aging*. 49:215e1-8.

Clement N, Braae A, Turton J, Lord J, Guetta-Baranes T, Medway C, Brookes KJ, Barber I, **Patel T**, Milla L, Azzopardi M, Lowe J, Mann D, Pickering-Brown S, Kalsheker N, Passmore P, Chappell S and Morgan K (2016) Investigating splicing variants uncovered by next-generation sequencing the Alzheimer's disease candidate genes, *CLU*, *PICALM*, *CR1*, *ABCA7*, *BIN1*, the *MS4A* locus, *CD2AP*, *EPHA1* and *CD33*. *Journal of Alzheimer's Disease & Parkinsonism*. 6:1-7.

Brookes KJ*, **Patel T***, Zapata-Erazo G, Barber I, Braae A, Clement N, Guetta-Baranes T, Chappell S and Morgan K (2016) Identifying polymorphisms in the Alzheimer's related *APP* gene using the MinION sequencer. *Next Generation Sequencing & Applications*. 3:125

Barber I, Garcia-Cardenas J, Sakdapanichkul C, Deacon C, Zapata-Erazo G, Guerreiro R, Bras J, Hernandez D, Singleton A, Guetta-Baranes T, Braae A, Clement N, **Patel T**, Brookes KJ, Medway C, Chappell S, Hardy J, Mann D, ARUK Consortium and Morgan K (2016) Screening exons 16 and 17 of the amyloid precursor protein gene in sporadic early-onset Alzheimer's disease. *Neurobiology of Aging*. 39:220e1-7.

Contents

Abstract ..... **i**
Publications ..... **ii**
Acknowledgements ..... **vii**
List of tables ..... **viii**
List of figures ..... **ix**
List of abbreviations ..... **x**
Chapter 1: Introduction **1**
1.1 Background and statistics 1
1.2 Symptoms 1
1.3 Pathogenesis 2
1.3.1 Amyloid beta plaques 4
1.3.2 Neurofibrillary tangles 7
1.3.3 Other hypotheses 10
1.4 Diagnosis 11
1.5 Biomarkers 14
1.6 Treatments 20
1.7 Types of Alzheimer's disease 23
1.7.1 Early-onset Alzheimer's disease 23
1.7.2 Late-onset Alzheimer's disease 26
1.8 Genetics of Alzheimer's disease 27
1.8.1 Apolipoprotein E 27
1.9 Genome-wide association studies 28
1.10 Susceptibility genes and pathways 32
1.11 New biological pathways 37
1.12 NGS in complex disease 39
1.14 Aims 44
Chapter 2: General methods **45**
2.1 Lab-based methods 45
2.1.1 DNA extraction 45
2.1.2 DNA quality control 46
2.1.3 Primer design 47
2.1.4 Polymerase chain reaction 47
2.1.5 Sanger sequencing 48
2.1.6 TaqMan genotyping assay for *APOE* 48
2.2 Bioinformatic file formats 50
2.2.1 FASTQ files 50
2.2.2 SAM and BAM files 50

2.2.3 VCF files 50
2.2.4 PLINK files 54
2.3 Power calculations 54
2.3.1 Power to detect variants 54
2.3.2 Power to detect associations 55
Chapter 3: Whole-exome sequencing pipeline 57
3.1 Introduction 57
3.1.1 Next-generation sequencing 57
3.1.2 Whole-exome sequencing process 58
3.1.3 Quality control 61
3.1.4 Alignment 63
3.1.5 Processing 63
3.1.6 Variant calling and filtering 64
3.1.7 Annotation 64
3.2 Aims 65
3.3 Methods 65
3.3.1 Sample preparation 65
3.3.2 Next-generation sequencing analysis 68
3.3.3 Initial quality control 68
3.3.4 Preparing reference files for input 70
3.3.5 Indexing reference sequence 71
3.3.6 Alignment 72
3.3.7 Pre-processing 72
3.3.8 Aligned quality control 73
3.3.9 Add read group information 73
3.3.10 Mark duplicates 76
3.3.11 Realignment around indels 77
3.3.12 Fix mate information 78
3.3.13 Base recalibration 79
3.3.14 Variant calling 81
3.3.15 Joint genotyping 83
3.3.16 Variant quality score annotation 84
3.3.17 Variant quality score recalibration 85
3.3.18 Functional annotation 88
3.3.19 Quality checks 92
3.3.20 Filtering variants 94
3.3.21 Visualisation 94
3.4 Results 95
3.4.1 Sample preparation 95
3.4.2 Quality control 95

3.4.3 Base quality recalibration 104
3.4.4 Variant quality recalibration 104
3.4.5 Quality checks 106
3.4.5 Functional annotation 108
3.5 Discussion 110
3.5.1 Quality control 110
3.5.2 Alignment and QC 112
3.5.3 BQSR 113
3.5.4 Variant calling and recalibration 113
3.5.5 Quality checks 114
3.5.6 Functional annotation 115
3.5.7 Conclusion 116
Chapter 4: Exome-wide analysis in Alzheimer's disease 118
4.1 Introduction 118
4.1.1 Genetics of Alzheimer's disease 118
4.1.2 Sequencing studies in Alzheimer's disease 119
4.1.3 Exome-wide association and burden 120
4.1.4 Polygenic risk scores 121
4.2 Aims 123
4.3 Methods 123
4.3.1 QC in VCFtools 123
4.3.2 Remove singletons 124
4.3.3 SnpEff annotation and filtering 125
4.3.4 Convert to PLINK format 126
4.3.5 Check sex 126
4.3.6 PSEQ 127
4.3.7 Summary statistics in PSEQ 128
4.3.8 Single variant analysis 130
4.3.9 Logistic regression analysis with covariates 131
4.3.10 Burden analysis 132
4.3.11 Polygenic risk scoring 133
4.4 Results 136
4.4.1 Quality control 136
4.4.2 Exome-wide association 137
4.4.3 Logistic regression 142
4.4.4 Burden analysis 147
4.4.5 Polygenic risk scores 149
4.5 Discussion 158
4.5.1 Quality control 161
4.5.2 Whole-exome association 163

4.5.3 Logistic regression analysis 165
4.5.4 Burden analysis 169
4.5.5 Polygenic risk scores 170
4.5.6 Conclusion 180
Chapter 5: Candidate genes 182
5.1 Introduction 182
5.1.1 Pathway analysis 182
5.1.2 Alzheimer's risk genes 183
5.1.3 Screening for causative mutations 193
5.1.4 Genetic pleiotropy 194
5.2 Aims 195
5.3 Methods 195
5.3.1 Pathway analysis 195
5.3.2 Screening for causative mutations 196
5.3.3 Association analysis 197
5.3.4 Burden analysis 198
5.3.5 Sequencing the *TREM2* R47H mutation 198
5.4 Results 199
5.4.1 Pathway analysis 199
5.4.2 Screening of familial AD genes 203
5.4.3 Screening of other neurodegenerative disease genes 205
5.4.4 Single variant association in candidate genes 208
5.4.5 Burden analysis 209
5.4.6 *TREM2* R47H mutation 211
5.5 Discussion 215
5.5.1 Pathway analysis 215
5.5.2 Screening familial AD genes for known causative mutations 217
5.5.3 Screening other neurodegenerative disease genes 218
5.5.4 Association analysis 219
5.5.5 Burden analysis 221
5.5.6 Sanger sequencing confirmed heterozygous *TREM2* R47H mutation 223
5.5.7 Conclusion 223
Chapter 6: General discussion 225
References 240
Appendix 268

Acknowledgements

I would like to thank Kevin Morgan, Keeley Brookes and Tamar Guetta-Baranes for their invaluable supervision, support, biscuits and endless pep talks. I am especially grateful to Kevin for providing me with so many amazing opportunities during my PhD. Thank you to Tamar and Keeley for everything you've taught me and for letting me loose in the lab! A special thanks to Sally, from whom I learnt so much in such a short time.

I am very grateful to the Neuroscience Support Group at the QMC and the University of Nottingham School of Life Sciences for funding this PhD studentship. Of course, I must thank the individuals who donated their tissue and their families, as without them none of this research would have been possible. Additionally, the BDR, ARUK and the Alzheimer's Society who have provided tissue and funding to enable a lot of this work.

Thanks also to the other members in the lab, past and present, for making the last 3 years such a fun and incredible experience – Jim, Sultan, Naomi, Anne, Mel and Chris. Thanks to the students who have contributed to this work.

Huge thanks go to my parents for their love and support and for always being proud of me no matter what. I'd also like to thank my family and friends for always being there to listen to the rants, outbursts and just general craziness. I'd especially like to thank my best friend Luke for absolutely everything, I couldn't have done it without you.

For my uncle, Vijoo Kaka, who always believed in me. Miss you always.

List of Tables

Table 1.1 NINCDS-ADRDRA criteria for diagnosing AD 13
Table 1.2 *APOE* haplotypes 28
Table 1.3 GWAS meta-analysis result 31
Table 2.1 Power to detect associations 55
Table 3.1 Phred scale 62
Table 3.2 Demographics of BDR cohort 67
Table 3.3 Variant metrics 107
Table 4.1 Association analysis results prior to correction 139
Table 4.2 Exome-wide logistic regression analysis 144
Table 4.3 Exome-wide burden analysis 148
Table 4.4 Polygenic risk score outliers and neuropathology 156
Table 4.5 PRS risk genes 159
Table 5.1 *TREM2* R47H primer sequences 200
Table 5.2 Mutation screening in the familial AD genes 204
Table 5.3 Mutation screening in other neurodegenerative disease genes 206
Table 5.4 Association analysis in candidate genes 210
Table 5.5 Burden analysis in candidate genes 212
Table 6.1 Variants selected as potential candidates for further investigation 227

List of Figures

Figure 1.1 Healthy brain compared with Alzheimer's brain 3
Figure 1.2 Neurons in healthy brain compared to Alzheimer's brain 5
Figure 1.3 APP processing pathways 8
Figure 1.4 Neuroimaging scans 12
Figure 1.5 Biomarkers in Alzheimer's disease 19
Figure 1.6 Biological pathways 33
Figure 1.7 Spectrum of genetic variants in complex disease 36
Figure 1.8 Spectrum of genetic variants in AD 42
Figure 2.1 FASTQ file format 51
Figure 2.2 SAM file format 51
Figure 2.3 VCF file format 52
Figure 2.4 PLINK files 53
Figure 2.5 Estimated power to detect associations 56
Figure 3.1 Illumina library preparation and sequencing-by-synthesis 60
Figure 3.2 Gel testing DNA quality 66
Figure 3.3 Next-generation sequencing pipeline 69
Figure 3.4 HaploytypeCaller method 83
Figure 3.5.1 FastQC report for BB001 FASTQ data 97
Figure 3.5.2 MultiQC report for all samples FASTQ data 98
Figure 3.6.1 FastQC report for BB001 aligned BAM data 101
Figure 3.6.2 MultiQC report for all samples BAM data 102
Figure 3.7 SamStats results 104
Figure 3.8 BQSR scores before and after recalibration 105
Figure 3.9 Tranches for filtering variants 106
Figure 3.10 Sequencing coverage per chromosome 108
Figure 3.11 VEP functional annotation on all variants 109
Figure 4.1 PSEQ Phenotype file 129
Figure 4.2 Manhattan plot of association prior to correcting for covariates 138
Figure 4.3 PRSice output 150
Figure 4.4 ROC curve and prediction model 152
Figure 4.5 Comparison of PRS between cases and controls 154
Figure 5.1 Pathway analysis using STRING 201
Figure 5.2 *TREM2* R47H sequencing chromatograms 214

List of Abbreviations

ABCA1	*ABCA1* ATP-binding cassette protein, family A, member 1
ABCA7	ATP-binding cassette protein, family A, member 7
ACh	Acetylcholine
AChE	Acetylcholinesterase
AD	Alzheimer's disease
AICD	Amyloid precursor protein intracellular domain
AKAP9	A-kinase anchoring protein 9
ALS	Amyotrophic lateral sclerosis
APOE	Apolipoprotein E
APP	Amyloid precursor protein
ARUK	Alzheimer's Research UK
Aβ	Amyloid beta
BACE1	Beta secretase 1
BIN1	Myc box-dependant-interacting protein 1
bp	Base pair
BQSR	Base quality score recalibration
CASS4	Cas scaffolding protein family member 4
CD2AP	Cluster of differentiation 2-associated protein
CD33	Sialic acid binding immunoglobulin-like-lectin-3
CELF1	CUG triplet repeat RNA binding protein Elav-like, member 1
CERAD	Consortium to Establish A Registry for Alzheimer's Disease
CLU	Clusterin
CNS	Central nervous system
CR1	Complement receptor 1
CSF	Cerebral spinal fluid
CT	Computed tomography
DLB	Dementia with Lewy bodies
DNA	Deoxyribonucleic acid
EOAD	Early-onset Alzheimer's disease
ESE	Exonic Splicing Enhancer
ESP	Exome Sequencing Project
ESS	Exonic Splicing Silencer
EVS	Exome Variant Server
ExoSAP	Exonuclease I - Shrimp Alkaline Phosphatase
FAD	Familial Alzheimer's disease
FDG-PET	Fludeoxyglucose positron emission tomography
FERMT2	Fermitin Family Member 2
fMRI	Functional magnetic resonance imaging
FTD	Frontal temporal dementia
GBA	Glucosylceramidase beta
GRN	Granulin
GWAS	Genome-wide association study
HDL	High density lipoprotein
HLA-DRB	Human leukocyte antigen D-related beta chain
IGAP	International Genomics of Alzheimer's Project
INDEL	Insertion/deletion
INPP5D	Inositol polyphosphate-5-phosphatase D
IQR	Interquartile range
KASP	Kompetitive Allele Specific Polymerase chain reaction
LD	Linkage disequilibrium

LDL	Low density lipoprotein
LOAD	Late-onset Alzheimer's disease
MAF	Minor allele frequency
MAPT	Microtubule associated protein Tau
MCI	Mild cognitive impairment
MEF2C	Myocyte enhancer factor 2C
MRI	Magnetic Resonance Imaging
mRNA	Messenger ribonucleic acid
MS4A	Membrane-spanning 2-domain family, subfamily A
NA	Not available
NFT	Neurofibrillary tangle
ng	Nano gram
NGS	Next-generation sequencing
NICE	National Institute for Health and Clinical Excellence
NINCDS-ADRDRA	National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association
NMDA	N-methyl-D-aspartate
NME8	Nucleoside diphosphatase kinase family member 8
NTC	Non-template control
OR	Odds ratio
PCA	Posterior cortical atrophy
PCR	Polymerase chain reaction
PD	Parkinson's disease
PET	Positron emission tomography
PiB	Pittsburgh compound B
PICALM	Phosphatidylinositol binding clathrin assembly protein
PLD3	Phospholipase D family member 3
PRS	Polygenic risk score
PSEN	Presenilin
PTK2B	Protein tyrosine kinase 2 beta
RIN3	Ras and Rab interactor 3
RNA	Ribonucleic acid
Rpm	Rotations per minute
SD	Standard deviation
SORL1	Sortilin-related receptor L
TE	Tris-EDTA buffer solution
TF	Transcription factor
T_m	Annealing temperature
TREM2	Triggering receptor expressed on myeloid cells 2
UCSC	University of California, Santa Cruz
UNC5C	Unc-5 netrin receptor C
UTR	Untranslated region
VCF	Variant call file
VEP	Variant Effect Predictor
VSQR	Variant quality score recalibration
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
WT	Wild type
ZCWPW1	Zinc finger CW-type and PWWP domain-containing 1
µl	Microlitre

Chapter 1: Introduction

1.1. Background and statistics

Dementia is now the leading cause of death in the UK, costing the economy nearly £26 billion per year (Lewis *et al.,* 2014; Office for National Statistics, 2017). Dementia incidence doubles every 5 years from ages 65 to 90+ years (Corrada *et al.,* 2010; Jorm and Jolley, 1998). Alzheimer's disease (AD) is the most common form of dementia, accounting for more than two-thirds of all dementia cases amongst the elderly. In the UK approximately 850,000 people are currently living with the disease, a figure expected to rise to over 1 million in the next decade (Prince *et al.,* 2015). Worldwide, an estimated 35 million people suffer from AD, likely to increase almost 4-fold to 115.4 million by 2050 (Ferri *et al.,* 2005). The lifetime risk of developing AD at age 65 is estimated to be 9% for men and 17% for women (Seshadri and Wolf, 2007). The incidence of AD is greater in women than in men, however this does not appear to be merely a result of increased longevity (Alzheimer's Association, 2014). AD is an incurable neurodegenerative disorder, in which the death of brain cells characteristically results in memory loss and cognitive decline (Burns and Iliffe, 2009). German psychiatrist Alois Alzheimer first described the disease in the early 20th century during an autopsy of a 51 year old woman suffering from short-term memory loss and strange behavioural symptoms. The autopsy revealed cerebral atrophy, 'senile' plaques and neurofibrillary tangles, now well-known characteristics of AD (McKhann *et al.,* 1984).

1.2. Symptoms

AD progressively damages neurons in the brain over time, affecting its ability to function normally, however the full aetiology is still largely unknown. The rate of disease progression varies greatly from person to person. Individuals with AD live an average of 8 years after diagnosis but can survive as long as 20 years (Zanetti *et al.,* 2009). While it is known that AD pathology begins before symptoms appear, the

prodromal phase usually presents with mild cognitive impairment (MCI) where some memory loss is observed but daily living is not affected (Petersen *et al*. 1999, Alzheimer's Association, 2014). It is estimated that 10-15% of individuals diagnosed with MCI progress to a clinical or pathological diagnosis of AD per year (Petersen *et al.,* 2009). Further symptoms develop with increasing neuronal degeneration to the extent that daily functioning is impaired (Li *et al.,* 2014). Early stage symptoms include difficulty remembering recent events, struggling to follow a conversation, misplacing items, trouble with planning and getting lost. As the disease develops further, clinical symptoms increase in severity and range from mood swings to behavioural changes, inability to recognise family and friends and losing awareness of surroundings. The later stages are characterised by a profound loss of memory, impaired cognition and communication, disorientation, confusion and eventually, difficulty speaking, swallowing, and walking. As symptoms become increasingly severe, full time care is required and the disease is ultimately fatal (Förstl & Kurz 1999; Alzheimer's Association, 2014).

1.3. Pathogenesis

At the macroscopic level, AD patients exhibit regional brain atrophy, with lesions appearing and the fluid-filled ventricles are enlarged due to severe neuronal cell death. Figure 1.1 illustrates the key differences in brain structure of AD patients compared to healthy controls. Shrinkage of the cerebral cortex and hippocampus are also observed as a result of this neuronal loss. Neuroimaging studies have demonstrated that the progression of neuronal death correlates with the onset and escalation of clinical symptoms as the disease advances. (Whitwell, 2010). AD initially affects the hippocampus, which is involved in memory formation. The disease then spreads to the temporal lobe thereby causing impairments in language comprehension and communication (Wenk, 2003). Damage to regions of the frontal cortex leads to impairment of

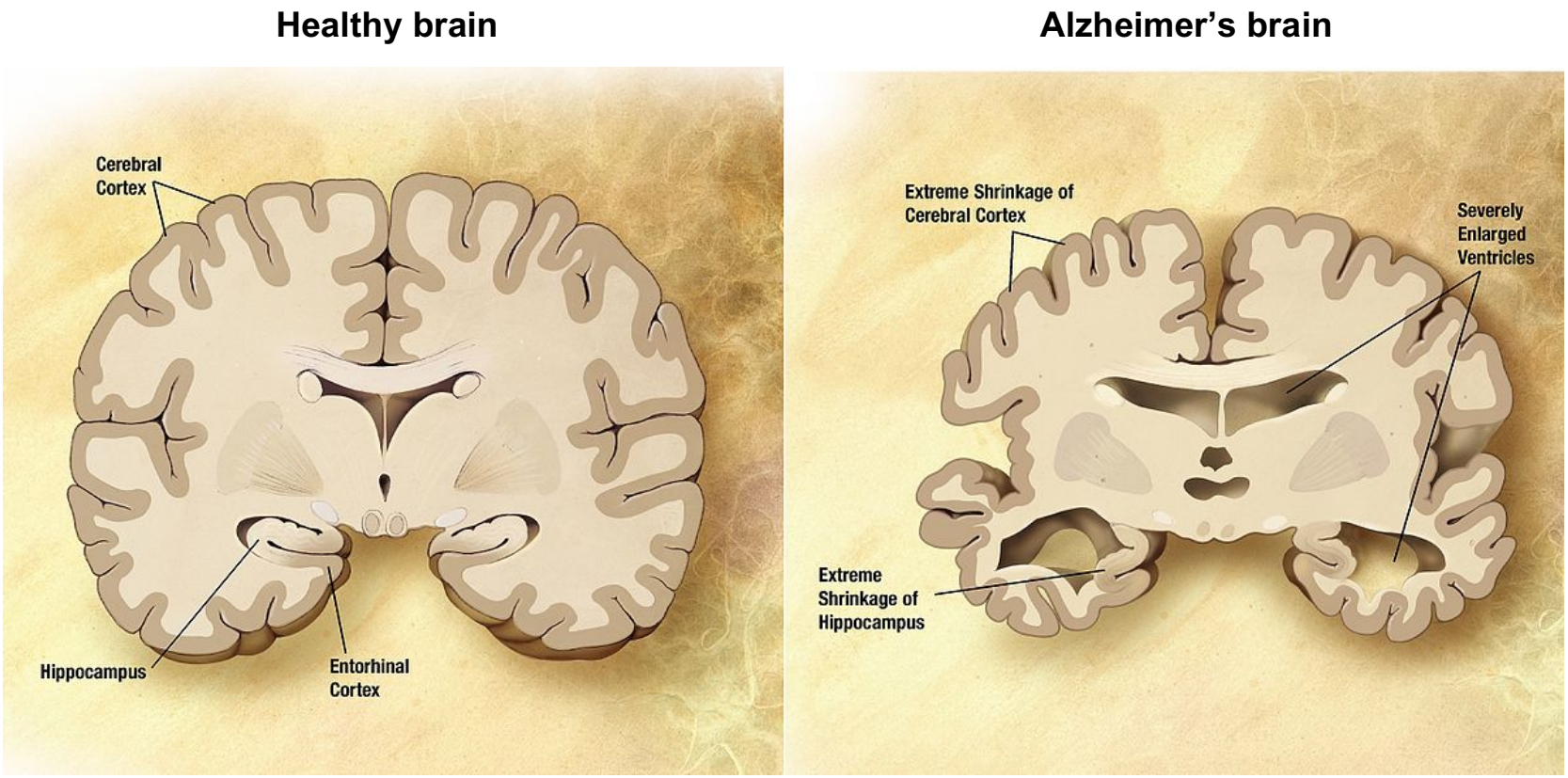


Figure 1.1. Graphic representation comparing a healthy brain and a brain with Alzheimer's disease. The brain with advanced AD (right) shows severe cortical shrinkage, especially in the hippocampal region, suggesting neuronal cell death. The fluid-filled ventricles are enlarged in comparison to the healthy brain (left). Image obtained from Ali *et al*. (2015).

cognition and logical thought, mood swings and other behavioural changes as the cingulate gyrus is also affected (Whitwell *et al.,* 2007). Increased cell death in parietal and occipital lobes, results in poor visuospatial skills, orientation and sensory perception and also often visual hallucinations (Förstl and Kurz, 1999). In the later stages, the disease affects the brainstem, which regulates several critical autonomic functions, such as balance, breathing, blood pressure, digestion and consciousness; finally reaching the cerebellum to affect coordination (Thal *et al.,* 2002).

A definitive diagnosis of AD is only possible upon a post-mortem examination of brain tissue which identifies amyloid beta (Aβ) plaques and neurofibrillary tangles (NFT) of hyperphosphorylated tau protein (Goedert and Spillantini, 2006). AD affected brains present with extracellular aggregates of Aβ as well as intracellular tangles made up of hyperphosphorylated tau protein, as shown in Figure 1.2. These pathological hallmarks both aggregate in large quantities propagating throughout the brain in a predictable pattern. Although Aβ plaques are characteristic of AD, they can be present in cognitively normal individuals and AD brains can also be devoid of plaques. Furthermore, the extent of Aβ deposition is not directly associated with the severity of clinical symptoms, however the pattern of brain atrophy correlates well with the progression of NFT pathology (Whitwell, 2010).

1.3.1 Amyloid beta plaques

Amyloid is a protein found throughout the body; its normal function is still unknown. This peptide is produced when the amyloid precursor protein (APP) is cleaved by α-, β- and γ-secretases. The toxic form, known as Aβ, is the main component of amyloid plaques seen in AD brains (Bird, 2013). In healthy brains, these protein fragments are broken down and cleared by microglia. However, in AD these fragments form insoluble aggregates, which disrupt the connections between neurons at synapses (Pozueta *et al.,* 2013).

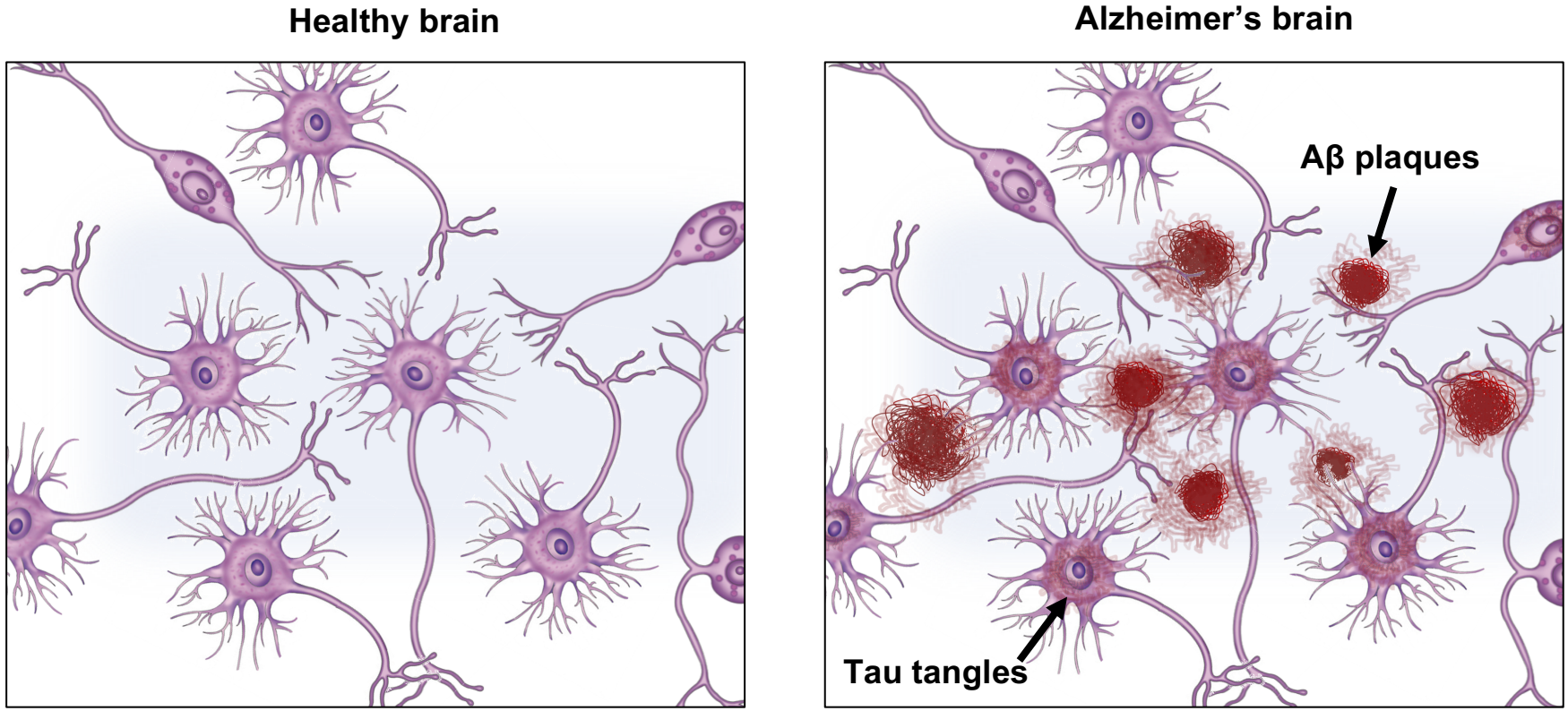


Figure 1.2. Differences in brain pathology between normal and AD affected brains. The structure and localisation of amyloid plaques and neurofibrillary tangles can be seen in relation to neurons in the brain. In comparison to the normal healthy brain (left), AD affected brains (right) often present with an accumulation of aggregated Aβ in the extracellular regions, whilst hyperphosphorylated tau protein builds up to form NFTs within the neurons. Adapted from Mahdi *et al*. (2006).

APP is cleaved via two pathways, the amyloidogenic pathway or non-amyloidogenic pathway (Figure 1.3). The non-amyloidogenic pathway is the predominant APP processing pathway (Figure 1.3a). It requires initial cleavage by α-secretase within the Aβ region, which prevents Aβ formation. It is followed by subsequent cleavage by γ-secretase in the transmembrane region, resulting in a soluble APP-α, a shortened P3 peptide and the APP intracellular domain (AICD) (Lichtenthaler, 2012; Postina, 2012).

In the amyloidogenic pathway (Figure 1.3b), β-secretase, also known as β-site APP cleaving enzyme 1 or BACE1, is the first protease to cleave APP to produce the soluble fragment sAPP-β. Sequential cleavage by γ-secretase at the transmembrane domain produces Aβ peptides and again, the AICD. Proteolysis by γ-secretase occurs at several sites, resulting in Aβ peptides with varying lengths, between 36-43 amino acids long (Marsden *et al.,* 2011). The size of the Aβ fragment can have important implications for AD. The predominant form is the soluble $A\beta_{40}$ (~90%), normally found circulating in plasma, CSF and brain interstitial fluid, followed by $A\beta_{42}$ (~10%) (Jan *et al.,* 2008). Isoform $A\beta_{42}$ is hydrophobic in nature and therefore it is considered the pathogenic form, however senile plaques are comprised of both $A\beta_{40}$ and $A\beta_{42}$ aggregates with greater accumulation of the latter (Gu and Guo, 2013).

The amyloid cascade hypothesis, first proposed in 1991, was a highly influential theory to explain AD pathogenesis (Hardy and Allsop, 1991). It is still widely accepted as one of the most probable theories despite the failure of Aβ-related treatments in clinical trials (Karran *et al.,* 2011). The hypothesis states that abnormal production of Aβ initiates the pathophysiological cascade that leads to AD (Tanzi and Bertram, 2005). Aβ deposition is understood to be toxic to cells, however the mechanism of its toxicity is unknown. One hypothesis posits that the aggregates may block cell-to-cell signalling at synapses, causing loss of

connections between neurons. This leads to neuronal cell death and the characteristic “shrinking” of the cortex seen in brain scans.

A review of the traditional Aβ hypothesis implicated an intermediate oligomeric species of amyloid as the main culprit of neuronal death in AD. Research has suggested that an intracellular form of oligomeric $A\beta_{42}$ can acutely inhibit synaptic transmission by activating casein kinase 2 (Moreno *et al.,* 2009). It has been hypothesised that misfolded Aβ forms prion-like oligomers that can spread to other Aβ molecules, causing them to misfold (Nussbaum *et al.,* 2013). Although insoluble plaques can cause synaptic dysfunction, the formation of plaques could form a defence mechanism to moderate the ability of soluble Aβ to spread, by sequestering it into an insoluble form (Castellani *et al.,* 2009; Treusch *et al.,* 2009). It is becoming clear that soluble Aβ oligomers are likely to be contributing to neurotoxicity in AD more significantly than fibrils (Klein, 2013). There is also evidence that suggests Aβ could induce tau misfolding, addressing the potential relationship between Aβ and tau in AD pathogenesis (Bloom, 2014; Nussbaum *et al.,* 2013).

1.3.2 Neurofibrillary tangles

Tau is a highly soluble protein, mostly found in axons of neuronal cells, which interacts with tubulin to promote assembly into microtubules and modulates their stability (Goedert and Spillantini, 2006). Microtubules help maintain cell structure and transport nutrients and organelles across the neuron (Spillantini and Goedert, 2013). Tau is encoded by the gene microtubule-associated protein tau (*MAPT*). There are six splice isoforms of *MAPT* distinguished by the number of binding domains. These positively charged domains bind to the negatively charged microtubule, therefore variants with a greater number of domains more effectively stabilise microtubules (Ballatore *et al.,* 2007). Tau also controls stability via phosphorylation, regulated by protein kinases - CDK5, GSK3, PKA, PKC and PKN (Churcher, 2006). During the course of AD tau becomes

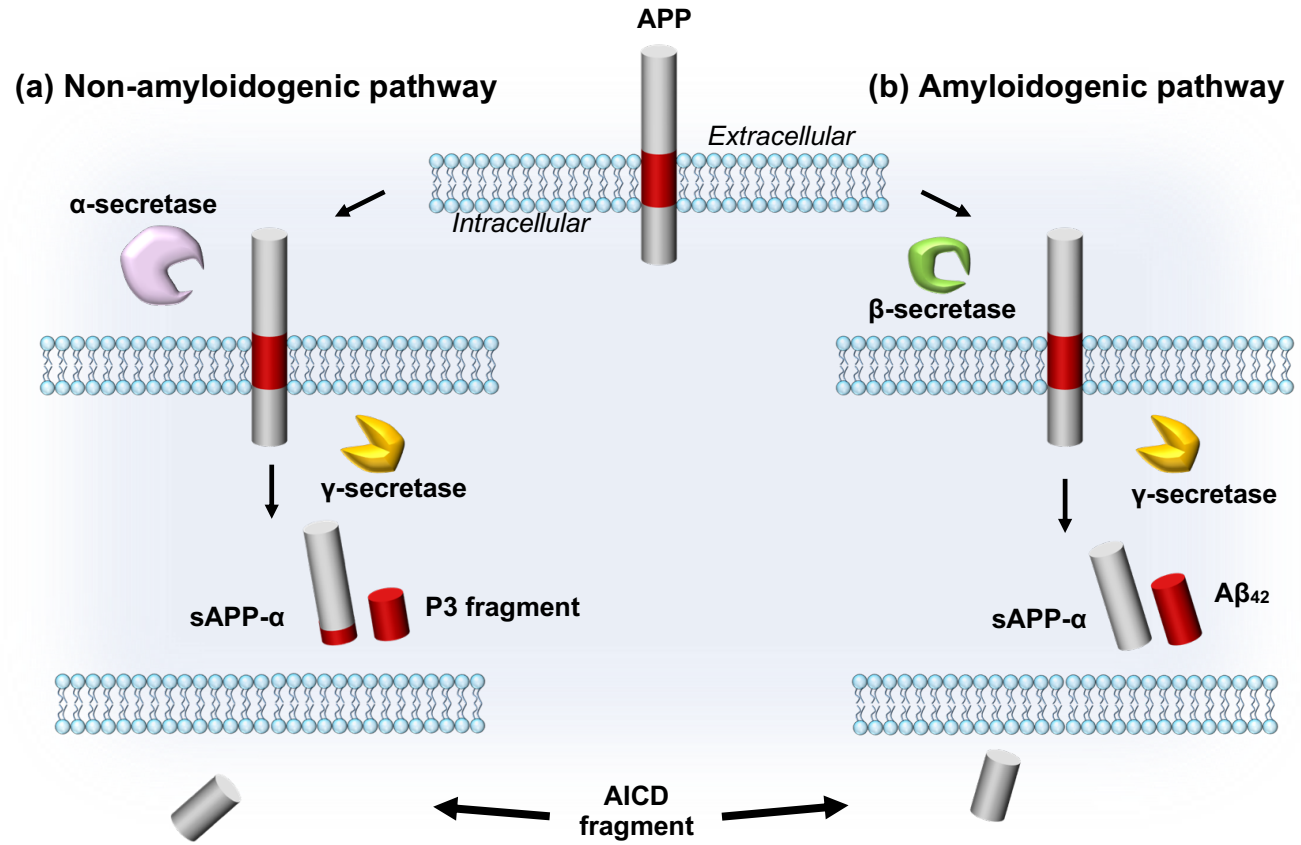


Figure 1.3. The amyloid precursor protein (APP) processing pathways. The APP polypeptide (silver) is a cell membrane protein that is cleaved by proteases via two pathways. In (a) the non-amyloidogenic pathway, the enzyme α-secretase cleaves APP in the Aβ region (red) to release a soluble fragment (sAPP-α) and the C-terminal is cleaved by γ-secretase to produce an APP intracellular domain (AICD) and P3 fragment. In (b) the amyloidogenic pathway β-secretase cleaves APP to produce fragment sAPP-β and $A\beta_{40}$ or $A\beta_{42}$, with γ-secretase again producing AICD. Adapted from Mu and Gage (2011).

hyperphosphorylated reducing its binding affinity to microtubules, resulting in destabilisation (Sato-Harada *et al.,* 1996). In AD affected brains, tau hyperphosphorylation increases the propensity of tau to accumulate in the somatodendritic region of neurons in paired helical filaments, leading to the hallmark neurofibrillary tangles (NFT) (Spillantini and Goedert, 2013). Insoluble tau inclusions are formed in the cytoplasm of neuronal cells, disrupting important transport systems within the cell (Goedert *et al.,* 1995; Roy *et al.,* 2005). These molecules can no longer move along the axon, depriving the cell of vital nutrients and eventually leading to neuronal cell death. Tau aggregates can become self-propagating, spreading through the cortex in a characteristic pattern during disease progression (Nussbaum *et al.,* 2013; Spillantini and Goedert, 2013).

Criteria proposed by Braak and Braak classified the progression of AD-related neurofibrillary pathology into six stages based on the location and extent of NFT formation, known as Braak staging (Braak *et al.,* 2006; Braak and Braak, 1991). In stages I-II NFT involvement is confined to entorhinal regions of the brain. The limbic regions are affected as the tangles accumulate in stages III-IV, affecting the hippocampus which is involved in memory formation, correlating with the cognitive decline observed in AD (Spires-Jones and Hyman, 2014). Stage V-VI signify the advanced stages where NFTs are found throughout the neocortex. NFTs occur naturally with normal ageing, therefore cognitively healthy individuals can often be classified with Braak staging 1-3 (Knopman *et al.,* 2003).

NFTs are also present in other dementias apart from AD including frontotemporal Dementia (FTD), progressive supranuclear palsy, Down's syndrome and Pick's disease, referred to as tauopathies (Neumann *et al.,* 2009). Mutations in the *MAPT* gene are known to cause an inherited form of FTD with Parkinsonism linked to chromosome 17 (FTDP-17), highlighting the high disease penetrance of

tau protein mutations (Hutton *et al.,* 1998; Spillantini and Goedert, 2013).

Aβ plaques and NFTs of hyperphosphorylated tau, can also be seen in the brains of elderly individuals without clinical symptoms of AD (Villemagne *et al.,* 2008). Coupled with the failure of Aβ- and tau-related drugs to reverse cognitive deficits in clinical trials, the role of Aβ and tau as causative factors is being reconsidered. It may be more likely that these pathological features are consequences of the disease process, resulting from other, yet unknown factors.

1.3.3 Other hypotheses

Several theories exist that aim to shed light on the pathogenesis of AD, however none can conclusively explain what causes AD. The oldest theory is the “cholinergic hypothesis” which is the basis of most FDA approved drug treatments used today. It suggests that AD symptoms arise from a deficiency in the production of neurotransmitter acetylcholine (Ach). While cholinesterase-inhibiting drugs can be effective in managing symptoms, they have not provided any significant benefits in terms of halting or reversing disease progression (Francis *et al.,* 1999). This led to the conclusion that ACh deficiencies are not a cause of AD, but rather an effect of the disease.

The “glutamatergic hypothesis” is based on the notion that overstimulation of NMDA receptors causes excessive glutamate release. This excitotoxicity can lead to cell death, which also occurs in other neurological diseases like Parkinson’s disease and Multiple Sclerosis (Lipton, 2006). However, NMDA receptor antagonists based on this hypothesis have shown very little benefit for alleviating symptoms.

The ‘‘vascular hypothesis’’ aims to supersede the amyloid beta hypothesis positing that vascular dysfunction plays a major role in AD development. A multitude of evidence suggests that exposure to

several vascular risk factors are associated with increased risk of AD. Arterial and systolic hypertension, high cholesterol levels and obesity have all been linked to an increased risk of developing AD (Kivipelto and Solomon, 2006; Launer *et al.,* 2000; Xu *et al.,* 2011). Additionally, patients with atherosclerosis are 3 times more likely to develop AD (Hofman *et al.,* 1997). This has led to the classification of such cases as vascular dementia. It is hypothesised that poor vascular health leads to hypoxia, oxidative stress and blood brain barrier dysfunction which leads to increased inflammation and neuronal damage. Moreover, it is suggested that this leads to subtle changes in Aβ metabolism which lead to the pathogenesis of AD (Xu *et al*. 2011).

1.4 Diagnosis

Several frameworks have been established to standardise the diagnostic procedure for identifying clinical signs of AD, the most prevailing of which is the NINCDS-ADRDA criteria, published in 1984 by NINCDS (National Institute of Neurological and Communicative Disorders and Stroke) and ADRDA (Alzheimer's Disease and Related Disorders Association) (McKhann *et al.,* 1984). It is constantly undergoing review to incorporate the best diagnostic tools to distinguish AD from other neurodegenerative conditions and dementias (Dubois *et al.,* 2007). It has effectively increased the reliability of diagnosing probable AD, with sensitivity as high as 81% in clinical pathological studies (Blacker *et al.,* 1994; Knopman *et al.,* 2007). A review of the clinical criteria called for revisions to include results from magnetic resonance imaging (MRI), positron emission tomography (PET) imaging and genetic information relating to mutations in *APP* and *PSEN* genes. Biomarker evidence was also integrated into the diagnostic guidelines, however due to some of the drawbacks they are limited for use in research settings only. Distinguishing features of other dementias (FTD, DLB, vascular dementia, PCA) and differentiating between possible AD and mild cognitive impairment was also proposed to improve diagnostic accuracy for AD (McKhann *et al.,* 2011).

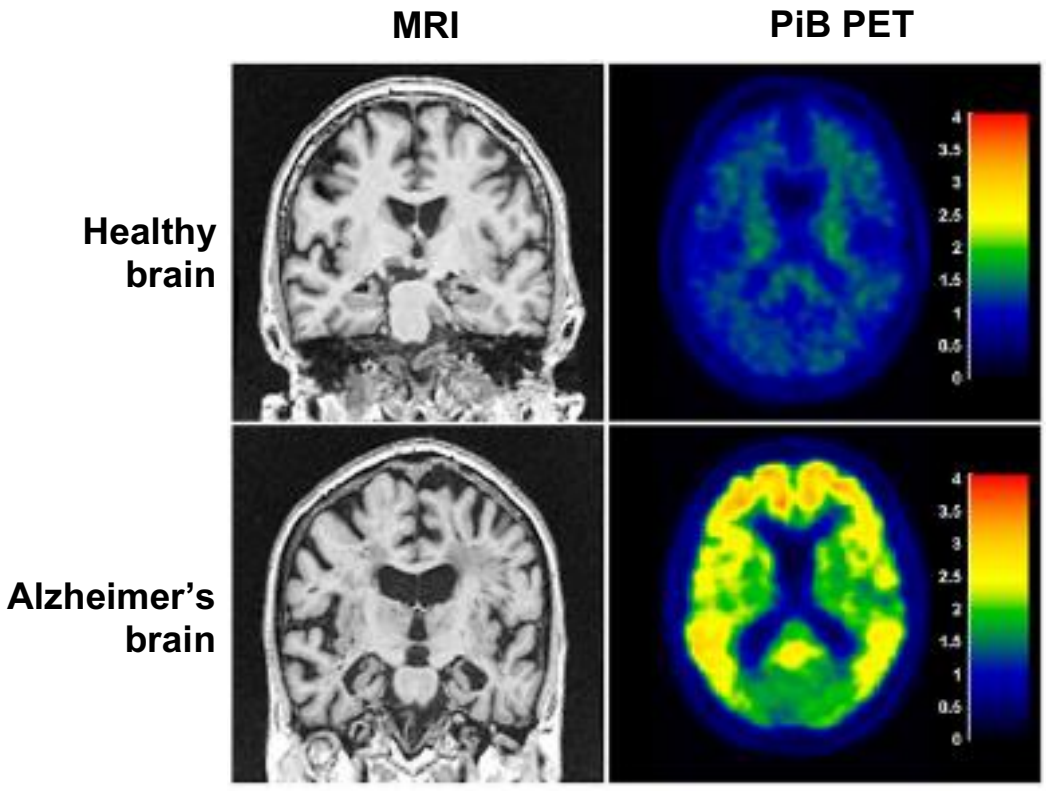


Figure 1.4. Neuroimaging scan comparing a normal and AD brain. MRI scans and PiB-PET with amyloid imaging show apparent differences between the control (top) and AD brain (bottom). Severe enlargement of ventricles and hippocampal atrophy can be seen in the AD brain in both types of imaging, consistent with brain changes found post mortem, compared to the healthy brain (NINCDS-ARDRA). Figure obtained from Mayo Clinic website with permission.

Structural imaging using MRI can identify the characteristic brain changes in AD to reach a more definitive diagnosis (Jobst *et al.,* 1998). PET has been used in combination with radiotracers for imaging cerebral Aβ deposition in the brain. Pittsburgh compound B (^{11}C-PiB) is the most studied radioligand, which binds specifically to fibrillar Aβ peptides with high affinity (Minati *et al.,* 2009). PiB retention rates allow quantitative measurement of fibrillar amyloid levels which strongly correlate with Aβ plaque load (Vlassenko *et al.,* 2012). Figure 1.4 shows imaging scans from MRI and PiB-PET, with obvious degeneration in the AD brain compared to the healthy individual. These imaging techniques are able to detect pathological changes in pre-clinical stages facilitating early diagnosis (Johnson *et al.,* 2012).

Table 1.1. NINCDS-ADRDA criteria for clinical diagnosis of AD. Summary of NINCDS-ADRDA core clinical criteria used to assess certainty of AD diagnosis. Individuals are classified at four stages using data from cognitive tests such as MMSE, neuroimaging scans, physical symptoms, family history, histopathology and diagnosis of any comorbid diseases. Adapted from McKhann *et al*. (2011).

Stage	Symptoms
Unlikely AD	• Sudden onset dementia with focal neurologic symptoms, seizures or gait disturbance in early stages • Negative Aβ PET scan • MMSE score > 27
Possible AD	• Dementia with atypical onset • Any comorbid disease unlikely to be origin of dementia • Positive Aβ PET scan
Probable AD	• Neuropsychological examination (low MMSE score) • Impairments in 2 or more cognitive domains • Onset of symptoms between 40 and 90 years of age • Absence of other disease that could be causing dementia
Definite AD	• Fits clinical criteria for probable AD • Histopathologic evidence of AD from autopsy or biopsy • MMSE score < 10

The severity of AD can be quantitatively assessed using a Mini Mental State Examination (MMSE); a 30-point questionnaire designed to test a number of different mental abilities, including memory, attention and language (Folstein *et al.,* 1975). The test can be used to assess changes over time in an individual already diagnosed with AD to give insight into progression and scores are also considered by clinicians when determining the best course of treatment (Pangman *et al.,* 2000). Test scores are influenced by educational background, age and other impairments such as vision, hearing or physical disabilities (Pangman *et al.,* 2000). A score above 27 out of the total 30 indicates a normal level of cognition, with lower scores suggestive of cognitive impairment. Scores below this in certain tasks can suggest mild (19-24 points), moderate (10-18 points) or severe (<9 points) cognitive impairment as a result of dementia (Mungas, 1991; Yaari and Corey-Bloom, 2007). Studies have demonstrated that the MMSE has been able to differentiate between different types of dementias. Patients with AD are found to score significantly lower on orientation and recall tasks when

compared to patients with DLB, vascular dementia and Parkinson's disease dementia (Jefferson *et al.,* 2002). Table 1.1 shows the updated criteria required for classification of AD, as determined by family and patient history, neuropsychological tests like MMSE to assess 8 domains of cognitive function, brain imaging scans and biomarker profiling (Dubois *et al.,* 2007).

1.5 Biomarkers

The diagnostic accuracy for classifying AD based on current clinical criteria varies greatly, with median sensitivity and specificity estimates of 87% and 58%, respectively (Beach *et al.,* 2012). However, AD pathology can exist for decades prior to onset of the first symptoms. Therefore, diagnosis can occur too late in the disease process for potential disease-modifying drugs to be effective. AD diagnosis is still largely based on clinical exclusion criteria. Consequently, there is a need for reliable biomarkers to detect preclinical signs of AD for early diagnosis, monitoring disease progression and the effect of treatments (Mattsson, 2011). Moreover, identifying at-risk individuals before symptoms can manifest could help delay or halt the subsequent neurodegeneration and eventual cognitive decline.

CSF biomarkers

Cerebrospinal fluid (CSF) surrounds the CNS, circulating hormones and nutrients to the necessary areas of the brain and removing waste products. This direct contact with the brain can allow pathological processes to be monitored by measuring AD-related protein levels in CSF. There are three established CSF biomarkers that have been shown to reliably diagnose AD with sensitivity and specificity above 80% (Blennow and Hampel, 2003). Levels of $A\beta_{42}$, total tau (t-tau) and phosphorylated tau (p-tau) in the CSF correlate with presentation of plaques and tangles, however this is not specific to AD (Llorens *et al.,* 2016). CSF $A\beta_{42}$ shows a significant reduction in AD patients compared to controls (Forlenza *et al.,* 2015; Humpel, 2011); suggested to be a

result of increased Aβ aggregation into amyloid plaques (Motter *et al.,* 1995). Tau protein levels increase gradually with age, however both t-tau and p-tau are observed at significantly greater levels in AD patients, proving to be good prognostic markers (Formichi *et al.,* 2006; Sharma and Singh, 2016).

Although these biomarkers have high accuracy for distinguishing AD from controls, studies have shown limited ability to discriminate AD from other major forms of dementia, with the exception of FTD (Ewers *et al.,* 2015; Kaerst *et al.,* 2013). This can be attributed in part to a significant overlap in dementia pathologies and also the fact that many cases may present with mixed aetiology (Rosén *et al.,* 2013). P-tau is the major component of NFT and is better able to discriminate AD from other dementias with greater sensitivity and specificity than t-tau (Hampel *et al.,* 2004; Koopman *et al.,* 2009). CSF tau has proven to be a reliable predictive marker of conversion from MCI to AD, with high CSF tau levels found in 90% of MCI converters, but not non-converters (Blennow, 2004). Analysing CSF $A\beta_{42}$, t-tau and p-tau levels in combination provides significantly greater diagnostic accuracy that is predictive of future cognitive decline (Mattsson *et al.,* 2009). Soluble TREM2 (sTREM2) is also being investigated for its use as a CSF marker of inflammatory changes and microglial activation. It was shown to correlate with increased CSF t-tau and p-tau levels (Suárez-Calvet *et al.,* 2016), highlighting its potential as a marker of inflammation in the early symptomatic phase of AD. However, obtaining CSF requires an invasive lumbar puncture which reduces the likelihood of patient follow up.

Imaging biomarkers

Neuroimaging provides a non-invasive but rather expensive alternative to CSF sampling. This technique is useful for generating structural and functional information about the brain to monitor disease progression. MRI and CT measure loss of brain volume, cortical atrophy and NFTs which correlate with neuronal loss. MRI is able to distinguish between

atrophy in early AD and normal ageing with high sensitivity and specificity (Scheltens *et al.,* 2002). It can also be used to track progression to MCI and AD and uses different atrophy patterns to discriminate between AD and other dementias (Nobili and Morbelli, 2010).

PET has been successfully used in combination with several radiotracers for imaging Aβ within the brain, including ^{11}C-PiB as previously described. PET combined with 18F-2-deoxy-2-fluoro-D-glucose (FDG) was experimentally tested to measure glucose levels as a potential metabolic marker for early diagnosis. Severe reductions of glucose metabolism in the parieto-temporal regions has been reported in AD patients compared to healthy controls, correlating with cognitive decline (Mistur *et al.,* 2009; Nobili and Morbelli, 2010). It allowed prediction of MCI to AD conversion with 75% accuracy (Herholz *et al.,* 2011). However, FDG-PET is less sensitive in preclinical disease than PiB-PET, which also strongly correlates with hippocampal atrophy as measured by MRI. Unfortunately, imaging techniques require expensive equipment and experienced personnel for their routine application in AD diagnosis.

Translocator protein (TSPO) is a mitochondrial membrane protein, that has been investigated for use as a microglial activation marker (Alam *et al.,* 2017). It is present at low concentrations in resting microglia and is upregulated in response to brain injury and inflammation. The tracer ^{11}C-PK11195 is still the most widely used tracer for PET imaging of inflammation, despite its low binding specificity and brain uptake. Second generation TSPO ligands, such as ^{11}C-DAA1106 and ^{11}C-PBR28 have shown increased binding in AD patients compared to controls, with the latter also correlating with clinical severity and loss of grey matter (Zimmer *et al.,* 2014). Increased binding to TSPO by ^{11}C-PBR28 is associated with greater cognitive decline and cortical atrophy (Kreisl *et al.,* 2016). However, these ligands still do not have sufficient binding affinity for activated microglia in the brain as TSPO expression

is also increased in reactive astrocytes and other immune cells, making it difficult to differentiate cell types (Vivash and OBrien, 2016). Despite the limitations of targeting microglial activation with TSPO tracers, PET imaging of neuroinflammation could provide an effective marker for tracking disease progression after onset.

Blood-based biomarkers

Alternatively, biomarkers found in blood plasma are far more accessible and repeat sampling is comparatively straightforward. Many studies have examined $A\beta_{40}$ and $A\beta_{42}$ plasma levels in early AD, however conflicting results were obtained, indicating that levels can be elevated, reduced or unchanged between AD and control patients (Hansson *et al.,* 2010; Mayeux *et al.,* 1999; van Oijen *et al.,* 2006). A panel of molecular markers associated with AD, previously identified by mass spectrometry, was examined in clinical plasma samples to assess their potential diagnostic utility (Richens *et al.,* 2014). The panel proposed 9 plasma proteins that could determine the presence or absence of AD with 87.5% accuracy, identifying keratin 9 and SPARCL1 as the most promising markers. Other studies have also found biomarker patterns that could distinguish AD from controls or MCI, by measuring the levels of several plasma analytes including clusterin (Doecke *et al.,* 2012; Ray *et al.,* 2007).

The advent of a cheap and rapid blood test to detect AD, based on a panel of biomarkers, is looking more feasible in the near future. However, although this offers a potentially high-throughput and inexpensive screening approach, findings need to be corroborated in larger studies first. Due to the blood-brain barrier, there is a lower concentration of brain-derived proteins in the blood than the CSF. Therefore, the replication of blood biomarker results has been difficult in comparison to CSF. Plasma contains tens of thousands of proteins, making it difficult to detect those at low concentrations. Additionally, plasma is derived from peripheral tissues and consequently not reflective of Aβ production in the brain (Olsson *et al.,* 2016). Conflicting

results indicate that Aβ peptides in plasma have little clinical use as diagnostic markers (Thambisetty and Lovestone, 2010).

Other biomarkers

Circulating miRNAs in the blood have been proposed as potential biomarkers for AD. They are non-coding RNAs involved in post-transcriptional regulation of gene expression, which may play a role in modulating the expression of AD-related genes (Delay *et al.,* 2012). Several groups have identified miRNA expression signatures that are altered in AD in brain, blood and CSF (Annese *et al.,* 2018; Cheng *et al.,* 2015; Leidinger *et al.,* 2013; Riancho *et al.,* 2017). Leidinger *et al*. identified a panel of 12 miRNAs through NGS of blood miRNA transcriptomes that was able to differentiate between AD and control patients with 93% accuracy. The accuracy was reduced to 74-78% when trying to distinguish AD from other neurological diseases (Leidinger *et al.,* 2013). Cheng *et al*. identified a further 16 miRNAs, however there was no overlap in the miRNA signatures. Although there is growing evidence for the role of miRNAs as potential diagnostic markers for AD, there appears to be no consensus panel of miRNAs that can provide an accurate profile of disease progression (Zhao, Bhattacharjee, *et al.,* 2015). Additionally, it remains difficult to determine whether these changes are a cause or consequence of AD pathogenesis.

Inflammatory markers have also been investigated, with a promising study showing that a combination of 18 inflammatory biomarkers in plasma were able to diagnose AD with nearly 90% accuracy (Ray *et al.,* 2007). It has also been proposed that a multi-biomarker panel may provide greater accuracy to generate a profile of disease progression (Humpel, 2011).

Current guidelines for AD diagnosis primarily use MRI scans to support clinical diagnoses, with CSF biomarker sampling if further discrimination

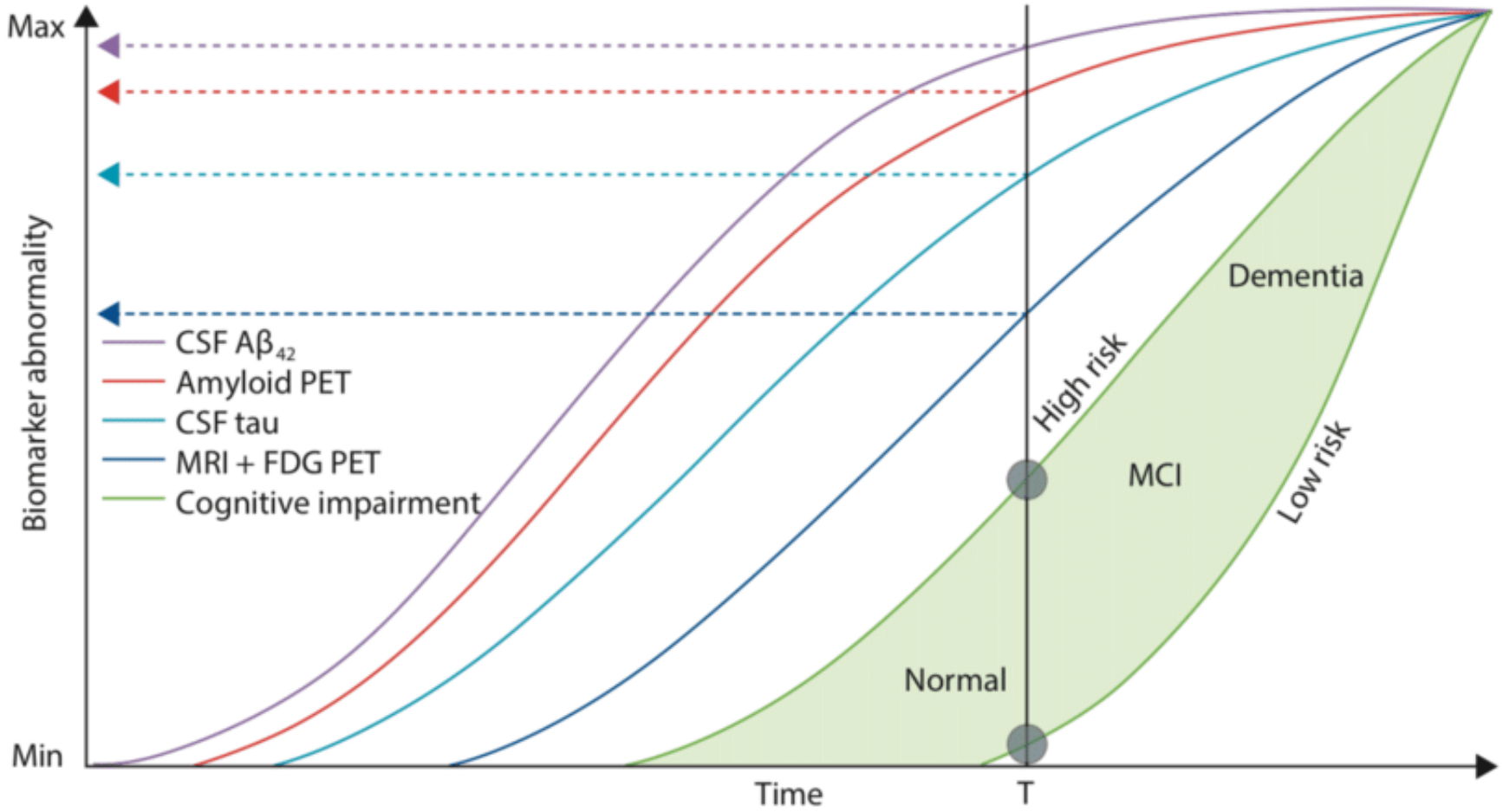


Figure 1.5. Pathological cascade for dynamic biomarkers of Alzheimer's disease. The dynamic changes in AD biomarkers over the course of disease progression (time). Amyloid beta is identified by CSF Aβ42 (purple) or amyloid PET imaging (red). Tau-mediated effects are measured by elevated CSF tau levels (light blue). Extent of neurodegeneration is determined by structural MRI and FDG-PET (dark blue) and cognitive impairment (green area) is expressed as a zone due to the variability between individuals. Image obtained from Jack *et al*. (2013).

is required. Figure 1.5 illustrates the changes in biomarkers over the course of the disease, from normal cognition to advanced dementia (Jack *et al.,* 2013). Individual changes can also be observed in other dementias, which highlights the need for combined biomarker analysis in the preclinical stages to identify at risk individuals prior to cognitive impairments. Although the biomarkers have their advantages, they each have drawbacks that limit their frequent use as diagnostic markers for AD. Since blood testing is the most cost-effective and accessible method to detect changes in AD, finding reliable plasma biomarkers is desirable. Plasma biomarker panels have the potential to address the issues posed by current methods to greatly improve diagnosis and monitoring of progression and treatment efficacy. Additionally, circulatory miRNA and inflammatory biomarkers are being developed and a combination of these could provide a more accurate diagnosis in future.

1.6 Treatments

There is currently no cure for AD as the aetiology of the disease is still largely unclear. Available treatments are not highly effective and only provide temporary symptomatic relief of between 6-12 months in most cases. There are two classes of FDA-approved drugs used for the management of symptoms in AD; acetylcholinesterase (AChE) inhibitors and N-methyl-D-aspartate (NMDA) receptor antagonists. Acetylcholine plays an important role in learning and memory. In the cholinergic hypothesis, cognitive symptoms are caused by the breakdown of the neurotransmitter by AChE, which is depleted in AD brains due to the loss of cholinergic neurons (Wenk, 2003). Reversible AChE inhibitors (Donepezil, Rivastigmine and Galantamine) act to prevent the breakdown of acetylcholine by blocking AChE activity to increase levels of the chemical at synapses (Čolović *et al.,* 2013). The National Institute for Health and Care Excellence (NICE) provides guidelines for the administration of drugs by the NHS. AChE inhibitors are recommended during the initial stages of AD, as they are most

effective at managing mild to moderate symptoms of AD, with 40-70% of patients experiencing improvements (National Institute for Health and Clinical Excellence, 2011).

Memantine is an NMDA receptor (NMDAR) antagonist that targets the glutamatergic system to regulate glutamate activity. Excessive amounts of glutamate are released when neurons are damaged in AD, resulting in excitotoxicity and consequently neuronal cell death. The drug binds to NMDAR with greater affinity than Mg^{2+} to counteract the voltage-dependent block on the receptors by excess glutamate. This prevents a prolonged influx of Ca^{2+} without affecting the action of NMDARs by higher concentrations of glutamate, thereby protecting cells from chronically elevated glutamate levels. Memantine is recommended for the treatment of moderate to severe AD and can replace or act as an adjuvant to AChE inhibitors (National Institute for Health and Clinical Excellence, 2011). The drug can slow progression of symptoms helping to ease the difficulty of daily activities and some evidence indicates it can also help with behavioural symptoms such as aggression, agitation and delusions (Wilkinson, 2012). Donepezil is the only AChE inhibitor also approved to treat all stages of AD as a combination of Donepezil and Memantine have shown statistical significance in their effect, although clinical effectiveness remains marginal at best (Ihalainen *et al.,* 2011; Raina *et al.,* 2008).

Given that the amyloid cascade hypothesis has been the most prominent theory of AD for decades, most drugs in development target the Aβ protein. There have been no new drugs approved since 2003 and recent clinical trials have seen very little success to date with drugs that aim to reduce Aβ production, prevent aggregation and increase clearance (Hung and Fu, 2017; Mehta *et al.,* 2017). Monoclonal antibody Solanezumab, developed by Eli Lilly, produced promising results in Phase 3 trials when the worsening of cognitive symptoms appeared to slow down in patients with mild AD compared to the placebo (Doody *et al.,* 2014). The drug binds to soluble Aβ, causing

$A\beta_{42}$ to solubilise in order to maintain the equilibrium in CSF and reduce levels of this toxic form (Farlow *et al.,* 2012). There was no effect in moderate AD patients, possibly indicating that treatments should be administered much earlier in the progression as irreversible neuronal loss may have already occurred (Le Couteur *et al.,* 2016). However, a second Phase 3 trial focusing on mild AD patients failed to show positive results.

The amyloid cascade hypothesis suggests that removing Aβ will reverse or prevent the clinical symptoms, however, treatments targeting Aβ have failed to successfully reduce Aβ burden (Le Couteur *et al.,* 2016). Aducanumab from Biogen Inc. is another antibody that targets amyloid by clearing Aβ; however, it has shown some promise in a small-scale Phase 1b clinical trial of 165 participants with Aβ build-up who were in the earliest stages of AD. All patients who received the drug experienced a reduction in amyloid deposition in the brain and those with the highest dose had no detectable amyloid deposits in their brain after a year (Sevigny *et al.,* 2016). Slower cognitive decline was also observed in most patients that were administered the drug. Furthermore, cognitive function remained stable in AD patients compared to controls. This gave some credence to the notion that clearance of Aβ deposits can stall progression of disease. Although these findings were profound, a larger scale trial with 2700 patients from 20 countries will garner results in 2020 and will determine the efficacy of Aducanumab as a preventative treatment.

Given the repeated failures of drugs targeting the Aβ pathway, there has been increased interest in developing compounds that target tau, especially as it is more closely correlated with AD pathogenesis. Therapeutics designed to target the tau pathway mainly focus on reducing or preventing the tau hyperphosphorylation that leads to protein misfolding (Panza *et al.,* 2016). AADvac-1 (Axon) is a high affinity antibody and the first anti-tau vaccine to enter clinical trials. It works by inducing an immune response against tau to produce

antibodies targeting abnormal forms of the protein. AADvac-1 is currently in Phase II trials using patients with mild-to-moderate AD (Novak *et al.,* 2016; Panza *et al.,* 2016).

1.7 Types of Alzheimer's disease

AD can be categorised into two forms by age when symptoms first occur. Generally, onset of symptoms occurs after the age of 65, affecting about 6% of the population (Burns and Iliffe, 2009). This represents the majority or around 95% of all AD cases (Piaceri *et al.,* 2013). Onset before age 65 is classified as early-onset AD (EOAD), which only accounts for about 5% of cases (Mendez, 2012; Minati *et al.,* 2009).

1.7.1 Early-onset Alzheimer's disease

The disease can be further differentiated into familial and sporadic AD, with familial cases predominantly being the early-onset form (EOFAD) and accounting for 13% of all early-onset cases (Ertekin-Taner, 2007). It is inherited in an autosomal-dominant manner, as a result of rare but highly penetrant mutations in one of three genes involved in Aβ production – *APP*, *PSEN1* or *PSEN2* (Bertram *et al.,* 2010). The majority of highly penetrant mutations in Mendelian disorders reside within protein-coding regions (Choi *et al.,* 2009). To date, a total of 231 pathogenic mutations have been identified that cause familial AD, listed in the *Alzheimer's Disease and Frontotemporal Dementia Mutation Database* (http://www.molgen.vib-ua.be/ADMutations). There are 32 known pathogenic variants of *APP*, 221 in *PSEN1* and 19 in *PSEN2* that cause EOFAD AD (Lanoiselée *et al.,* 2017). Most mutations in these genes increase the production of the neurotoxic isoform of Aβ, prone to aggregation.

Genetic mutations in EOFAD have provided support for the amyloid hypothesis as mutations in *APP* and Presenilin genes are directly or indirectly involved in the production, regulation or aggregation of Aβ

peptides. $A\beta_{40}$ is normally soluble but mutations in these genes can increase the production of hydrophobic isoform $A\beta_{42}$ (Nussbaum *et al.,* 2013). Genetic studies have demonstrated that the high penetrance of such mutations which result in Aβ dysfunction is sufficient to cause EOFAD. An interesting link between AD and Down's syndrome revealed that patients with trisomy 21 almost universally exhibit AD-like traits by the age of 40 (Lott and Head, 2005). This provides further support for the genetic basis of EOFAD and the amyloid cascade theory, as the *APP* gene is located on chromosome 21, in triplicate in Down's syndrome. This increases the likelihood of abnormal Aβ production, leading to the characteristic features of AD.

EOAD in its sporadic form shares the same traits as LOAD aside from age of onset. Despite differences between familial and sporadic AD they are clinically indistinguishable (Piaceri *et al.,* 2013). This may be due to yet undetermined mutations in known or even unknown genes, with varied levels of penetrance (Antonell *et al.,* 2013).

Amyloid precursor protein

Amyloid precursor protein or APP on chromosome 21q21.3 is an integral membrane protein, with isoforms ranging between 639 and 770 amino acids in length. Proteolysis of this molecule by secretase enzymes generates Aβ, the principal component of senile plaques seen in AD pathology. The majority of *APP* mutations are found on exons 16 and 17, corresponding to secretase cleavage sites and the transmembrane domain (Bekris *et al.,* 2010). Mutations in the vicinity of β- or γ- cleavage sites appear to alter the ratio of $A\beta_{42}$ to $A\beta_{40}$ or total Aβ, whereas substitutions within the peptide seem to directly increase production of the toxic isoform $A\beta_{42}$ (Bekris *et al.,* 2010; St George-Hyslop, 2000). A substitution mutation p.A673T adjacent to the β-secretase cleavage site reduces the formation of toxic Aβ peptides by ~40% *in vitro* and protects against cognitive decline in the elderly without AD (Jonsson *et al.,* 2012). The strong protective effect of this variant against AD suggests that reducing cleavage via the

amyloidogenic pathway could be a target for therapeutic intervention.

Presenilin 1

The presenilin genes (*PSEN1* and *PSEN2*) form the catalytic core of the multiple subunit complex, γ-secretase. It comprises of 4 main components - PSEN1 or PSEN2, Nicastrin, anterior pharynx defective 1 (APH-1) and presenilin enhancer protein 2 (PEN2), required for stability (Bekris *et al.,* 2010). Presenilins play an important role in the pathogenesis of familial AD, as they are involved in the APP processing pathway that gives rise to the neurotoxic Aβ peptide.

PSEN1 is located on chromosome 14q24.3 and mutations in this gene account for the majority of EOFAD cases, referred to as familial AD type 3. Most *PSEN1* mutations result in complete penetrance, however, a common mutation (Glu318Gly) displays only partial penetrance. Therefore, the mutation is considered a genetic risk factor which predisposes individuals; the incidence rate of the mutation is 8.7% in patients with familial AD (Taddei *et al.,* 2002). A study found that transgenic mice overexpressing a mutant *PSEN1* gene accumulate greater amounts of $A\beta_{42}$ in the brain, highlighting the importance of the gene in the regulation of Aβ in AD (McGowan *et al.,* 2006). *Psen1* knockout mice showed normal α- and β-secretase activity during APP processing, however γ-secretase cleavage of the TM domain did not occur, resulting in a 5-fold decrease in Aβ peptide production (De Strooper *et al.,* 1998). This implicated *PSEN1* as a potential target for therapeutic intervention in AD.

Presenilin 2

PSEN2, located on chromosome 1q31-q42, shares 67% homology with *PSEN1* in its primary structure (Levy-Lahad *et al.,* 1996). Mutations in this gene can lead to rare cause of EOAD type 4 and are thought to be specific to family pedigrees (Ertekin-Taner, 2007). Despite lacking extensive study, distinct allelic variants have been identified in *PSEN2*. The mutation p.N141I is shown to alter APP metabolism by disrupting γ-

cleavage to affect the C-terminus of the Aβ peptide. The result is an increased rate of $A\beta_{42}$ deposition into plaques (Tomita *et al.,* 1997). The Aβ deposition was also found to manifest itself as cerebral amyloid angiopathy in one family (Nochlin *et al.,* 1998). Another mutation p.D439A has been suggested to change the endoproteolytic processing of PSEN2 near the C-terminus of APP (Lleo *et al.,* 2001).

1.7.2 Late-onset Alzheimer's disease

The majority of AD cases are late-onset (LOAD), which is often referred to simply as AD. Hence from hereon, the term AD will be used to describe the late-onset form and any mention of early-onset or familial forms will be referred to as such.

AD is primarily sporadic in nature; however familial cases have been described. This sporadic form is genetically complex in comparison with familial AD in that the heritability is between 50-80% as determined by twin studies (Wilson *et al.,* 2011). Therefore it is likely that environmental factors, such as diet and lifestyle, are also contributing to risk of AD (Gatz *et al.,* 2006). AD is not part of the normal aging process however prevalence increases significantly with age. The risk of developing AD doubles every 5 years after 65, affecting an estimated 1 in 6 people over the age of 80, making this the most significant risk factor (Ferri *et al.,* 2005). Other factors that have been implicated include exposure to pesticides, metals and air pollution, as well as head injury, exercise, hypertension and hypercholesterolemia (Dosunmu *et al.,* 2007; Kivipelto and Solomon, 2006).

Low educational level has also been associated with an increased risk of AD. Epidemiological studies have investigated the correlation between educational background and incidence of AD, however results are inconsistent (Andel *et al.,* 2005; Cobb *et al.,* 1995; Kokmen *et al.,* 1996; Zhou *et al.,* 2006). Nonetheless, this led to theories suggesting that better educated individuals might partake in more mentally stimulating activities throughout their life, such as reading, playing

instruments, having a stimulating job or frequent social interactions. The hypothesis is that greater brain activity develops more cell-to-cell connections.

1.8 Genetics of Alzheimer's disease

AD has a strong genetic component, yet relatively little is understood about the heritability of the disease (Gatz *et al.,* 2006). There appears to be no overlap with EOFAD genes, suggesting a distinct route of pathogenesis in the sporadic form (Tanzi and Bertram, 2005). Despite no mutations currently known to have full penetrance in AD, some genetic polymorphisms are known to act as predisposing factors (Farrer, 1997). AD is therefore a complex disease that does not follow a Mendelian inheritance pattern in that it is caused by a mutation in a single gene, but rather a combination of genetic and environmental factors.

1.8.1 Apolipoprotein E

The *APOE* (Apolipoprotein E) locus was first identified as a genetic risk factor for AD through linkage analysis (Corder *et al.,* 1993). It is located on chromosome 19q13.32 and occurs in three isoforms that confer different levels of risk for AD; ε2, ε3 and ε4 (Corder *et al.,* 1994). The isoforms differ by single amino acid substitutions caused by two variants in exon 4, rs429358 (Cys112Arg) and rs7412 (Arg158Cys), which determine the haplotypes as shown in Table 1.2. The population frequencies for the ε2, ε3 and ε4 alleles are 8.4%, 77.9% and 13.7%, respectively, with ε4 frequency increased to 40% in AD samples (Farrer, 1997). The most common allele is ε3 which is considered to have a neutral effect. The most recognised risk factor for LOAD is the inheritance of the *APOE* ε4 allele. The risk of developing AD for individuals possessing one ε4 allele is increased by 3-fold and this susceptibility increases 12 to 15-fold in individuals that are homozygous for the allele (Bertram *et al.,* 2007). Between 40 and 80% of individuals with AD are heterozygous for *APOE* ε4, this does not however

guarantee that a person will develop AD (Farrer, 1997). The rare ε2 allele confers a protective effect and has been associated with up to 40% decrease in AD risk (Farrer, 1997). It has also been associated with a delayed age of onset in individuals that go on to develop AD (Corder *et al.,* 1994).

Table 1.2. *APOE* haplotypes, summary of *APOE* ε2, ε3 and ε4 allelic status. The *APOE* allelic status is determined by two SNPs, rs429358 and rs7412, which are inherited in different combinations. Adapted from Kim *et al*. (2009).

			rs7412	
			Arg158Cys	
			Major (C)	**Minor (T)**
rs429358	**Cys112Arg**	**Minor (C)**	ε4 (15%)	ε1 (<0.1%)
		Major (T)	ε3 (77%)	ε2 (8%)

APOE is predominantly expressed by astrocytes and microglia in the central nervous system (CNS) and facilitates the transport of lipids to neurons in the brain (Corder *et al.,* 1994). The three isoforms exert differential effects on Aβ metabolism through conformational changes to the APOE structure (Liao *et al.,* 2017). The ε4 isoform has been shown to physically interact with Aβ, disrupting its conformational properties (Holtzman *et al.,* 2000). This causes deficits in Aβ clearance, thereby increasing the aggregation and enhancing plaque formation (Sleegers *et al.,* 2010). Consequently, ε4 carriers exhibit greater plaque and tangle burden than other *APOE* polymorphisms (Liu *et al.,* 2013).

1.9 Genome-wide association studies

Genome-wide association studies, known simply as GWAS, investigate variants or SNPs (single nucleotide polymorphisms) across the entire genome for association with a trait. They have identified hundreds of susceptibility loci implicated in complex diseases since the first study was published in 2005 (Manolio, 2010). A tag SNP represents a group

of SNPs in high linkage disequilibrium (LD) within a genomic region, called a haplotype. LD is explained as the occurrence of two or more SNPs on the same chromosome, commonly inherited together in a population (Bush and Moore, 2012). Presence of the tag SNP confirms presence of other SNPs in the haplotype, without genotyping every SNP in the region. GWAS employs this method to efficiently genotype hundreds of thousands of SNPs across the genome. The true causative variants are likely to be in LD with the associated tag SNP (Rosenthal and Kamboh, 2014). Due to the small effect sizes of these SNPs on the trait, large sample sizes are required to examine an array of common genetic variants, with MAF (minor allele frequency) >5% and evaluate the likelihood of association between variants and phenotypic traits (Clarke *et al.,* 2011). If an allele is significantly more or less frequent in cases than controls, the variant is considered to be associated with the disease. These variants represent susceptibility loci in the human genome that can influence the risk of developing disease (Manolio, 2010). Despite using a non-candidate driven approach, GWAS have generally yielded consistent and reproducible findings (Bertram *et al.,* 2010).

Initial GWAS of LOAD were underpowered due to small sample sizes and therefore failed to detect any signals above genome-wide significance, aside from established risk factor *APOE* (Beecham *et al.,* 2009; Bertram *et al.,* 2008; Carrasquillo *et al.,* 2009; Coon *et al.,* 2007; Grupe *et al.,* 2007; Li, Rowland, *et al.,* 2008; Reiman *et al.,* 2007).

The first two large-scale GWAS were published in 2009, having identified three independent signals associated with AD risk that reached genome-wide significance. The GERAD (Genetics and Environmental Risk in Alzheimer's Disease) consortium found two novel SNPs in *CLU* ($P=1.4x10^{-9}$, OR=0.84 [0.79-0.89]) and *PICALM* ($P=1.9x10^{-8}$, OR=0.85 [0.80-0.90]) using a two-stage approach of independent sample sets to achieve genome-wide significance (Harold *et al.,* 2009). This was also employed by the EADI (European

Alzheimer's Disease Initiative) which replicated the signal in *CLU* ($P=7.5x10^{-9}$, OR=0.86 [0.81-0.90]) and identified a further association in *CR1* ($P=3.5x10^{-9}$, OR=1.21 [1.14-1.29]) (Lambert *et al.,* 2009). Studies aiming to validate these results found the associations to be highly replicable (Carrasquillo *et al.,* 2010; Jun *et al.,* 2010; Seshadri *et al.,* 2010a).

A new signal found to significantly affect AD risk was found in *BIN1* ($P=1.59x10^{-11}$, OR=1.15 [1.11-1.20]) and the associations in *CLU* and *PICALM* were confirmed (Seshadri *et al.,* 2010a). A further five signals were detected in *ABCA7* ($P=5.0x10^{-21}$, OR=1.23 [1.17-1.28]), the *MS4A* locus (*MS4A6A*/*MS4A4E*) ($P=1.2x10^{-16}$, OR=0.91 [0.88-0.93]), *CD2AP* ($P=8.6x10^{-9}$, OR=1.1 [1.07-1.15]), *CD33* ($P=1.6x10^{-9}$, OR=0.91 [0.88-0.93]) and *EPHA1* ($P=6.0x10^{-10}$, OR=0.90 [0.86-0.93]) using a three-stage meta-analysis (Hollingworth *et al.,* 2011). These nine susceptibility loci were later confirmed in a combined study (Naj *et al.,* 2011).

A meta-analysis of the multiple large-scale GWAS datasets, comprising of 17,008 AD case and 37,154 control individuals, revealed 11 new susceptibility loci in addition to *APOE* and the 9 already known that reached genome-wide significance (Lambert *et al.,* 2013). Genes associated with the loci include *CELF1*, *CASS4*, *HLA-DRB1-HLA-DRB5*, *PTK2B*, *SORL1*, *INPP5D*, *NME8*, *FERMT2*, *SLC24A4*, *MEF2C* and *ZCWPW1*. Table 1.3 contains association results of the meta-analysis for all susceptibility loci with their population attributable fraction (PAF) and studies that identified the associations. PAF quantifies the contribution of a risk factor to a disease; it is the proportional reduction in disease risk if exposure was excluded (Naj et al., 2011). Collectively, these genes explain ~60% of AD risk, with *APOE* estimated to contribute 25% and the GWAS genes accounting for a further ~30% of LOAD risk (Ebbert *et al.,* 2014; Medway and Morgan, 2014; Naj *et al.,* 2011). These findings, although consistent in

Table 1.3. Alzheimer's disease GWAS meta-analysis results. Summary of results from Alzheimer's GWAS stage 1, stage 2 and overall meta-analyses for SNPs reaching genome-wide significance after stage 2. The nearest coding gene, genomic position, P-value, odds ratio, population attributable fraction (PAF) and original identifying study are listed. PAF represents the incidence of disease that is not due to exposure. Key: CI, 95% confidence intervals; PAF, population attributable factor (Lambert *et al*. 2013).

Gene	Variant	Genomic Location	P-value	Odds ratio (95% CI)	PAF	Study
ABCA7	rs4147929	19:1063443	1.06E-15	1.15 (1.11-1.19)	2.8	Hollingworth *et al*. (2011)
BIN1	rs6733839	2:127892810	6.43E-44	1.22 (1.18-1.25)	8.1	Seshadri *et al*. (2010)
CASS4	rs7274581	20:55018260	2.46E-08	0.88 (0.84-0.92)	1.1	Lambert *et al*. (2013)
CD2AP	rs10948363	6:47487762	5.20E-11	1.10 (1.07-1.13)	2.3	Hollingworth *et al*. (2011)
CD33	rs3863444	19:51727962	2.97E-08	0.94 (0.91-0.96)	1.8	
CELF1	rs10838725	11:47557871	1.12E-08	1.08 (1.05-1.11)	2.4	Lambert *et al*. (2013)
CLU	rs9331896	8:27467686	2.77E-25	0.86 (0.84-0.89)	5.3	Harold *et al*. (2009)
CR1	rs6656401	1:207692049	5.69E-24	1.18 (1.14-1.22)	3.7	Lambert *et al*. (2009)
EPHA1	rs11771145	7:143110762	1.12E-13	0.90 (0.80-0.93)	3.1	Hollingworth *et al*. (2011)
FERMT2	rs17125944	14:53400629	7.94E-09	1.14 (1.09-1.19)	1.5	Lambert *et al*. (2013)
HLA-DRB1/HLA-DRB5	rs9271192	6:32578530	2.94E-12	1.11 (1.08-1.15)	3.2	
INPP5D	rs35349669	2:234068476	3.17E-08	1.18 (1.05-1.11)	4.6	
MEF2C	rs190982	5:88223420	3.23E-08	0.93 (0.90-0.95)	2.7	
MS4A6A	rs983393	11:59923509	6.14E-16	0.90 (0.87-0.92)	4.2	Hollingworth *et al*. (2011)
NME8	rs2718058	7:37841534	4.76E-09	0.93 (0.90-0.95)	2.9	Lambert *et al*. (2013)
PICALM	rs10792832	11:85867875	9.32E-26	0.87 (0.85-0.89)	5.3	Harold *et al*. (2009)
PTK2B	rs28834970	8:27195121	7.37E-14	1.10 (1.08-1.13)	3.1	Lambert *et al*. (2013)
SLC24A4/RIN3	rs10498633	14:92926952	5.54E-09	0.91 (0.88-0.94)	1.5	
SORL1	rs11218343	11:121435587	9.73E-15	0.77 (0.72-0.82)	1.1	
ZCWPW1	rs1476679	7:100004446	5.58E-10	0.91 (0.89-0.94)	3.2	

Caucasian groups, have been difficult to replicate in other populations (Jun *et al.,* 2010).

To date, GWAS have successfully identified over 20 new susceptibility loci that affect the risk of developing AD. Although they can serve as a powerful tool for elucidating associations between genes and complex traits, they also have their limitations. GWAS typically identify common variants which only exert small effects (Bush and Moore, 2012; Visscher *et al.,* 2012). As a result, the susceptibility loci do not account for much of the 'missing heritability' in complex diseases. The majority of GWAS variants are located in non-coding regions, which can lie hundreds of kilobases from a known causative gene. Additionally these studies cannot specify whether genes are truly causal to the disease (Manolio, 2010). Most GWAS signals are highly replicable, yet the causative functional genetic effect remains to be identified in the majority of instances (Bush and Moore, 2012). Importantly however, they have led to the identification of new biological pathways linked to AD pathogenesis, including synaptic signalling and endocytosis, cholesterol metabolism and inflammation (Morgan, 2011; Tosto and Reitz, 2013).

1.10 Susceptibility genes and pathways

The amyloid cascade hypothesis has been the dominant theory of AD pathogenesis for more than 2 decades. However, the failures of several drug treatments targeting Aβ pathologies suggests that Aβ may not be the primary causative factor in AD. This is further supported by a lack of correlation between the numbers of plaques and severity of the disease (Morris *et al.,* 2014). Although the role of aberrant APP processing is well understood in EOFAD, current data distracts support from the theory that toxic $A\beta_{42}$ is the primary cause of all AD.

The new genetic risk factors identified by GWAS were found to be associated with pathways already hypothesised to underlie AD such as APP processing, as well as implicating some new pathways in the

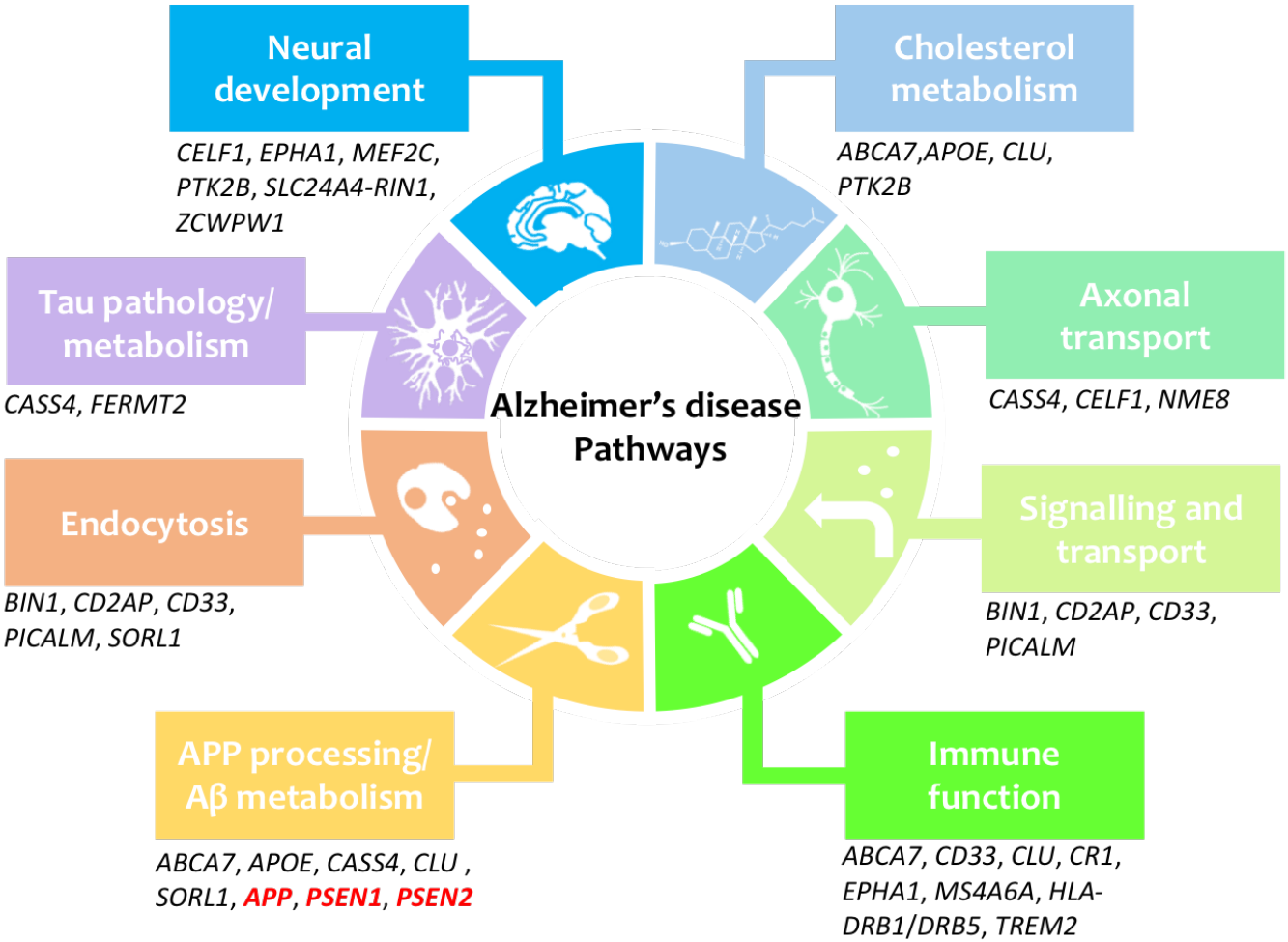


Figure 1.6. Biological pathways implicated in Alzheimer's disease. New genes and disease pathways implicated from GWAS and NGS. There is considerable overlap of risk genes between pathways with several of the identified risk factors being involved in multiple pathways. Several genes are also involved in the amyloid pathway, suggesting the Aβ may have still play a role in sporadic AD. Adapted from Morgan (2011) and Tosto and Reitz (2013).

disease pathogenesis (Reitz, 2014). Figure 1.6 illustrates the associated pathways identified in GWAS. Broadly these pathways can be grouped into the following; Aβ metabolism, tau pathology, axonal transport, endocytosis, cholesterol metabolism, neural development, immune function, signalling and extracellular transport. There is considerable overlap with several genes being involved in multiple pathways, highlighting potential mechanisms by which the disease could manifest (Tosto and Reitz, 2013).

In addition to *APOE*, *CLU* and *ABCA7*, newly identified risk genes *CASS4* and *SORL1 also* appear to be involved in APP processing and Aβ metabolism. While this does not demonstrate that changes in Aβ metabolism are the primary cause of AD, this indicates that the regulation of the Aβ pathway remains a core constituent in AD pathogenesis. *ABCA7*, *APOE* and *CLU* variants are also crucial to the cholesterol metabolism pathway, which has been linked to the APP processing pathway (Morgan, 2011; Tosto and Reitz, 2013).

Similarly, GWAS identified several genes encoding proteins which function in axonal transport and tau metabolism. Whilst the involvement of tauopathy and breakdown of axonal transport have been well characterised in AD, the identification of these new loci is key to understanding the complex pathogenesis of the disease. *CASS4* and *FERMT2* are associated with tau metabolism, with CASS4 being involved in both tau metabolism and axonal transport (Reitz, 2014).

The genes linked to immune response and inflammation pathways include *HLA-DRB1-HLA-DRB5*, *INPP5D* and *MEF2C*, with *CLU*, *CR1*, *ABCA7*, *EPHA1*, *MS4A* and *CD33* (Hollingworth *et al.,* 2011). This pathway has recently gained much popularity due to the identification of novel risk genes that provide evidence for the neuroinflammation theory of AD.

PICALM, *BIN1*, *CD2AP* and *CD33* play important roles in synaptic signalling and endocytosis (Morgan, 2011; Seshadri *et al.,* 2010a).

SORL1 has been associated with both endocytosis and Aβ metabolism (Lambert *et al.,* 2013). Endocytosis is essential for proper transport of Aβ around the cell, for example to lysosomes to ensure toxic products are removed. Dysfunction in this pathway could affect the proper removal of misfolded proteins, contributing to the build-up of such products.

New pathways implicated by GWAS data also include hippocampal synapse function (*MEF2C* and *PTK2B*), axonal transport (*CASS4*, *CELF1* and *NME8*) and microglial cell function (*INPP5D*) (Reitz, 2014).

Identification of novel pathways underlying AD pathogenesis provides the potential for new therapeutic targets; however, these genes would need to be assessed for functionality to determine their roles in AD (Bettens *et al.,* 2013). Despite the identification of new disease pathways, the functional relevance of many of the genes also remains elusive. These genes are described in further detail in Chapter 5.

The graph in Figure 1.7 represents the spectrum of genetic variants believed to confer risk in common diseases, with their relative allele frequency against effect size. Current NGS technology has the capability to detect variants within the central region shown in the figure, ranging from common frequency variants with small effects like GWAS and more importantly, lower frequency variants with a high effect size.

The heritable component of EOFAD is as high as 92-100% (Guerreiro and Hardy, 2014). It has been established that EOFAD is caused by extremely rare but highly penetrant mutations in one of three genes *APP*, *PSEN1* or *PSEN2*. Such rare variants are difficult to detect in common diseases as exceedingly large sample sizes in the tens of thousands would be required to detect significant associations.

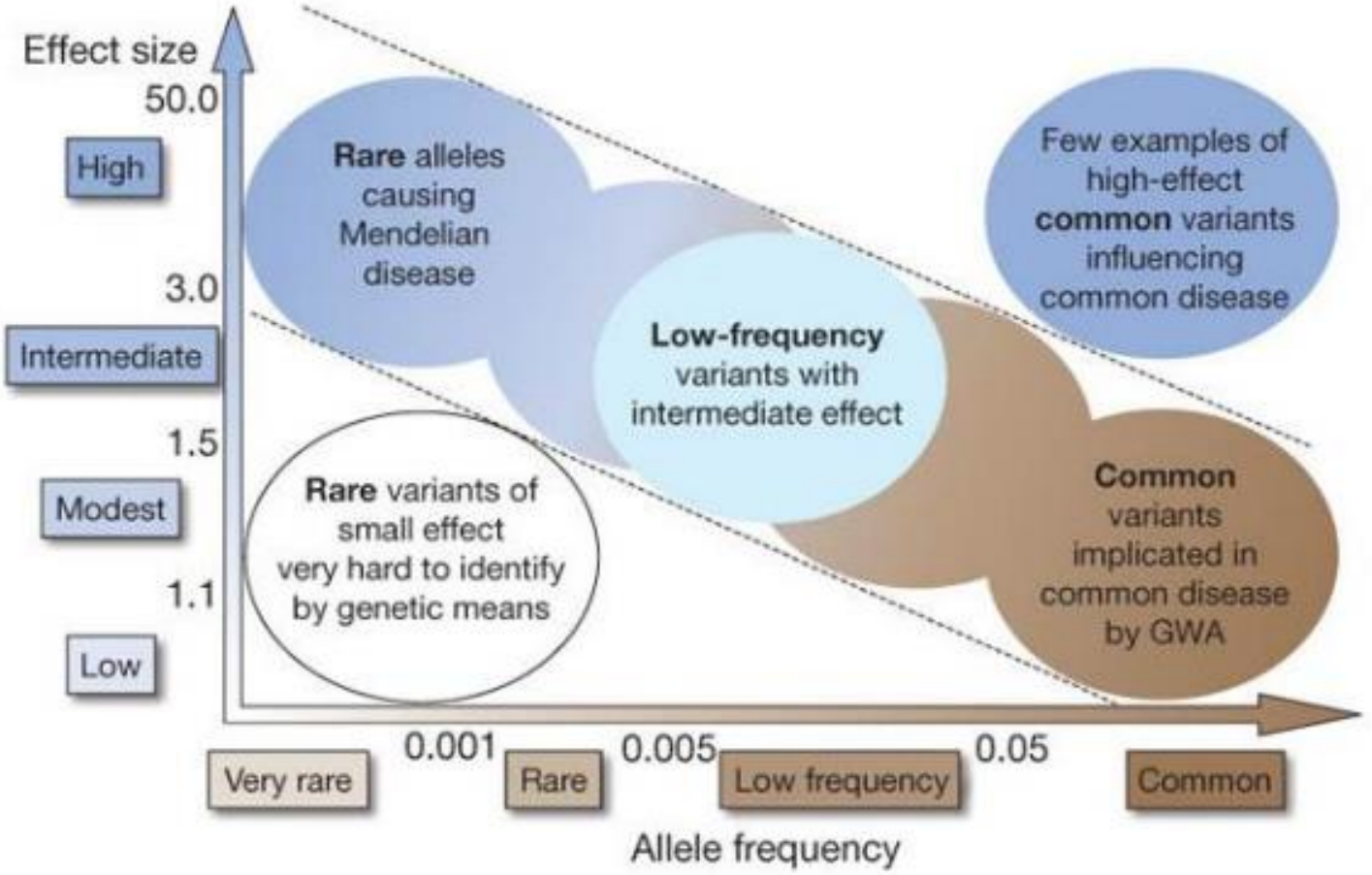


Figure 1.7. Variants associated with complex disease risk. The spectrum of genetic variants thought to be involved in the aetiology of complex diseases. Variants are plotted by effect size in terms of odds ratio against allele frequency. Current sequencing technologies are able to identify rarer variants within the dotted lines. Figure taken from (Manolio, 2010).

1.11 New biological pathways

Immune system function

Emerging evidence from both GWAS and NGS have suggested a causal role for the inflammatory pathway in AD pathogenesis. Microglia play a significant role in inflammation and neuroprotection. As macrophages, microglia remove cellular debris, damaged neurons and misfolded proteins (Bird, 2013). It is also suggested that microglia phagocytose Aβ aggregates, and thus impairments in their function will lead to the accumulation of plaques (Heneka *et al.,* 2015). The 'microglial dysfunction hypothesis' suggests increased microglial senescence due to ageing results in reduced neuroprotection which correlates with increased AD risk (Streit and Xue, 2012). It has been posited that damage from Aβ pathology can also increase microglial senescence. It has also been established that Aβ deposition is known to activate microglia, releasing inflammatory cytokines which cause neuroinflammation (Heneka *et al.,* 2015). The Aβ plaques block cell signalling at synapses which can lead to increased toxicity from neurotransmitter build up. Indeed both microglia and astrocytes are activated in response to excess neurotransmitters and can act to reduce excitotoxicity (Streit and Xue, 2012).

Endocytosis and synaptic function

Neurons are heavily dependent on complex molecular trafficking and endocytosis of proteins which are critical to neuronal and synaptic functions. These can be grouped together in the endocytic pathway, which is critical to normal APP processing as APP, Aβ and APOE are internalised through the endo-lysosomal pathway (Chapuis *et al.,* 2013). Dysregulation of this pathway in AD can affect the production and clearance of Aβ in the cell. Furthermore, the build-up of waste proteins indicates that autophagic and lysosomal processes may be defective in AD (Moreau *et al.,* 2014). Synaptic activity is also disrupted in AD and has been shown to alter Aβ generation by influencing the endocytosis

of APP (Cirrito *et al.,* 2008). Tau is also a key component of the neuronal transport system; hyperphosphorylation of this protein leads to the collapse of microtubules due to NFTs. This results in the disruption of molecule trafficking along the axon. While it is unclear whether endocytic dysfunction is directly associated with tau pathology, a recent study suggests that AD pathology is influenced by GWAS genes within this pathway (Kimura and Yanagisawa, 2017).

Cholesterol metabolism

The exact mechanisms through which cholesterol affects Aβ production and metabolism remains elusive; however, *in vivo* and *in vitro* studies suggest an essential role in AD pathogenesis (Gamba *et al.,* 2012). The brain has the highest concentration of cholesterol, where it constitutes a major component of neuronal cell membranes and is a main determinant of membrane fluidity. In this way cholesterol is essential for proper neuronal functioning (Orth and Bellosta, 2012; Zhang and Liu, 2015). Increased cholesterol levels in the lipid bilayers have been reported to directly bind APP, promoting the conversion of Aβ from soluble monomers to form neurotoxic aggregates via the amyloidogenic pathway (Beel *et al.,* 2010; Kakio *et al.,* 2001). Extracellular cholesterol is also known to accumulate within senile plaques and NFTs of AD patients. APOE, Aβ as well as cholesterol co-localise within the core of the fibrillary plaques, indicating that both APOE and cholesterol may be essential for plaque formation (Burns *et al.,* 2003). Concurrently, elevated levels of cholesterol may contribute to early amyloidogenesis and development of AD (Burns *et al.,* 2006). Equally, decreased cellular cholesterol levels are observed in the non-amyloidogenic pathway, thus it suggests that regulating cholesterol levels could influence AD pathology (Gamba *et al.,* 2012). This link was further highlighted by the GWAS, which implicated genes involved in cholesterol metabolism, including *ABCA7*, *CLU* and *PTK2B*.

1.12 NGS in complex disease

GWAS examines the "Common Disease – Common Variant" hypothesis, which argues that the accumulation of multiple high frequency variants with relatively low effects are the major contributors to susceptibility in complex diseases (Reich and Lander, 2001). However, this does not explain much of 'missing heritability' observed in AD. Conversely the "Common Disease – Rare Variant" hypothesis states that multiple rare mutations with greater penetrance, within the same gene or several genes, are more likely to cause a disease phenotype (Pabinger *et al.,* 2014). It is plausible that a combination of rare and common variants will determine individual disease susceptibility (Schork *et al.,* 2009). GWAS have identified a number of susceptibility loci, however the causative variants remain unknown.

Finer mapping of these loci using high-throughput next-generation sequencing (NGS) technology could elucidate the true causative variants driving the association signal and also detect rarer variants previously missed by GWAS (Visscher *et al.,* 2012). The previous gold standard technology was Sanger sequencing, a chain-termination method that uses modified dideoxyribonucleotidetriphosphates (ddNTPs) to terminate elongation of DNA strands (Sanger *et al.,* 1977). A surge in demand for low-cost sequencing has driven development in the field of genomics towards faster, more accurate and substantially cheaper technology (Kircher and Kelso, 2010). The advent of massively parallel sequencing has facilitated NGS to read millions of DNA fragments from a single sample concurrently (Grada and Weinbrecht, 2013). Despite the low costs of NGS in comparison to Sanger sequencing, whole-genome sequencing (WGS) is still costly and computationally expensive in terms of data storage and processing (Goh and Choi, 2012). To attain a compromise between cost-effectiveness and genome-wide comprehensiveness, whole-exome sequencing (WES) is considered a suitable alternative (Ng *et al.,* 2010). This represents a targeted approach that only captures protein-coding

regions, which account for approximately 1% of the human genome yet harbour an estimated 85% of known pathogenic mutations associated with disease (Choi *et al.,* 2009).

Increasing evidence supports the “Common Disease – Rare Variants” hypothesis that low frequency variants could have greater effect sizes and therefore contribute significantly to the heritable component of AD (Rivas *et al.,* 2011). The single base resolution of NGS has allowed the discovery of rare mutations in both Mendelian and complex forms of disease (Bamshad *et al.,* 2011). Rare genetic variants found in a small number of individuals are likely to contribute to diseases such as AD (Ng *et al.,* 2010). Therefore, by identifying the genetic variants of individuals with AD, WES has the potential to uncover novel rare variants associated with disease risk.

Rare variants associated with altered LOAD risk have been identified in the genes *UNC5C*, *AKAP9*, *ADAM10*, *APP* and *TREM2*. Coding variants in *UNC5C* and *AKAP9* were associated with increased risk of developing AD (Wetzel-Smith *et al*, 2014). *ADAM10* mutations were shown to increase plaque load and Aβ levels in Tg2576 mice (Suh *et al*, 2013). A rare APP mutation rs6370847 (A673T) discovered through WGS of the Icelandic population confers protection against LOAD. It was associated with decreased cognitive decline in healthy controls (Jonsson *et al.,* 2012), likely by reducing APP cleavage via the amyloidogenic pathway to decrease Aβ production.

The most significant finding from NGS in recent years has been the identification of the rs75932628 variant causing a rare substitution mutation R47H (Arg>His) in TREM2 (Guerreiro, Wojtas, *et al.,* 2013; Jonsson *et al.,* 2013). The variant was extremely rare with MAF of 0.2%, yet it was only observed in AD cases with an effect size of 3-fold, similar to *APOE* (Abduljaleel *et al.,* 2014). TREM2 has an anti-inflammatory function expressed within microglial cells. The microglia themselves are capable of phagocytic removal of extracellular debris thus preventing the build-up of misfolded proteins such as Aβ which can

become toxic (Hickman and El Khoury, 2014). Other variants in the *TREM* cluster of genes have also been implicated in altered LOAD risk. Missense variant rs3747743 (p.S144G) in *TREML2*, independent of the *TREM2* finding, was correlated with protection against AD (Benitez *et al.,* 2014)

WES has also identified population-specific rare variants in the genes associated with LOAD risk exhibiting greater effect sizes than the GWAS tag SNPs (Li, Rowland, *et al.,* 2008; Vardarajan *et al.,* 2015). They have also discovered rare variants in the risk genes that were not detected by GWAS, highlighting the importance of sequencing approaches to uncover rarer variants associated with disease (Tosto and Reitz, 2013). Rare variants in *CLU* showed greater association with AD risk than the common variant identified by GWAS (Bettens *et al.,* 2012). These findings also suggest that both rare and common risk variants within a gene could play a role in AD susceptibility (Pottier *et al.,* 2012). *SORL1* was found to carry mutations that cause EOAD with autosomal dominant inheritance (Pottier *et al.,* 2012). It is a neuronal receptor involved in vesicle trafficking within the cell; it directs APP to endocytic pathways for recycling which affects Aβ generation (Rogaeva *et al.,* 2007). Its role in APP sorting could explain the impact of *SORL1* mutations on EOAD, as the protein has a protective function against Aβ secretion (Nicolas *et al.,* 2016).

Addressing some of the limitations faced by both GWAS and NGS, a recent study used an exome microarray to investigate rare and common coding variants in association with LOAD (Sims *et al.,* 2017). They observed three new genome-wide significant associations of rare variants in the genes *PLCG2*, *ABI3* and *TREM2*. These genes are highly expressed in microglia and enriched for immune function, which provides further evidence to implicate the microglial-mediated immune response in the pathogenesis of AD.

Figure 1.8 illustrates the current state of AD genetics knowledge, with all risk genes shown by their relative frequency and effect in AD. The

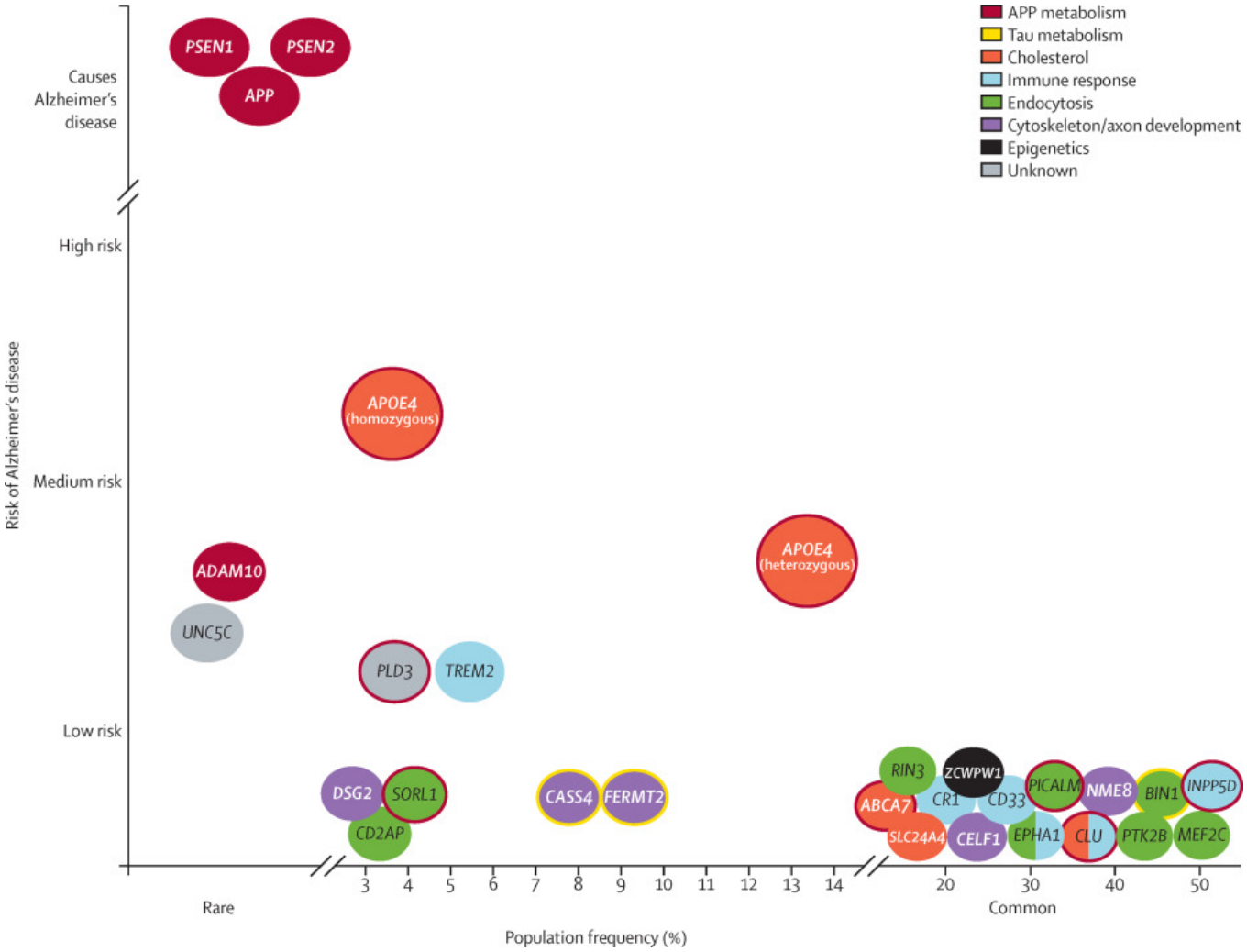


Figure 1.8. Genes containing AD risk variants categorised by frequency and effect. The spectrum of known genetic variants associated with AD. Variants are plotted by the observed effect size against allele frequency in the population. Variants in FAD genes (top left) are extremely rare but highly penetrant. Whereas GWAS variants (bottom right) are common with small effects. Figure obtained from (Karch and Goate, 2015).

Mendelian genes *APP*, *PSEN1* and *PSEN2* represent the rare, highly penetrant causative mutations identified in EOFAD. The cluster of genes in the bottom right are the risk genes identified by GWAS, which are common in frequency yet exert small effects. *APOE* is the only common variant with higher risk. Several risk genes containing rare variants have been identified with varying effect sizes, however, using NGS we now want to uncover rarer variants with moderate to high risk of AD.

Sequencing studies have uncovered high risk variants within these genes, providing evidence for the contribution of rare variants to AD risk. Therefore, NGS will be a valuable resource for ongoing and future studies that is likely to unravel a significant part of the hidden heritability of AD (Tosto and Reitz, 2013). Studies are now beginning to integrate other factors such as imaging data, cognitive traits, gene expression and methylation with genetic mutations. BDR (Brains for Dementia Research) collects comprehensive neurocognitive and clinical information on patients which provides in-depth phenotyping. The application of WES on such well-characterised samples may also prove very useful in delineating the genetic basis of neurodegenerative disease.

1.13 Project Aims

Following on from GWAS, which addressed common variation associated with disease, we are beginning to utilise NGS and custom-designed genotyping arrays to explore the contribution of rare variants. This approach highlighted the role of *TREM2* in AD risk and emerging NGS data is set to generate more genes of interest. This project aims to investigate genetic variation in Alzheimer's disease using whole-exome sequencing of the BDR cohort. Implementation of a robust DNA sequencing analysis pipeline was used to generate high quality variant call data **(Chapter 3)**. Exonic variants were assessed for association with AD, individually and in burden within genes to discern potentially causative variation **(Chapter 4).** Furthermore, PRS were generated to explore the ability to classify individual risk using genetic profiles and neuropathological information provided by the BDR resource. Associated genes of interest were further explored to predict potential interactions with known AD risk genes using *in silico* pathway analysis tools. AD risk genes were then assessed more closely for associated variants individually and in burden, to increase the power for detection **(Chapter 5)**. Genes linked to other neurodegenerative diseases were also included to consider potential genetic pleiotropy. The ultimate aim is to generate and prioritise potential candidates for functional assessment in future work.

Chapter 2: General Methods

2.1. Lab-based methods

This section details the general experimental methods which form the basis of any work carried out in this thesis. Where a technique is only relevant to a particular chapter it is described in the methods section of that chapter.

2.1.1. DNA extraction

Frozen post-mortem brain tissue was obtained from the Brains for Dementia Research network. DNA was extracted using a phenol chloroform method. All reagents were obtained from QIAGEN except the phenol chloroform which was from Sigma Aldrich. Approximately 100mg fragment of frozen brain tissue was used for the extraction. The sample was kept cold on dry ice and manually homogenised with a scalpel. The homogenised tissue was added to a 1.5ml Eppendorf with 500µl lysis buffer, 100µl of 20mg/ml Proteinase K and 10µl of 100mg/ml RNAse A and incubated overnight for 16-18 hours at 50°C in a GrantBio thermoshaker with gentle shaking.

Following overnight incubation, 500µl of cold phenol chloroform was added to the sample and mixed thoroughly before centrifuging. All centrifugation steps were performed in an Abnet PRISM benchtop microcentrifuge at full speed (15,900g) for 5 minutes unless otherwise stated. Phenol chloroform separates the homogenised solution into three distinct phases; the aqueous layer contains the DNA; a white interphase layer separates the proteins and the bottom organic phase contains the cellular debris. The top aqueous was carefully pipetted into a 1.5ml Phase Lock gel (PLG) tube (5PRIME). Another 500µl of phenol chloroform was added to the PLG tube for a second phenol chloroform step and mixed until it formed a clear homogenised solution before centrifuging at full speed for 5 minutes. The gel contained in the PLG tube trapped the interphase and organic layers beneath the gel,

allowing the aqueous layer to be easily transferred to a clean 1.5ml Eppendorf.

Then 50µl of 3M sodium acetate at pH 5.2 was added and the tube inverted before adding 500µl of chilled 100% ethanol. The tube was inverted several times to aid precipitation of the DNA from the mixture. The sample was centrifuged at full speed for 15 minutes to form a DNA pellet. The ethanol was decanted so as not to dislodge the pellet and 500µl of 75% ethanol was added to wash the DNA. Another centrifugation step was performed at full speed for 15 minutes and the ethanol removed. The DNA pellet was air dried at room temperature for 30 minutes to ensure that all remaining ethanol had evaporated as this can inhibit downstream applications. 100µl Tris-EDTA elution buffer was added before incubation at 50°C with shaking for 40 minutes until the pellet had fully resuspended. The DNA was stored at -20°C for up to 2 weeks or -80°C for longer periods of time.

2.1.2. DNA quality control

For quality check, samples were diluted 1 in 10 with nuclease-free water to ensure concentrations were within the detection range for the spectrophotometer. All extracted DNA samples were quantified on a Nanodrop spectrophotometer 2200 (Thermo Scientific) using a Quant-iT dsDNA Broad Range Assay Kit (Life Technologies) according to manufacturer's protocol. The Nanodrop 2200 uses PicoGreen fluorescence to measure the concentration of double stranded DNA present in the sample rather than all DNA, providing a more accurate reading. Samples were quantified in triplicate to ensure readings were reliable. The minimum concentration required for sequencing was 10µg.

DNA integrity was measured on the BioAnalyser TapeStation using 1µl of sample mixed with 10µl buffer in PCR tubes and loaded into the machine with a genomic DNA screentape. The TapeStation outputs a DNA Integrity Number (DIN) score which provided a quantitative measure of DNA quality. Samples were required to have a DIN score

≥6.0 and appear as a clean band on the gel. Dilutions were repeated for any sample below the DIN threshold. If any smearing was observed the sample was re-extracted if tissue was available.

2.1.3. Primer design

Primers were designed in Primer 3 v4.0.0 (http://bioinfo.ut.ee/primer3/) using reference sequences obtained from the NCBI Genome Browser. They were designed to ensure approximately 150 base pairs were included both upstream and downstream of the variant. All primers were checked for SNPs within the binding sites using SNPCheck (https://secure.ngrl.org.uk/SNPCheck/snpcheck.htm). Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) was then used to identify potential non-specific binding by the primers.

2.1.4. Polymerase Chain Reaction

Polymerase chain reaction (PCR) was performed to amplify DNA fragments. All reactions consisted of 1X Roche PCR buffer (Roche Diagnostics, Basel, Switzerland), 0.2mM dNTPs (Thermo Scientific), 1pmol/µl of each forward and reverse primer (Invitrogen), 0.2µl of 5U/µl Taq Polymerase (Roche Diagnostics) and 10-20ng of DNA made up to 20µl with molecular grade dH_2O. A no-template control (NTC) without DNA was also included. All PCRs were performed in a Veriti 96 Well Thermal Cycler (Applied Biosystems, Carlsberg, CA, USA). The standard thermal cycling program included an initial denaturation step at 94°C for 2 minutes, followed by 35 cycles of 94°C for 30 seconds, primer-specific annealing temperature for 30 seconds and 72°C for 45 seconds, with a final extension step at 72°C for 10 minutes.

PCR products were examined on a 1% agarose gel containing 0.5µl/ml ethidium bromide in 1X TAE buffer, alongside a 1kb DNA ladder (Thermo Scientific) at 80V for approximately 30 minutes. Electrophoresis gels were visualised using a UV transilluminator to ensure only one product was present and of the expected size.

2.1.5. Sanger Sequencing

PCR products were enzymatically purified with ExoSAP-IT (Affymetrix, Santa Clara, CA, USA) prior to sequencing. Exonuclease I hydrolyses primers and Shrimp Alkaline Phosphatase dephosphorylates excess dNTPs. The reaction consisting of 5µl of PCR product and 2µl of ExoSAP-IT, was incubated in the thermal cycler at 37°C for 15 minutes and 80°C for 15 minutes to inactivate the enzymes. Treated PCR products were sequenced using the Sanger di-deoxy method. Sanger sequencing remains the gold standard for validation of DNA variants.

Each sequencing reaction was made up of 2µl of cleaned PCR product, 1X BigDye Terminator Sequencing Buffer (Applied Biosystems), 2µl BigDye Terminator v3.1, 5pmol/µl forward primer and molecular grade dH_2O up to 10µl. A Veriti 96 Well Thermal Cycler was used to perform the thermal cycling, which consisted of 25 cycles of denaturation at 96°C for 30 seconds, 50°C for 15 seconds and 60°C for 4 minutes. Sequencing products were then purified by running through Performa DTR Gel Filtration cartridges (Edge Biosystems) that were centrifuged at 1767g for 3 minutes prior to adding PCR products. The cartridges were centrifuged again at the same speed for 3 minutes to remove excess BigDye and dNTPs from the eluate, which was dried at 90°C. The products were sequenced by capillary electrophoresis on an ABI 3130 Genetic Analyser (Applied Biosystems) by the Molecular Diagnostics department.

Electropherograms were visualised in Chromas Lite (http://technelysium.com.au/wp/) and raw sequence data was aligned to the reference sequence using the Clustal Omega online tool (http://www.ebi.ac.uk/Tools/msa/clustalo/). Aligned sequences were examined to determine genotypes for variants of interest.

2.1.6. TaqMan assay for *APOE* genotyping

Genotyping for *APOE* SNPs rs429358 and rs7412 was performed manually using TaqMan assays (ABI, Foster City, CA, USA). The SNPs

are located in GC-rich regions, making them difficult to genotype on platforms such as the NeuroX chip (Jacobsen *et al.,* 2002). SNP specific primers and probes were designed by ABI. All equipment was sterilised by irradiating under UV light for a minimum of 20 minutes prior to use. Light sensitive reagents – assay mixes and TaqMan master mix were thawed at room temperature under foil to prevent degradation. All procedures were performed in a sterile laminar flow hood.

The Mx3000P real-time quantitative PCR (qPCR) machine was primed by switching the lamp on to warm up at least 20 minutes before use. Each genotyping reaction contained 3µl ddH_2O, 4.5µl TaqMan master mix, 0.5µl assay mix (1-rs429358, 2-rs7412) and 2µl of DNA at an approximate concentration of 20ng/µl. Reagent volumes were multiplied by the total number of reactions per assay to calculate required volumes for the TaqMan master mixes. NTCs and Sanger confirmed positive controls of known wildtype and mutant homozygous and heterozygous genotypes were included on each plate. A 96 well optic PCR plate was loaded with 8µl of master mix and 2µl of DNA sample, positive control or nuclease free water for NTCs. Plates were sealed with optical caps and vortexed to ensure reagents were mixed well. The plate was placed in the qPCR machine, plate plan set up as required and thermal cycle set to run for 10 minutes at 95°C, then 60 cycles of 92°C for 15 seconds and 60°C for 1 minute.

The correct alleles for each SNP were determined by comparing HEX and FAM fluorescence curves against the known positive controls. HEX represents the C allele and FAM is the T allele, therefore if only one signal is present the sample is homozygous for that allele (CC or TT) and if both are present it is heterozygous (TC). Genotype calls were defined manually from the fluorescence curves and used to calculate haplotypes to determine *APOE* status of samples.

2.2. Bioinformatics file formats

Several file formats have been developed to store and manipulate NGS data and are now commonly used by the majority of NGS analysis tools.

2.2.1. FASTQ files

FASTQ files stores raw sequence data from Illumina reads in a text-based format with Phred quality scores per individual nucleotide. Reads are given a unique identifier and header as shown in Figure 2.1. The nucleotide sequence for each read is given with a quality score in Sanger FASTQ format. These scores are encoded as ASCII characters which represent the quality of each base. Paired-end reads are stored in separate files for forward and reverse reads, with reads in exactly the same order.

2.2.2. SAM and BAM Files

The majority of alignment programs output aligned FASTQ data in Sequence Alignment/Map (SAM) format. This is a tab delimited text file with a header containing details about the reference genome and an alignment section which includes the aligned read sequences, mapping position and quality scores, orientation and position of the paired mate read and other information. The SAM file stores a large amount of information that can be difficult to store. Compression of the file creates a Binary Sequence Alignment/Map (BAM), which is significantly smaller in size and can be indexed to allow quick retrieval of alignment regions.

2.2.3. VCF files

The standard file for storing variant information from NGS sequencing studies is variant call format (VCF), developed by 1000 Genomes (Danecek *et al.,* 2011). It is a text file containing a header with the VCF version and any filters applied to the data. Each row in the file represents a single variant and includes related information such as ID,

Header	@MG00HS20:797:HH5TGBCXX:1:1101:1500:2068 1:N:0:GCCACATA
Sequence	AGGGTTGCCGTGCCGGATGTCCTACGCGGAACCGGCGCTTGCCAAGGCTGGGGCGGTAGTGG
Additional Information	+
Quality Score	IIHIHHHHIIIIIIHIIDHIHII1DHHHIHHHIHIIDH?GHHIHIHHGHHIIIIIHDHHH?

Figure 2.1. Each FASTQ read entry comprises four lines of information. The header begins with "@" and includes a sequence identifier and details about read group. Raw sequence base calls are given with optional additional information. The final row contains the quality value assigned for each individual base call. The quality scores are represented by ASCII characters 33-126, which correspond to Phred scores of 0-93.

Chromosomes

```
@SQ   SN:chrMT    LN:16571
@SQ   SN:chr1     LN:249250621
@SQ   SN:chr22    LN:51304566
@SQ   SN:chrX     LN:155270560
@SQ   SN:chrY     LN:59373566
```

Alignment Information

```
@PG   ID:bwa      PN:bwa  VN:0.7.10-r789  CL:bwa mem -t 128 hg19.fasta read1.fastq read2.fastq
```

Read Information

```
MG00HS20:797:HH5TGBCXX:1:110:15:268 99    chr3  52472000    60    101M  =     52472049    150
```

Sequence

```
AGGGTTGCCGTGCCGGATGTCCTACGCGGAACCGGCGCTTGCCAAGGCTGGGGCGGTAGTGG
```

Quality Score

```
IIHIHHHHIIIIIIHIIDHIHII1DHHHIHHHIHIIDH?GHHIHIHHGHHIIIIIHDHHH?
```

1	QNAME	Individual read identifier
2	FLAG	Indicates that the read is mapped as the first in a pair
3	RNAME	Chromosome number
4	POS	Chromosomal position according to reference
5	MAPQ	Read mapping quality in Phred scaling
6	CIGAR	All 100bp matched
7	RNEXT	Consecutive reads mapped to the same reference
8	PNEXT	Position of next read in relation to this read
9	TLEN	Insert length (negative for mate pair reverse reads)
10	SEQ	Base calls
11	QUAL	Quality scores in ASCII coding

Figure 2.2. The SAM file header lists chromosomes 1-22, X, Y and MT and their sequence length and alignment information. Technical information such as alignment program name (PN), version number (VN) and command line (CL) are included. The read information, sequence and quality score are comprised of the fields in the table. These 11 fields are required for each aligned read and its mate-pair.

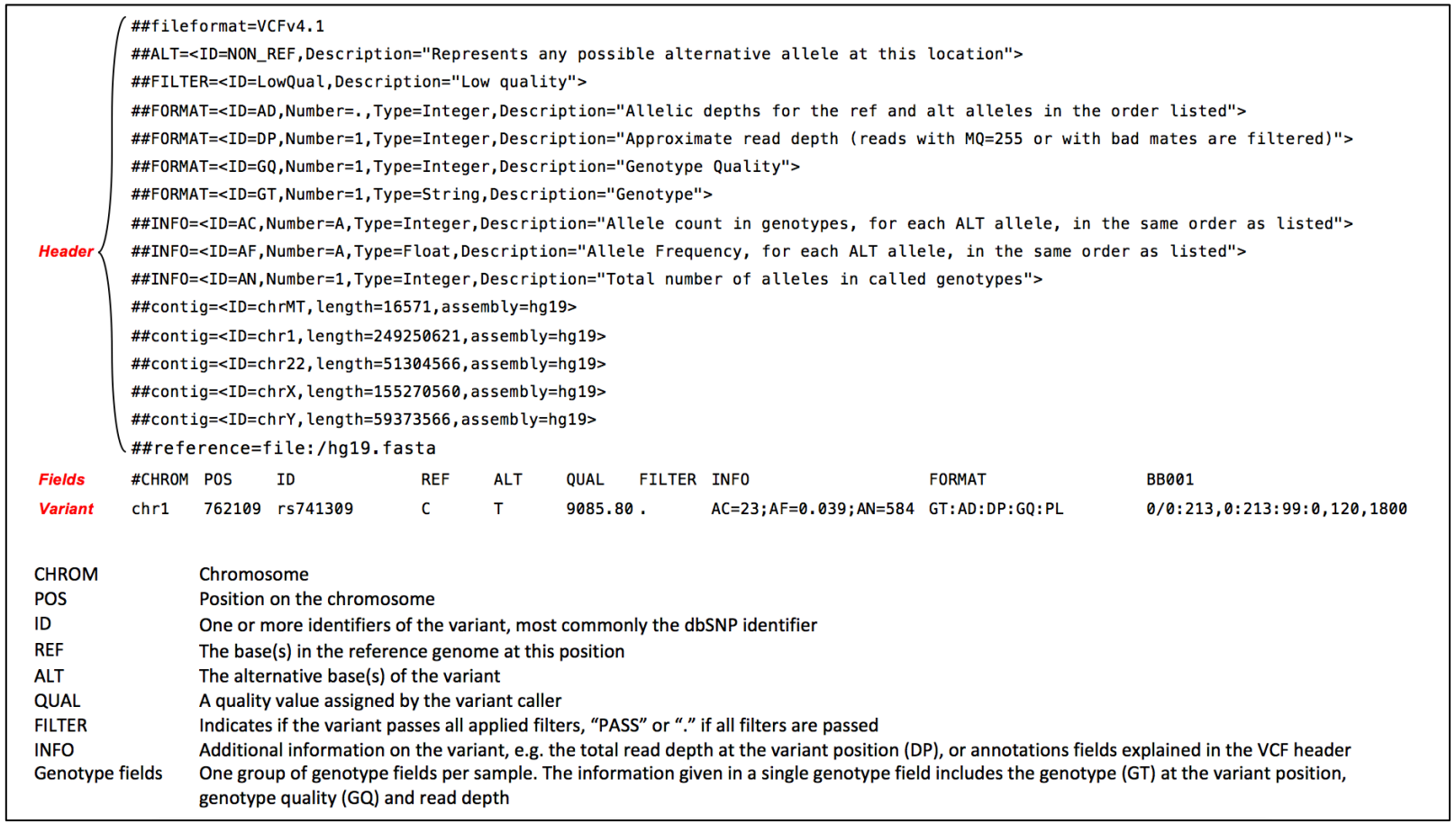


Figure 2.3. The VCF file header contains important metrics on VCF version. Information about quality filters applied to the data, variant information fields, chromosome contigs and link to the reference genome file are included. This is followed by variant meta information with one line per variant. The fields are shown in the table and additional information should be explained in the header. Genotype and quality value for each sample are listed, with an example shown for two samples. The information provided for sample BB001 contains the genotype (GT) as 0/0 or homozygous wildtype, allele depth (AD), depth of coverage (DP), genotype quality scores (GQ) and a Phred-like score (PL).

(a) MAP file

Chrom	Variant	Distance	Position
1	rs540936498	0	762109
1	rs41285790	0	865628
1	rs139570490	0	865738
1	rs774532669	0	866344
1	rs143282473	0	871159

(b) PED file

FID, IID		PID, MID, Sex, Phenotype				Genotypes					
BB001	BB001	0	0	1	1	CC	GG	AA	CC	GG	CC
BB003	BB003	0	0	1	2	CT	GG	AA	CG	GT	CT
BB004	BB004	0	0	1	1	CC	GA	AA	CC	GT	CT
BB005	BB005	0	0	2	2	CC	AA	AA	CG	GT	CC

(c) Covariates file

				Covariates					
FID	IID	Sex	Phenotype	AAR	AAD	APOE_E4_alleles	APOE_E2_alleles	Braak_Tangle_Stage	CERAD
BB001	BB001	1	1	77	78	0	0	2	0
BB003	BB003	1	2	94	95	0	0	3	1
BB004	BB004	1	1	91	92	1	0	1	0
BB005	BB005	2	2	84	87	1	0	4	3

Figure 2.4. MAP, PED and covariates file formats used by PLINK. MAP and PED files do not include a header whereas the covariates file requires a header as shown. (a) MAP file contains all the variants in the data with their ID and chromosomal location to identify them and genetic distance in Morgans. (b) PED file lists all individuals in the dataset and their genotypes for each variant in the MAP file. It also includes Paternal (PID) and Maternal ID (MID) to link families if provided, sex and phenotype. (c) Covariates file containing information about covariates to adjust for in analyses. Family ID (FID) and individual ID (IID) are used to identify each sample. The data in these files requires binary or numerical coding of variable factors. Sex is coded as '1' for male and '2' for female. Phenotype is '1' for controls and '2' for AD cases, with '-9' or '0' denoting different or missing information. Age at recruitment (AAR) and age at death (AAD) were continuous variables, used as a substitute to correct for age. Number of ε4 and ε2 alleles are quantitative and range from 0 to 2 for homozygous samples. Braak tangle stage denotes the severity of tau tangles observed ranging between 0 and 6. CERAD score 0 to 4 determines neurocognitive impairment, with higher scores indicating more severe impairment.

chromosomal location, annotations and assigned quality values, as shown in Figure 2.3.

2.2.4. PLINK files

PLINK v1.07 (Purcell *et al.,* 2007) was used for association analyses, therefore files were converted from VCF to PLINK format. MAP and PED files are the standard file types as shown in Figure 2.4, along with a covariates file. The MAP contains variant information for all genotyped markers, their position and genetic distance. PED files store sample information such as family, individual, maternal and paternal identifiers if available, sex and phenotype encoded with binary values and genotypes for each marker in the MAP file. These files can be encoded in binary format to reduce file size and processing speed, known as BED, BIM and FAM. Covariates such as age, ethnicity and *APOE* ε4 genotype are used to correct for differences between individuals. They are encoded with binary or numeric values, with -9 representing missing data.

2.3. Power calculations

Statistical power is the likelihood that an effect will be detected if present. Calculations were used to determine the power required to detect variants with a specified minor allele frequency (MAF) and whether the study has enough power to detect an association with disease.

2.3.1. Power to detect variants

The capacity to detect variants of certain MAFs can be calculated using the equation below. With a sample size *n* of 292 individuals, this study has almost 100% power to detect variants with a MAF between 1 and 5% as shown in Table 2.1.

$$n \; x \; log \; (1 - MAF) = log \; (1 - POWER)$$

Table 2.1. The power of this study to detect variants of varying MAFs based on a sample size of 292 individuals. Almost 100% power is achieved to identify variants of MAF 1-5%, however this is halved for even rarer variants of MAF 0.1%. Larger samples sizes would increase the power to detect these.

MAF	Power (%)
0.001	44.3
0.005	94.7
0.01	99.7
0.02	~100
0.03	~100
0.04	~100
0.05	~100

2.3.2. Power to detect associations

The statistical power of this study was calculated using the software QUANTO v1.2.4 (Gauderman, 2002) for various MAFs and odds ratios. Figure 2.5 illustrates the estimated power to detect associations of rare variants with a trait. The study was specified as matched case-control in QUANTO, assuming a sample number of 100 cases and controls. The MAF ranged from 0.01 to 0.05 in increments of 0.01 for a log additive model with a 5% Type I error rate. Multiple odds ratios were tested to achieve significant power for detection. Variants with low frequency and odds ratios are difficult to detect in small sample studies; large sample sizes would be required to investigate these variants. However, it is possible for rare variants with large effect sizes, to be able to achieve significant power.

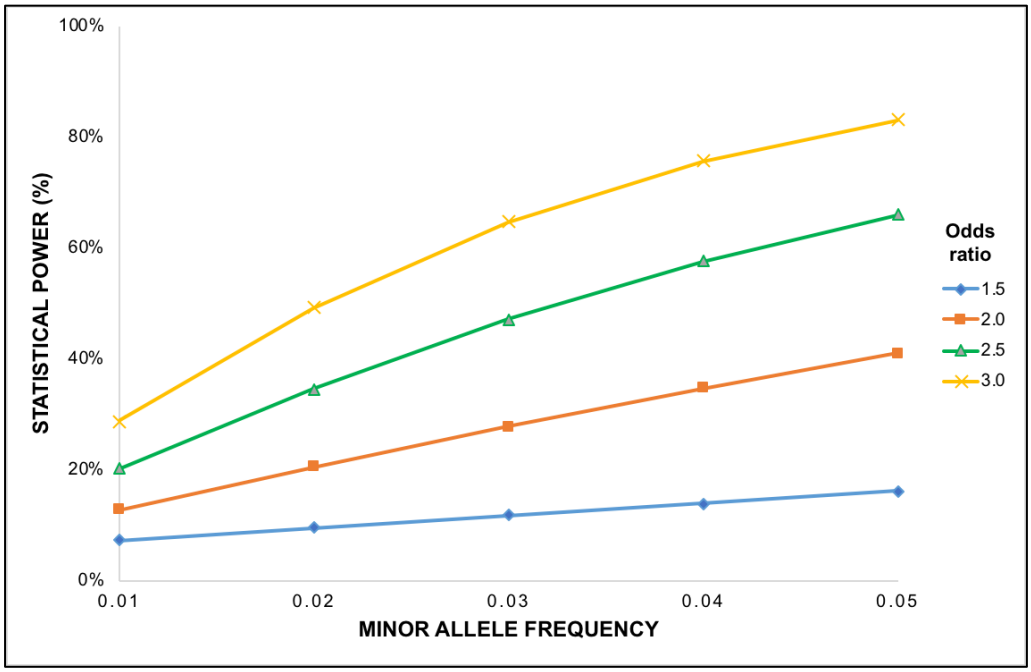


Figure 2.5. Estimated power to detect a disease association with rare variants of varying MAFs. The power of this study was calculated, assuming 100 cases and an equal number of controls. The power to detect greater effect sizes for rare variants increases with MAF. Results are shown here for four odds ratios (1.5, 2.0, 2.5 and 3.0). Rare variants with greater effect sizes can be detected with sufficient power.

Chapter 3: Whole-exome sequencing pipeline

3.1. Introduction

Genome-wide association studies (GWAS) have been fundamental in identifying hundreds of susceptibility loci associated with complex disease traits. They were designed to detect common variants (MAF > 5%) that associate with a phenotype. The majority of the trait-associated SNPs identified in GWAS are located within intronic or intergenic regions and often large distances from the nearest gene (Manolio, 2010). However, these variants have small effects and therefore only explain a small proportion of complex disease heritability (Visscher *et al.,* 2012). These SNPs may also be in linkage disequilibrium (LD) with several other variants in close proximity, making it difficult to pinpoint the causative variant (Klein *et al.,* 2012). As a result, GWAS associates a genomic location or LD block with a phenotype rather than a specific variant. Whilst highly associated with the disease, these SNPs rarely exhibit functional, disease-causing effects (Manolio, 2010). It is possible that rare variants within these loci that go undetected by GWAS, may have larger effect sizes and therefore account for the missing heritability (Manolio *et al.,* 2009; Rivas *et al.,* 2011). Advances in next-generation sequencing (NGS) have permitted exomes and entire genomes to be explored at single-base level. This has led to the use of NGS to map these genetic loci, allowing much rarer variants to be uncovered.

3.1.1. Next-generation sequencing

NGS technology has revolutionised the field of genetics by allowing large volumes of sequencing data to be generated in a significantly reduced time frame. Automated Sanger sequencing was used prior to NGS, which uses fluorescently-tagged dideoxyribonucleotide triphosphates (ddNTPs) to terminate elongation of DNA strands upon

incorporation of a modified ddNTP (Sanger *et al.,* 1977). This method is expensive and time-consuming and typically fails to detect insertions and deletions; however, with 99.9% accuracy it remains the “gold standard” for validating polymorphisms. In contrast, massively parallel NGS allows millions of fragments to be sequenced in a single run and is better suited for variant discovery (Koboldt *et al.,* 2013). The surge in demand for high-throughput sequencing has resulted in the advent of several different NGS platforms, at substantially lower costs (Kircher and Kelso, 2010).

Whole genome sequencing (WGS) attempts to sequence the entirety of the genome, allowing investigation of all variant types - SNPs, indels, structural variants, copy number variants (CNVs) and regulatory sequences (Sims *et al.,* 2014). Despite being significantly cheaper than Sanger sequencing, WGS is still costly and computationally expensive in terms of data storage and processing. Whole-exome sequencing (WES) offers a robust alternative with costs being 5x cheaper and generating 5% the amount of data as WGS. As a result, greater numbers of samples can be sequenced and analysed in a fraction of the time (Goh and Choi, 2012).

3.1.2. Whole-exome sequencing process

WES is a targeted method that captures protein coding regions. Exons are parts of a gene that are transcribed into mature messenger RNA (mRNA) after introns have been removed by RNA splicing; this mRNA undergoes translation to synthesise proteins. The exome only accounts for approximately 1-2% of the human genome, yet harbours an estimated 85% of known mutations underlying disease-related traits (Choi *et al.,* 2009; Ng *et al.,* 2009). However, the technique is still subject to a few technical limitations. Non-coding regulatory elements, such as promoters, enhancers and non-coding RNAs cannot be studied and there is limited ability to analyse CNVs and other structural variations (Sims *et al.,* 2014). WES exhibits non-uniform coverage as it is enriched for coding regions, meaning that variants in low coverage

areas can be missed (Bamshad *et al.,* 2011; Warr *et al.,* 2015). Another caveat is that errors in alignment or sequencing artefacts could be misinterpreted as rare variants leading to false positive findings (Altmann *et al.,* 2012). Therefore, Sanger sequencing is still widely used as a complementary strategy to validate NGS findings. Despite these drawbacks, WES is a well-justified method to discover rare variants underlying complex genetic disorders/traits (Bamshad *et al.,* 2011).

A number of different NGS platforms are available; although they share commonality in function and output, they differ in their biochemistry. With a cost per Mb of $0.10, output of >1000 million reads per run and lowest sequencing error rate of ~0.1%, Illumina is the current market leader in NGS technology (Goodwin *et al.,* 2016). Here we used the Agilent SureSelect kit to prepare libraries for sequencing of human genomic DNA (gDNA) samples on the Illumina HiSeq 2500. This process is illustrated in Figure 3.1.

The Agilent SureSelect (Agilent, USA) Human All Exon V4+UTRS in-solution capture kit was used for target enrichment of exonic regions and 5' and 3' untranslated regions (UTRs). The SureSelect method has been demonstrated to have the greatest target enrichment efficiency and depth of coverage when compared to other platforms, such as Illumina Nextera (Illumina, USA) and NimbleGen (Roche, USA) (Shigemizu *et al.,* 2015). It uses an in-solution hybrid capture methodology to enrich for target regions and generate a sequencing library. Fragmented gDNA was covalently bound to platform-specific adapters. Ligated products were amplified to generate sufficient signal for each sequencing reaction. RNA library "baits" were designed to be complementary to regions of interest. The library was incubated with these biotinylated RNA capture probes for 16 hours to allow hybridisation. Streptavidin-coated magnetic beads were added to bind to probes and pull target regions out of solution; unbound DNA was washed off. A further PCR amplification step enriched captured libraries, producing a sequence-ready library.

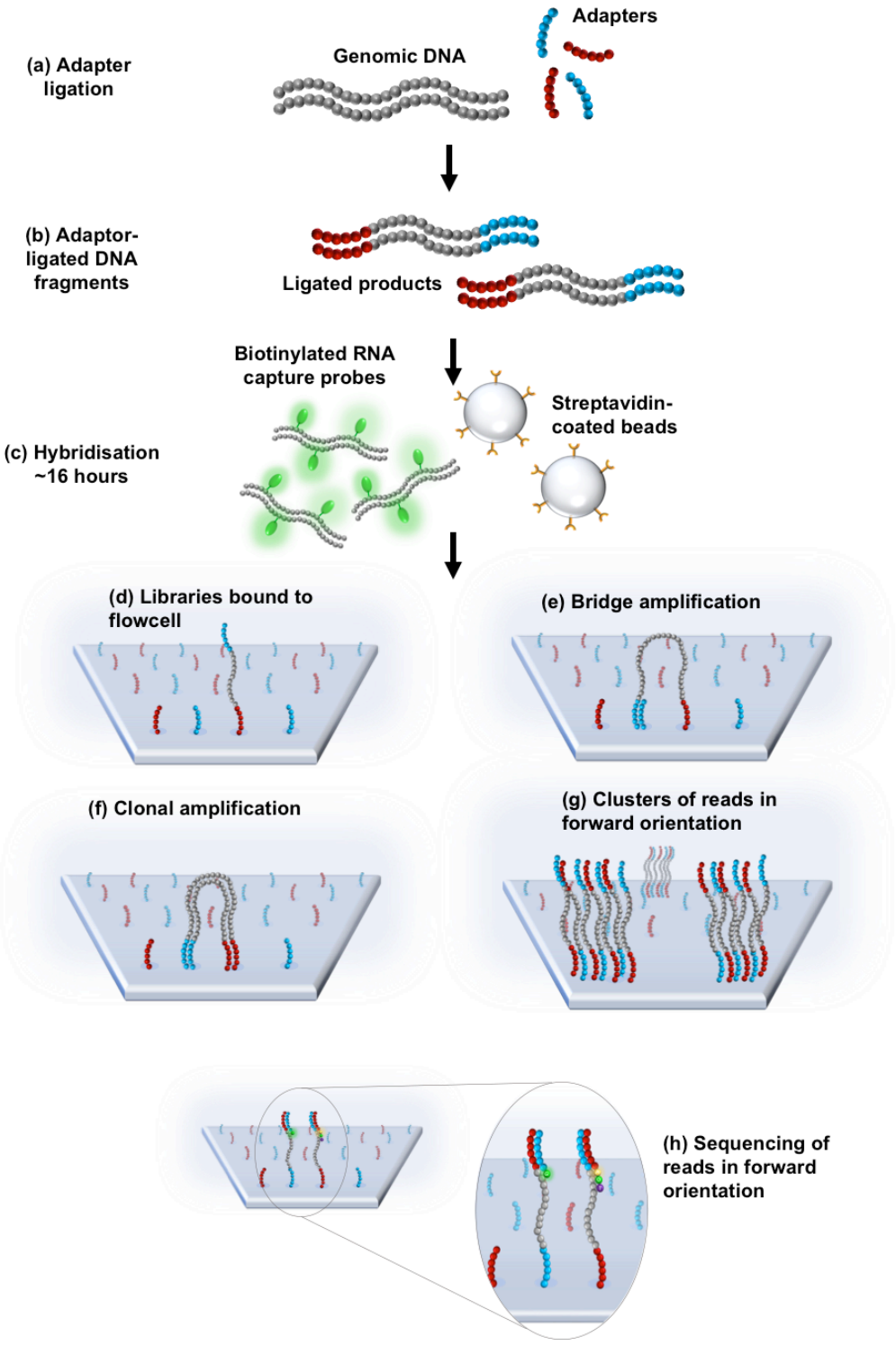


Figure 3.1. Schematic depicting Illumina library preparation and sequencing-by-synthesis technology using paired-end primers. (a) Paired-end adaptors are ligated to genomic DNA. (b) Ligated products are incubated with biotinylated RNA capture probes and hybridised for 16 hours. (c) Streptavidin-coated magnetic beads pull target regions out of solution. (d) Sequencing libraries hybridised to glass microfluidic flow cell. (e) Bridge amplification as strands fold over onto second primer. (f) Clusters of copied fragments generated by clonal amplification. (g) Reverse strands selectively cleaved to leave forward reads. (h) Sequencing primers anneal and forward reads are sequenced by synthesis. This is then repeated for reverse strands. Adapted from the Illumina sequencing workflow general use image with permission.

The Illumina HiSeq 2500 uses sequencing by synthesis technology. Sequencing libraries were hybridised to the glass surface in a flow cell, which has microfluidic channels containing immobilised primer sequences complementary to the library adapters. Each fragment is amplified on a solid surface to increase the signal for base imaging, to create clusters of copies of the same fragment, through a process known as bridge amplification. Here, a DNA strand folds over and hybridises to a second type of primer on the flow cell. Polymerases generate a complementary strand, resulting in two single-stranded copies tethered to the flow cell. This is performed in parallel, resulting in clonal amplification of all fragments and creates millions of clusters of copied fragments in close proximity.

Reverse strands are selectively cleaved and washed off leaving only clusters of fragments in the forward orientation. A sequencing primer is annealed to the adapters and sequencing begins to generate the first read. During each cycle, the four fluorescently-tagged nucleotides compete for addition to the nucleic acid chain. The reversible termination chemistry allows the fluorophore to be imaged once the nucleotide is incorporated and immediately cleaved to continue the process with the next base. This is repeated with bridge amplification of the reverse strand to sequence from each end of the template. This “paired-end” approach generates almost double the amount of data. Forward and reverse reads are paired, creating contiguous sequences. Base calls are determined by the emission wavelength and signal intensity during each cycle, resulting in highly accurate base-by-base sequencing with considerably lower error rates than other methods (Quail *et al.,* 2012). The number of cycles determines read length. Raw Illumina reads are output in FASTQ format, which stores base calls and quality score per base encoded using ASCII characters.

3.1.3. Quality control

Assessment of the sequencing output is vital to ensure that reads are of a suitable quality for downstream analyses. Initial assessment of raw

reads is necessary to identify problems with library preparation or the sequencing run (Guo *et al.,* 2014). Tools like FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) can highlight quality issues in NGS data, such as contamination of DNA material with adapter sequences and biases in base composition. FastQC performs checks for read length, per base quality scores, GC content, sequence duplication and overrepresented sequences in the raw data. Base calling accuracy is determined by a Phred scale quality score (Q); a logarithmic scale that measures the probability that a base is incorrectly called. Phred scores range from low to high, with higher quality scores indicating a smaller probability of error, as Table 3.1 shows. A score of Q30, with a 99.9% accuracy rate, is considered the benchmark for good quality data (Endrullat *et al.,* 2016).

Table 3.1. Phred quality scores. Phred scores are attributed to each base and determine the accuracy of the base call. A higher Phred score means a smaller probability of an incorrect base being called. A value of Q10 indicates a low quality score, with the cut-off threshold of Q30 commonly used to denote high quality base calls.

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
Q10	0.1	90%
Q20	0.01	99%
Q30	0.001	99.9%
Q40	0.0001	99.99%

Further quality control (QC) is performed once data is aligned to a reference. This determines accuracy of alignments, sequence duplication levels and sequence contamination. Mapping quality scores quantifies the probability that a read is misplaced and are also measured on a Phred scale. This is dependent on the read length, number of mismatches and multiple mapping locations (Li, Ruan, *et al.,* 2008). As paired end reads are longer in length, this significantly reduces the likelihood of ambiguity in mapping. Evaluating the quality profile of raw and aligned data allows identification of problems that need to be addressed to obtain high quality sequences.

3.1.4. Alignment

Millions of short reads are generated by the sequencer, which require alignment to a reference genome to be meaningful. This is the first step in NGS analysis pipelines, after which several processing steps are undertaken to adjust for sequencing errors and other potential bias. Due to large volumes of data generated by sequencers, alignment tools have been optimised to enable fast, highly accurate alignment with lower memory usage. Therefore, different tools are designed with various approaches to optimise speed and accuracy to achieve the best possible alignment. A mapping algorithm aims to match a read to a location in the reference sequence, whilst tolerating a certain level of mismatch to allow variant detection.

3.1.5. Processing

Aligned sequence data is stored in SAM (Sequence Alignment Mapped) format which stores position and quality information for all reads. There is no standard protocol for processing NGS data, however best practice guidelines have been proposed for refining datasets. These steps adjust for quality issues and sequencing errors to improve alignment and reduce potential false positive variant calls downstream. Removal of PCR duplicates is essential as amplification bias can cause problems with variant calling, resulting in some alleles being overrepresented. Indels within reads can often cause mappers to misalign with mismatches, leading to false positive calls. This can affect variant detection, therefore local realignment is performed using known variants to realign bases around these regions (DePristo *et al.,* 2011). Base quality scores should also be recalibrated to detect systematic errors made by the sequencer in order to obtain good variant calls. Picard tools and Genome Analysis Toolkit (GATK) suite of software are the most widely used for processing DNA sequencing data.

3.1.6. Variant calling and filtering

Variant calling is the identification of sites with an alternate base to the reference sequence. SNPs and indels are called separately using a database of known variants, such as dbSNP and 1000 Genomes where most variants are already catalogued. Generating a high-quality variant call set is essential for the discovery of real variants. A range of tools exist for variant calling and these can vary considerably in the algorithms they use. Simple fixed filter algorithms make a heterozygous variant call when the alternative allele is present in between 20 - 80% of reads (Nielsen *et al.,* 2011). GATK, FreeBayes and SAMtools rely on similar Bayesian approaches to calculate posterior probabilities of the true genotype, but differ in the prior probabilities used (Heinrich *et al.,* 2012). Other tools such as VarScan implement statistical methods to screen reads for mismatches that are combined into unique SNPs and indels if observed in multiple reads (Pabinger *et al.,* 2014). Each has their relative strengths and limitations, however GATK is increasingly recognised as the “gold standard” for variant discovery in large NGS datasets (De Summa *et al.,* 2017).

3.1.7. Annotation

Annotation provides important information to identify known and novel variants against established databases that are curated regularly, such as dbSNP and 1000 Genomes. Documented SNPs and indels are each given a unique Reference SNP cluster ID (Rs #) as an identifier. With the large amounts of data produced by NGS, annotating large numbers of variants requires computational tools to automate the process (Pabinger *et al.,* 2014). Variants can be annotated with a range of additional information which is useful for interpreting the biological function of variants, such as whether it resides within or near a gene, affects protein coding or has a putative functional effect. Importantly, the allele frequency in public databases can also be incorporated into variant annotation (Koboldt *et al.,* 2013).

Functional annotation tools help prioritise potential disease-causing variants for further analysis by classifying mutations as benign or deleterious (McCarthy *et al.,* 2014). Annovar and Ensembl's Variant Effect Predictor (VEP) are popular annotation programs, which use prediction tools to estimate functional impact by observing changes in the amino acid sequence. Sorting Intolerant from Tolerant (SIFT) and Polymorphism Phenotyping v2 (PolyPhen-2) use protein sequence homology to score variants on the likelihood that the mutation is damaging to the protein (Itan and Casanova, 2015). However, *in silico* functional characterisation is affected by choice of transcript sets, which can generate varying annotations between programs (McCarthy *et al.,* 2014). Additionally, variants of interest should be functionally tested *in vitro* to validate any *in silico* predictions as a deleterious prediction does not necessarily translate to causality (Guerreiro *et al.,* 2014).

3.2. Aims

This chapter describes in detail the bioinformatics pipeline implemented for processing of the NGS data, from raw sequence data to filtered variants. Functional annotation will aid further prioritisation of variants for downstream analyses.

3.3. Methods

Whole-exome sequencing was performed at University College London using the Agilent SureSelect XT Human All Exon V4 library kit, which targets 51Mb of exonic sequence. Samples were multiplexed and run on an Illumina HiSeq 2500. The Agilent SureSelect kit provides 99.8% coverage of intended target regions (Chilamakuri *et al.,* 2014).

3.3.1. Sample preparation

Human post-mortem brain tissue was obtained from UK brain banks in Bristol, King's College London, Manchester, Newcastle and Oxford; all contributing centres of the Brains for Dementia Research network. Written informed consent was obtained from donors at point of

recruitment into the study. All centres and the Coordinating Centre operate within the legal requirements determined by the Human Tissue Authority. The overall project and its management committees have been approved by the National Research Ethics Service. Each Brain Bank is also approved by the National Research Ethics Service and their respective institutional review boards. The BDR ethics approval number is 08/H0704/128+5. DNA was extracted from cerebellar tissue as described in Section 2.1 for an initial 382 BDR samples. A table indicating sample dropout at each quality control stage is shown in Appendix 7.1. All diagnoses were neuropathologically confirmed. *APOE* ε4 SNPs were genotyped using a TaqMan assay as described in Section 2.5. Quality checks were performed to ensure DIN scores of at least 7 and to ensure DNA was not subject to degradation as assessed on gels. An example gel from the Agilent TapeStation is shown in Figure 3.2. Samples passing this quality check were also quantified using the Quant-IT dsDNA Broad Range Assay kit to obtain accurate concentrations. If DNA concentration was greater than 250ng/µl the sample was diluted to a concentration of 3µg in 45µl for sequencing.

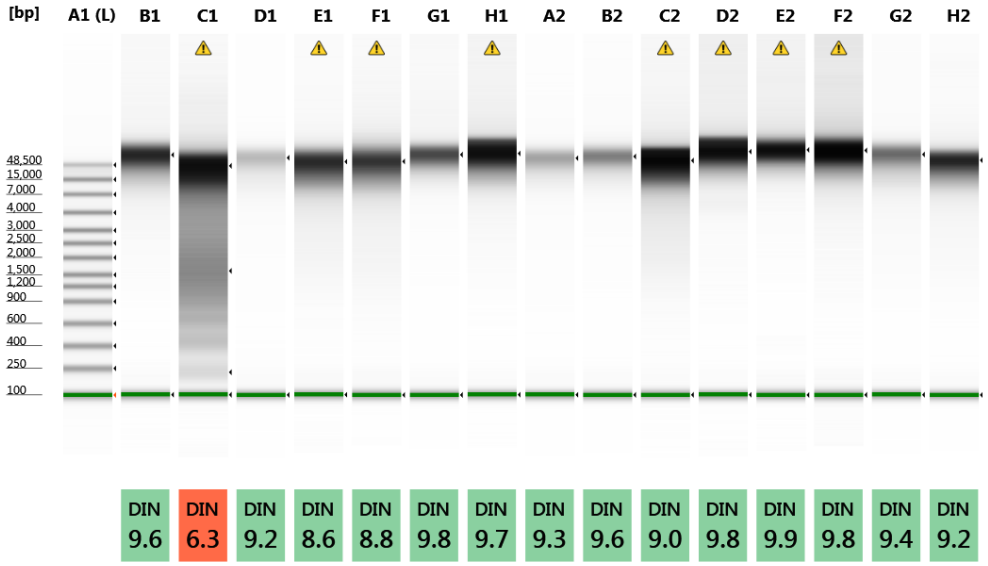


Figure 3.2. Quality check gel from Agilent TapeStation. Samples 1 to 15 were run on the TapeStation to generate quantitative measures for DNA quality. Smears indicated degradation of the DNA and DIN scores below 7 were considered poor quality. Sample 2 (C1) appears degraded with a low DIN score and therefore it required re-extraction.

Table 3.2. Demographics of BDR LOAD and control samples and other phenotypes. The cohort contains individuals from multiple centres. Samples are separated by (a) LOAD cases, (b) cognitively normal controls and (c) other phenotypes. Information provided includes the number of individuals (N) from each centre and percentage of sample size, mean age of death with standard deviation (Mean AAD+SD), the number and percentage of females per centre (Females), the number and percentages of individuals harbouring at least one APOE ε4 allele (APOE ε4+), APOE ε4 minor allele frequency (APOE ε4 MAF) and number and percentage of individuals with two ε4 alleles (APOE ε4ε4). Key: AAD, age at death; N, number of individuals; SD, standard deviation.

(a) Late-onset Alzheimer's disease cases.

Centre	N (%)	Mean AAD (± SD)	Females (%)	*APOE* ε4+ (%)	*APOE* ε4 MAF	*APOE* ε4ε4 (%)
Bristol	13 (8.8)	84.7 (±9.0)	5 (38.5)	10 (76.9)	0.27	1 (7.7)
KCL	34 (23.0)	83.2 (±6.4)	17 (50.0)	21 (61.8)	0.35	4 (11.8)
Manchester	36 (24.3)	81.2 (±8.8)	19 (52.8)	25 (69.4)	0.47	11 (30.6)
Newcastle	15 (10.1)	87.5 (±7.9)	8 (53.3)	12 (80.0)	0.32	2 (13.3)
Oxford	51 (34.5)	82.5 (±14.6)	25 (49.0)	33 (64.7)	0.43	9 (17.6)
All	149	83.0 (±10.7)	74 (49.7)	101 (67.8)	0.4	27 (18.2)

(b) Cognitively normal control samples.

Centre	N (%)	Mean AAD (± SD)	Females (%)	*APOE* ε4+ (%)	*APOE* ε4 MAF	*APOE* ε4ε4 (%)
Bristol	11 (20.0)	85.6 (±6.1)	5 (45.5)	7 (63.6)	0.14	0 (0.0)
KCL	19 (34.5)	81.3 (±10.5)	11 (57.9)	4 (21.1)	0.11	0 (0.0)
Manchester	7 (12.7)	86.3 (±9.4)	4 (57.1)	0 (0.0)	0.00	0 (0.0)
Newcastle	6 (10.9)	82.8 (±6.5)	2 (33.3)	2 (33.3)	0.08	0 (0.0)
Oxford	12 (21.8)	83.7 (±10.7)	5 (41.7)	1 (8.3)	0.04	0 (0.0)
All	55	83.5 (±9.2)	27 (49.1)	14 (25.5)	0.12	0 (0.0)

(c) Other phenotype samples.

Centre	N (%)	Mean AAD (± SD)	Females (%)	*APOE* ε4+ (%)	*APOE* ε4 MAF	*APOE* ε4ε4 (%)
Bristol	2 (2.4)	82.0 (±9.9)	1 (50.0)	2 (100.0)	0.25	1 (50.0)
KCL	15 (17.6)	74.3 (±24.0)	7 (46.7)	5 (33.3)	0.2	1 (6.7)
Manchester	29 (33.0)	72.6 (±25.2)	10 (34.5)	13 (44.8)	0.22	0 (0.0)
Newcastle	8 (9.4)	81.0 (±10.3)	0 (0.0)	5 (62.5)	0.36	0 (0.0)
Oxford	34 (38.6)	79.4 (±19.6)	16 (47.1)	12 (35.3)	0.2	1 (2.9)
All	88	78.3 (±17.4)	34 (38.6)	37 (42.0)	0.23	3 (3.4)

A further QC step was performed prior to sequencing; only samples with concentration greater than 1µg with an absence of smearing were run on the sequencer. A total of 292 DNA samples were successfully sequenced, for which demographics are shown in Table 3.2 by phenotype.

3.3.2. Next-generation sequencing analysis

Raw sequence data in FASTQ format was obtained for successfully sequenced samples. Two FASTQ files were output per sample, one with reads in the forward direction and the other containing paired reads in the reverse direction. The NGS analysis pipeline shown in Figure 3.3 was followed. As the majority of data analysis programs are implemented using a command line interface in a Unix operating system, all data processing analyses performed were undertaken in a Linux-based server operating Red Hat. The commands used in the following sections are shown for an individual file unless otherwise stated.

3.3.3. Initial quality control

Initial quality checks were performed on the FASTQ files from Illumina to determine raw data quality from the sequencer. FastQC (Babraham Bioinformatics) is a java-based program that provides information on sequence quality to identify any potential issues prior to analysis. It generates a QC report per sample that can help determine whether any problems originate in the sequencer or in the starting library material itself. This report includes information on GC composition, adapter content, overrepresented sequences, read length, read distribution, duplicate sequences and quality scores. The program was run for multiple samples in parallel on the server to increase processing speed. Paired read files were run through FastQC individually. The number of threads used was selected using --threads *N* and -o names the output file report.

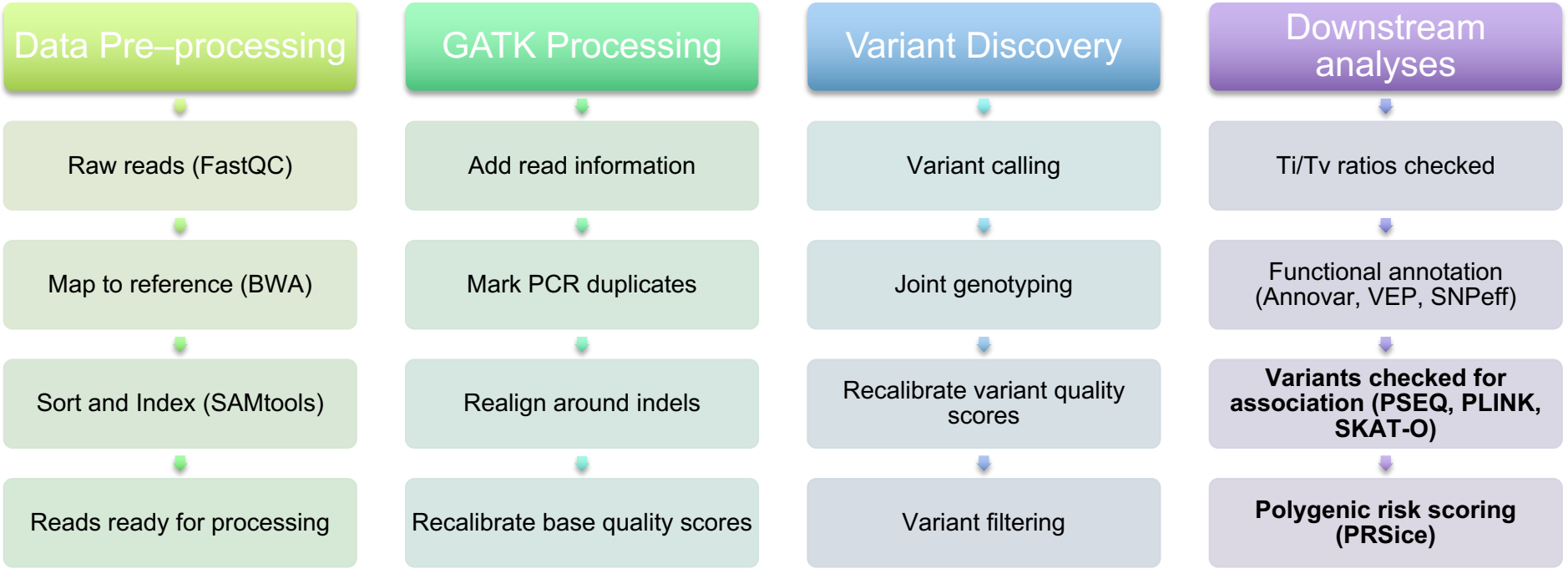


Figure 3.3. The bioinformatics pipeline implemented for processing the raw whole-exome sequencing FASTQ files generated by Illumina. Pre-processing steps were performed using various tools, which are indicated in brackets. The GATK Best Practices pipeline was utilised for processing of aligned reads to generate the highest quality variants calls possible. Downstream analyses to explore genetic variation in AD are indicated in bold. Adapted from Genome Analysis Toolkit Best Practices for DNA–Seq (DePristo et al. 2011).

```
fastqc --threads 8 <sample_r1>.fastq -o </Folder/sample_qc>
```

Options are:

- --threads: number of processors to use
- -o: output file

MultiQC (Ewels *et al.,* 2016) generates a combined report by collating results from all FastQC files. This summarises information from all samples onto shared plots, providing a comparative overview of data quality and allowing easier observation of trends and outliers.

```
multiqc <Folder>
```

3.3.4. Preparing reference files for input

The reference sequence used for alignment was human genome assembly build 19 (hg19) from UCSC, released in March 2009. The hg19 reference sequence is identical to NCBI GRCh37 and Ensembl references, however they differ in the annotations they use. The reference is made up of contigs from several different individuals to build a representation of the genome. Therefore, coordinates change between different assemblies as genomic regions are updated to better represent human variation. The most recent version is build 38 (hg38), however hg19 was used to avoid remapping as updates are released for hg38. Sequence files for individual chromosomes were downloaded from the UCSC website and concatenated to create the reference.

```
wget --timestamping
'ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/*'

cat chr*.fasta > hg19.fasta
```

The `wget` command retrieves content from the web server provided. Concatenation is executed using cat and output to a new file with > output.

3.3.5. Indexing reference sequence

Aligning NGS data is computationally demanding, therefore many programs require the reference genome or sequence files to be indexed before processing. Indexing allows regions of the sequence to be accessed efficiently without storing the entire file in memory, reducing processing power. The index file stores basic information about the reference file, such as contig name, number of bases and bytes per contig. Reference sequence indexing was required for steps involving BWA, Samtools and Picard tools. The algorithm for constructing the BWT index was selected as -a *bwtsw*, the Burrows-Wheeler Transform Smith-Waterman algorithm which is optimal for large genomes. BWA uses an FM-index, first constructing a suffix array of all strings in the input sequence, sorting rotations lexographically and storing the transformed sequence in an index. This enables fast pattern searching as only a small region of the reference needs to be decompressed to look for matches (Li and Durbin, 2010).

```
bwa index -a bwtsw -p hg19 hg19.fasta
```

Options are:

- -a: algorithm for constructing BWT index
- -p: prefix of the output database

For GATK, it is necessary to have a fasta index, different to that produced by BWA and a sequence dictionary. Samtools (http://samtools.sourceforge.net) generates this index using the option faidx, taking the reference fasta as input. The index contains information about the length, names and file offset for each contig in the file, to quickly locate the required reference bases in the sequence.

```
samtools faidx hg19.fasta
```

Picard tools (www.picard.sourceforge.net/index.shtml) are located in the folder /usr/local/src/picard/picard-tools-1.117. The CreateSequenceDictionary tool was used to generate the sequence

dictionary for the reference, describing the contents of the fasta file such as sequence name, length, genome assembly amongst other information.

```
java -jar picard/CreateSequenceDictionary \
REFERENCE=hg19.fasta \
OUTPUT=hg19.dict
```

3.3.6. Alignment

All sequencing reads were aligned to the human genome build hg19 using the Burrows-Wheeler aligner (BWA) (Li and Durbin, 2009). BWA uses indexing, a data compression method via Burrows-Wheeler Transform (BWT) to map short reads with speed, allowing for mismatches and gaps (Li and Durbin, 2009). Modern alignment algorithms use seeding to find maximal exact matches (MEM) for part of the read with the reference. This is very fast so identifies a subset of locations where a sequence is most likely to align to. Seeds are then extended to build optimal alignments using the affine-gap Smith-Waterman algorithm. While this step is slower it is also more accurate and reduces computational power as less extension is required. Therefore, overall BWA-MEM is less time-consuming and more robust to sequencing errors, making it better than other short read aligners (Li and Homer, 2010). The BWA-MEM alignment algorithm is invoked with the mem sub-command.

```
bwa mem -t 8 hg19.fasta sample_r1.fastq sample_r2.fastq > sample.sam
```

Options are:

- -t: number of threads to use

3.3.7. Pre-Processing

Aligned data is output in Sequence Alignment Mapped (SAM) format, generally 20-30Gb in size. To optimise storage and processing, SAM files were compressed to smaller binary SAM (BAM) files, which are between 4-10Gb in size. BAM files can be indexed so that portions of

the file can be accessed without loading the whole file. The data requires pre-processing in order to maximise technical accuracy and correct for potential sequencing errors highlighted during QC. Initial pre-processing steps include sorting, marking duplicates and recalibration of base quality scores, which are performed using Picard tools. The BAM first requires sorting by coordinate. Format conversion, sorting and indexing were performed using Picard tool SortSam.

```
java -Xmx4g -jar picard/SortSam \
INPUT=sample.sam  \
OUTPUT=sample.sorted.bam \
SO=coordinate \
CREATE_INDEX=TRUE \
VALIDATION_STRINGENCY=LENIENT
```

Options are:

- SO: sort order to sort file by coordinate {Coordinate, Unsorted, Queryname}
- CREATE_INDEX: generate an index for BAM file {True, False}
- VALIDATION_STRINGENCY: determines whether certain validation measures (e.g. reads, qualities, tags) can be bypassed to improve processing performance {Strict, Lenient, Silent}

3.3.8. Aligned quality control

The unprocessed BAM file was checked for quality using FastQC to determine any problems with alignment. Alignment statistics were also generated to calculate the level of coverage for each sample against the reference sequence.

```
fastqc --threads 8 sample.bam -o sample.aln.qc
```

3.3.9. Add read group information

A read group refers to a set of reads generated from a single lane in a sequencing run for a biological sample. Read groups are identified by a number of tags in the BAM file that allow differentiation of various

technical features associated with artefacts. GATK requires this information to be included to mitigate the effects of these artefacts during duplicate marking and base recalibration steps. It will fail with errors if this is not provided. Read group information comprises of several fields, such as sample name, sequencing platform, flowcell and library, as defined below.

SM = Sample

This field refers to the name of the sequenced sample. GATK tools treat read groups with the same SM value as the same sample.

PL = Platform

This is the platform used to generate the sequencing data.

PU = Platform unit

The platform unit holds three types of information: flowcell barcode, lane of the flowcell and sample barcode. These barcodes are unique identifiers of the sequencing technology used. This field is optional; however, it takes precedence over ID for base recalibration if present. These fields differentiate factors contributing to technical batch effects.

ID = Read group identifier

This tag is a unique identifier that helps distinguish which read group each read belongs to. In Illumina data, the ID is generally composed from the flowcell, lane name and number to create a globally unique identifier across all sequencing data.

LB = DNA preparation library identifier

The library field is used during marking of duplicates to determine whether read groups may contain molecular duplicates, in case of the same DNA library being sequenced on multiple lanes.

Flowcell barcode and lane information were obtained from the BAM file header to generate the PU field using Samtools. An example output is shown below.

```
samtools view -H sample.sorted.bam | grep '^@RG'
```

Options are:

- -H: header
- grep: search for string

```
@RG ID:0    PL:illumina
PU:Illumina0044_20080829_1_Pilot1_Ceph_12414_B_lib_1_2Kb_MP_Pilot1_Ce
ph_12414_B_lib_1_2Kb_MP   LB:Lib1 PI:2750 DT:2008-08-28T20:00:00-0400
SM:NA12414  CN:bcm
```

This information was collected for all samples prior to modifying read groups using Picard tool AddOrReplaceReadGroups, as the ID field is null as default. Specifying the maximum memory allocation for the process with -Xmx4g allows the program to make full use of the available resources, where 4g represents the heap size to be used by Java in Gb. BDR IDs were used as the sample name, library and sequencing platform were SureSelect and Illumina, respectively. Flowcell barcode and lane were obtained from the BAM file to update the PU field as MG00HS20_HH5TGBCXX. As PU contained the flowcell barcode, the read group IDs were given as the sample name with a prefix, KM_BB001.

```
java -Xmx4g -jar picard/AddOrReplaceReadGroups \
INPUT=sample.sorted.bam \
OUTPUT=sample.sorted.rg.bam \
RGLB=SureSelect \
RGPU=Platform \
RGSM=Sample \
RGPL=Illumina \
RGID=Sample
```

File validation

All BAM files needed to meet the formatting requirements for GATK. It should be sorted by coordinate, list read group information in the header and must pass validation by Picard ValidateSamFile. This can diagnose errors in the file and provide warnings about improper formatting, alignment and flag values that need to be addressed before proceeding.

```
java -Xmx4g -jar picard/ValidateSamFile.jar \
INPUT=sample.sorted.rg.bam \
MODE=SUMMARY
```

Options are:

- MODE: mode of output information {Verbose, Summary}

3.3.10. Mark duplicates

Duplicate reads present in sequencing data can cause problems in variant calling. These reads are exact copies that share the same sequence and alignment position, leading to overrepresentation of some alleles as a result. They generally arise from sample library preparation due to amplification biases in PCR, through inadequate amounts of starting material or enrichment methods. Mistakes in sequencing technology can also cause artefacts known as optical duplicates. This occurs when single amplification clusters are incorrectly detected as multiple clusters by the optical sensor of the sequencing instrument. These duplicate reads are not informative and therefore should not be counted as additional evidence for putative variants. The MarkDuplicates tool identifies and tags duplicates for each read to ensure GATK tools disregard them without removal. It differentiates between the primary and duplicate reads using an algorithm that ranks them by the sum of their base-quality scores. A metrics file indicating numbers of duplicates is also produced.

```
java -Xmx4g -jar picard/MarkDuplicates \
INPUT=sample.sorted.rg.bam \
OUTPUT=sample.sorted.rg.marked.bam \
METRICS_FILE=metrics \
CREATE_INDEX=TRUE \
VALIDATION_STRINGENCY=LENIENT
```

Options are:

- METRICS_FILE: output metrics information

3.3.11. Realignment around indels

GATK is considered the industry standard for identifying SNPs and indels in DNA sequencing data (McKenna *et al.,* 2010). It is a collection of command line tools developed for analysis of high throughput NGS data and variant discovery. The following steps that were performed with GATK required the Samtools index and sequence dictionary for the hg19 reference genome from Picard. The GATK resource bundle contains files required for working with human sequencing data and was obtained from ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/. This includes databases of known variants such as dbSNP (Sherry *et al.,* 2001), 1000 Genomes (Altshuler *et al.,* 2012) and HapMap (International HapMap Consortium *et al.,* 2003).

Algorithms used for initial alignment can produce various mapping artefacts. In particular, indels near the end of reads often cause mappers to misalign the bases, leading to false positive SNPs during variant detection. The local realignment process aims to minimise the number of incorrect mismatches and generate consistent alignments across all reads for variant calling. Using known variants to identify indels and realign bases around these regions can reduce the number of sequencing artefacts. This two-step process involves GATK tool RealignerTargetCreator using a known variant database to create a table of possible indels, identifying the intervals that require

realignment. The dbSNP database version 138 was used as the known sites to locate indel regions. Only indels from the file were considered.

```
$ java -Xmx4g -jar GenomeAnalysisTK.jar \
-T RealignerTargetCreator \
-R hg19.fasta \
-I sample.sorted.rg.marked.bam \
-o indels.intervals \
--known dbsnp.vcf
```

In the second step IndelRealigner determines the optimal consensus sequence for regions in the list and performs the local realignment of reads.

```
$ java -Xmx4g -jar GenomeAnalysisTK.jar \
-T IndelRealigner \
-I sample.sorted.rg.marked.bam \
-R hg19.fasta \
-targetIntervals indels.intervals \
-o sample.sorted.rg.marked.realign.bam
```

Options are:

- -T: GATK tool
- -R: path to reference sequence
- -I: path to input file
- -o: output filename
- --known: VCF file of known variants
- -targetIntervals: list of loci to be targeted

3.3.12. Fix mate information

Paired-end reads are generated from DNA fragments being sequenced from both ends. This results in two reads per fragment in forward and reverse-complement orientations which are referred to as “mate pairs”, which are a known distance apart. Depending on the initial fragment size and read length, these pairs could overlap. Picard FixMateInformation verifies the mate-pair information for paired-end

reads in the sorted BAM file and fixes them into the correct position again if necessary. This is because alignments may change during the realignment process due to indels altering the known distance between mate pairs. Although IndelRealigner fixed the position of reads that were realigned, it did not account for the mate-pairs of unadjusted reads. Therefore, this additional step was performed to ensure all paired reads are fixed.

```
$ java -Xmx4g -jar picard/FixMateInformation.jar \
INPUT=sample.sorted.rg.marked.realign.bam \
OUTPUT=sample.sorted.rg.marked.realign.fixed.bam \
SO=coordinate \
VALIDATION_STRINGENCY=LENIENT \
CREATE_INDEX=TRUE
```

Options are:

- SO: sort order

3.3.13. Base recalibration

Data generated by sequencers is subject to various sources of technical error, such that systematic errors can affect base quality. These quality scores are estimates of error emitted by the sequencer that quantify the likelihood of a correct base call. Recalibration is performed to obtain accurate variant calls, which rely on high base quality scores. They are used by variant calling algorithms to weigh up the evidence for a variant allele at a particular site. The base quality score recalibration (BQSR) process involves two key stages. Initially, BQSR utilises machine learning to build an empirical model of observed errors based on the data and known variants, then it uses this to adjust quality scores accordingly.

BaseRecalibrator uses the average Phred scores assigned by the sequencer to split reads into separate read group bins. These read bins undergo two passes to compute empirical scores. During the first pass, aggregate errors and observations are calculated for each bin. BQSR

assumes that any SNP in the data not in dbSNP is an error. The tool also incorporates covariates such as base position within the read and information from neighbouring bases to improve accuracy and outputs all of this information to a recalibration report.

```
$ java -Xmx4g -jar GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R hg19.fasta \
-I sample.sorted.rg.marked.realign.fixed.bam \
--knownSites dbsnp.vcf \
-o recalibration_report1.grp
```

The second pass involves building a second model to evaluate the covariation remaining after recalibration. BQSR performs recalibration on the data table in the first report to create a second report. This is used to generate before and after plots to visualise the effects of the recalibration process.

```
$ java -Xmx4g -jar GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R hg19.fasta \
-I sample.sorted.rg.marked.realign.fixed.bam \
-knownSites dbsnp.vcf \
-BQSR recalibration_report1.grp \
-o recalibration_report2.grp
```

Options are:

- -BQSR: base quality score recalibration

Both reports were used to plot scores before and after recalibration. The R libraries required (ggplot2, gplots, gsalib and reshape) were installed with R version 3.4.1 and GATK tool AnalyzeCovariates was used to generate the plots.

```
$ R
Install.packages(“ggplot2”)
Install.packages(“gplots”)
```

```
Install.packages("gsalib")
Install.packages("reshape")

$ java -Xmx4g GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
-R hg19.fasta \
-before recalibration_report1.grp \
-after recalibration_report2.grp \
-plots plot.pdf \
-csv output.csv
```

PrintReads rewrites the quality scores using the first report, which contains the covariation data required for recalibrating the base quality scores in the BAM file. The new score is computed using the formula below as the sum of global difference between the reported and empirical quality scores, plus the differences in quality bin scores and covariate effects. This yields more realistic base quality scores, improving the accuracy and robustness of variant calling.

$$Q_{actual} = GlobalQ + \Delta all + \Delta readGroup + \Delta quality + {}^{n}\Sigma_i \Delta covariate_i$$

```
$ java -Xmx4g -jar GenomeAnalysisTK.jar \
-T PrintReads \
-R hg19.fasta \
-I sample.sorted.rg.marked.realign.fixed.bam \
-BQSR recalibration_report1.grp \
-o sample.sorted.rg.marked.realign.fixed.recal.bam
```

3.3.14. Variant calling

The detection of germline variants requires ascertaining an individual's genotype at each potential variant locus. GATK HaplotypeCaller is capable of calling SNPs and indels simultaneously using a reference confidence model with known variants from dbSNP. It uses local *de novo* assembly of haplotypes when calling variants, allowing more accurate, incremental variant discovery on large cohorts (Van der Auwera *et al.,* 2013).

HaplotypeCaller identifies "ActiveRegions" in the genome with high

numbers of mismatches as potential variant sites using a sliding window approach, as illustrated in Figure 3.4. For each region, a De Bruijn-like graph of possible haplotypes is assembled using the read data and a likelihood score is computed to determine the most plausible paths. Each haplotype is realigned against the reference using the Smith-Waterman algorithm to identify potential variant sites, resulting in high sensitivity and precision (Pfeifer, 2016). Pairwise alignment of each read against the all possible haplotypes is performed using PairHMM, a Hidden Markov model algorithm. This yields a matrix of scores that express the likelihood of observing each read for the given haplotype, taking into account quality scores. These scores are marginalised to calculate per-read likelihoods of alleles at each potential variant site, by selecting the haplotype with the maximum likelihood that supports that read. These per-read allele likelihoods are aggregated, then Bayes' theorem is applied to determine the most likely genotype at each potential variant site. Sites with sufficient evidence for at least one of the variant alleles will be called as a variant and a genotype will be assigned.

HaplotypeCaller performs per-sample variant discovery to generate intermediate genomic VCF files containing raw, unfiltered SNP and indel calls. The --intervals parameter specifies the regions on which to perform variant calling; a list of coordinates corresponding to the exome targets used by the library capture kit was provided. The dbSNP file is also passed to annotate variants with their corresponding reference IDs.

```
$ java -jar GenomeAnalysisTK.jar \
-R hg19.fasta \
-T HaplotypeCaller \
-I sample.sorted.rg.marked.realign.fixed.recal.bam \
--emitRefConfidence GVCF \
--intervals SureSelect_all_exon_V4.bed \
--dbsnp dbsnp.vcf \
-o sample.g.vcf
```

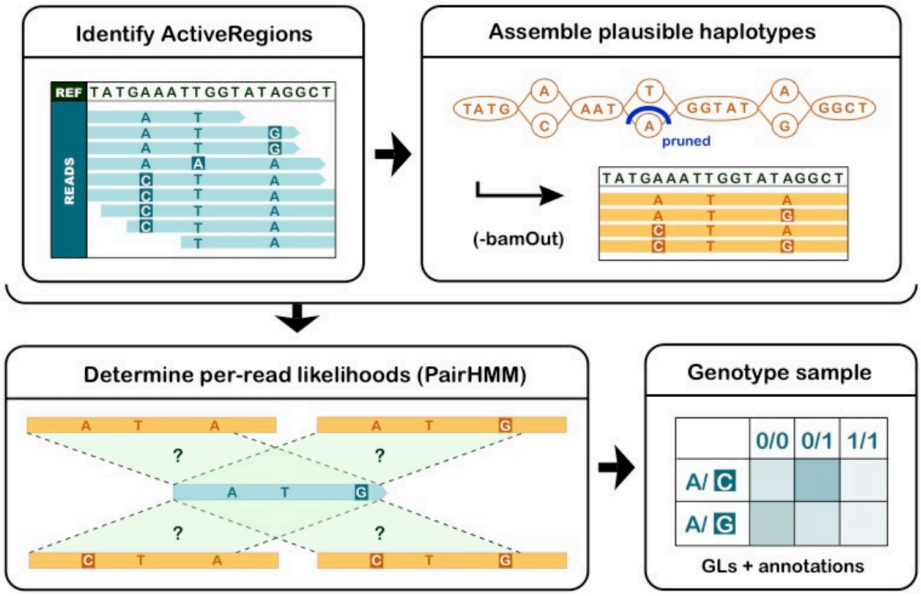


Figure 3.4. The algorithm used by GATK HaplotypeCaller to assign sample genotypes during variant calling. (1) Define ActiveRegions. (2) Determine haplotypes by assembly of the ActiveRegion. (3) Determine likelihoods of haplotypes given the read data. (4) Assign sample genotypes. Figure obtained from Poplin *et al*. (2017).

Options are:

- --emitRefConfidence: mode for emitting reference confidence scores
- --intervals: file containing coordinates of intervals to target
- --dnsnp: dbSNP reference variants file

3.3.15. Joint genotyping

The individual gVCFs require joint genotyping with GenotypeGVCFs and filtering before further analysis. GenotypeGVCFs combines the data from all samples to re-genotype samples as a cohort. This multi-sample approach produces a square matrix of genotypes with information on all sites of interest in all samples. Variants in regions with low coverage would not be called when processed separately; by accumulating evidence from all samples to detect low frequency variants present in multiple samples, with greater sensitivity. Singleton variants unique to a single sample are likely to be missed in low quality regions as a result, however this can be minimised by increasing

sample size. Joint calling also allows clearer distinction between homozygous reference sites and missing data, emitting a genotype call for every site where an individual in the dataset has evidence for variation. This step can be rerun for the cohort at any point with more samples. A list of all gVCF files is provided to GenotypeGVCFs and this produces a combined call set of raw SNPs and indels for filtering.

```
$ java -jar GenomeAnalysisTK.jar \
-R hg19.fasta \
-T GenotypeGVCFs \
-V list_of_vcfs.list \
-o 1_bdr_combined.vcf
```

Options are:

- -V: list of VCF files

3.3.16. Variant quality score annotation

Context annotations inform about properties of the sequencing data specific to variants. This largely includes quality information which can be used for filtering purposes with VariantAnnotator. Most annotations are added during the variant calling stage, however additional information can be provided. The dbSNP database was used to populate the ID columns with reference IDs from known variants. Coverage information was also added to get depth of coverage for variants.

```
$ java -Xmx32g -jar GenomeAnalysisTK.jar \
-T VariantAnnotator \
-R hg19.fasta \
-V 1_bdr_combined.vcf \
-o 2_bdr_combined_annotated.vcf \
-A Coverage \
--dbsnp dbsnp.sorted.149.vcf
```

Options are:

- -A: annotation

3.3.17. Variant quality score recalibration

Variant quality score recalibration (VQSR) uses established variant databases and annotations to distinguish true variant calls from false positive and false negatives, to produce a filtered call set. VariantRecalibrator uses high quality, curated variant databases 1000 Genomes Project and HapMap as training sets to learn the annotation profiles of a subset of “true” variants that are likely to be real. The dbSNP database is used to generate metrics based on the presence of variants in the resource. Information from several annotations is used to build a model to recognise how good and bad variants cluster along each of the dimensions. The individual annotation scores are integrated to generate a single score for each variant, a variant quality score log odds (VQSLOD) ratio, which determines likelihood of it being real. These rules are then applied to the variant call set using a Gaussian mixture model, to assign a new calibrated VQSLOD score to the raw SNPs and indels. Training sets required for recalibration of SNP scores are listed below. The dbSNP version used was v138 subset to exclude sites discovered after v129, omitting the 1000 Genomes Project data, which is used as a separate training set. Training sets required for recalibration of SNP scores:

- HapMap - true sites (Q15)
- Omni genotyping array - true sites (Q12)
- 1000 Genomes - non-true sites (Q10)
- dbSNP - known sites resource, not used in training (Q2)

```
$ java -Xmx32g -jar GenomeAnalysisTK.jar \
-R hg19.fasta \
-T VariantRecalibrator \
-input 2_bdr_combined_annotated.vcf \
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
hapmap_3.3.hg19.sites.vcf \
-resource:omni,known=false,training=true,truth=true,prior=12.0
1000G_omni2.5.hg19.sites.vcf \
-resource:1000G,known=false,training=true,truth=false,prior=10.0
```

```
1000G_phase1.snps.hg19.sites.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
dbsnp_138.hg19.excluding_sites_after_129.vcf \
-an QD \
-an FS \
-an SOR \
-an MQ \
-an MQRankSum \
-an ReadPosRankSum \
-an InbreedingCoeff \
-mode SNP \
-tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
-recalFile recalibrate_SNP.recal \
-tranchesFile recalibrate_SNP.tranches \
-rscriptFile recalibrate_SNP_plots.R
```

Options are:

- -resource: training set to use
- -an: annotation {QD - depth of coverage, FS - Fisher strand, SOR - strand odds ratio, MQ - read mapping quality, MQRankSum - mapping quality rank sum test, ReadPosRankSum - read position rank sum test, InbreedingCoeff - inbreeding coefficient}
- -mode: variant calling mode {SNP, Indel}
- -tranche: tranche threshold for % true positives
- -recalFile: recalibration information
- tranchesFile: tranches information
- rscriptFile: R script to generate recalibration plots

Variants are then filtered using ApplyRecalibration, based on score to exclude any that do not meet the selected sensitivity and specificity thresholds as determined by the tranche it falls in: 90%, 99%, 99.9%, 100%. The tranches represent the score above which the given percentage of variants in the training call set are included. VQSR is performed for SNPs and indels separately as they require different training sets and annotations. The above recalibration parameters are applied to SNPs first.

```
$ java -Xmx32g -jar GenomeAnalysisTK.jar \
-R hg19.fasta \
-T ApplyRecalibration \
-input 2_bdr_combined_annotated.vcf \
-mode SNP \
--ts_filter_level 99.0 \
-recalFile recalibrate_SNP.recal \
-tranchesFile recalibrate_SNP.tranches \
-o 3_bdr_annotated_recalsnps_rawindels.vcf \
```

The following training set was required for indel recalibration, along with the annotations shown below:

- Mills - true sites (Q12)

```
$ java -Xmx32g -jar GenomeAnalysisTK.jar \
-R hg19.fasta \
-T VariantRecalibrator \
-input 3_bdr_annotated_recalsnps_rawindels.vcf \
-resource:mills,known=true,training=true,truth=true,prior=12.0
mills.vcf \
-an QD \
-an FS \
-an SOR \
-an MQRankSum \
-an ReadPosRankSum \
-an InbreedingCoeff \
-mode INDEL \
-tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
-recalFile recalibrate_INDEL.recal \
-tranchesFile recalibrate_INDEL.tranches \
-rscriptfile recalibrate_INDEL_plots.R
```

Recalibration was applied to indels using these parameters.

```
$ java -Xmx32g -jar GenomeAnalysisTK.jar \
-R hg19.fasta \
-T ApplyRecalibration \
-input 3_bdr_annotated_recalsnps_rawindels.vcf \
-mode INDEL \
--ts_filter_level 99.0 \
```

```
-recalFile recalibrate_INDEL.recal \
-tranchesFile recalibrate_INDEL.tranches \
-o 4_bdr_annotated_recal_variants.vcf
```

Variants were tagged by the tranche they fell into and then filtered to retain only those above the selected threshold. To find a balance between sensitivity and specificity, multiple thresholds were tested and 99% was chosen. Variants with VQSLOD scores below this threshold were filtered out.

```
java -Xmx32g -jar GenomeAnalysisTK.jar \
-R hg19.fasta \
-T SelectVariants \
-V 4_bdr_annotated_recal_variants.vcf \
--excludeFiltered \
-o 5_bdr_annotated_recal_filtered.vcf
```

3.3.18. Functional annotation

Functional annotations can predict the effect a variant may have on biological function based on positional and genomic information. They are dependent on external databases of known variants, such as dbSNP, 1000 Genomes and the Exome Sequencing Project (ESP). These are readily updated so the most current annotations are available. The latest dbSNP database version 149 was used to obtain the most up-to-date reference ID annotations. The VCF was acquired from the NCBI repository and prepared for use by annotation tools.

```
$ wget
ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b149_GRCh37p13/VC
F/All_20161121.vcf.gz
```

Chromosome notations were altered to match that of the reference sequence by adding “chr” to the numerical values.

```
awk '{if($0 !~ /^#/) print "chr"$0; else print $0}' All_20161121.vcf
> dbsnp.awk.149.vcf
```

Lastly, the chromosome contigs were sorted into the same order as the reference sequence using Picard SortVcf, which required the sequence dictionary.

```
$ java -jar picard/SortVcf.jar I=dbsnp.149.vcf O=dbsnp.149.sorted.vcf
SEQUENCE_DICTIONARY=ucsc.hg19.dict
```

Variants were annotated with multiple tools to obtain as comprehensive annotations as possible. The most popular tools include Annovar (Wang *et al.,* 2010), Ensembl's Variant Effect Predictor (VEP) (McLaren *et al.,* 2016) and SnpEff (Cingolani *et al.,* 2012), which used various databases to add functional information.

ANNOVAR

Annovar is a Perl-based command-line program useful for generating gene, region and filter-based annotations. It has some beneficial functionalities, allowing variants to be annotated from multiple public databases, such as dbSNP, 1000 Genomes and ESP which can offer different information. Gene-based annotation identifies whether variants affect the amino acid sequence to cause protein-coding changes in genes. The program uses scores from SIFT, PolyPhen, MutationTaster, MutationAssessor, PhyloP conservation and Likelihood Ratio Test (LRT) to estimate deleterious impact of variants (Adzhubei *et al.,* 2013; Chun and Fay, 2009; Kumar *et al.,* 2009; Miosge *et al.,* 2015; Pollard *et al.,* 2010; Schwarz *et al.,* 2010). This is particularly useful for exome sequencing datasets to investigate predictable functional effects of coding mutations. Filter-based annotation also allows filtering subsets of variants against public databases to identify novel mutations. It relies on several databases which can be used as necessary; those required for human genome hg19 were downloaded from the server individually as shown.

```
perl annotate_variation.pl -buildver hg19 -downdb -webfrom annovar
refGene humandb/

perl annotate_variation.pl -buildver hg19 -downdb cytoBand humandb/
```

```
perl annotate_variation.pl -buildver hg19 -downdb genomicSuperDups
humandb/

perl annotate_variation.pl -buildver hg19 -downdb -webfrom annovar
esp6500siv2_all humandb/

perl annotate_variation.pl -buildver hg19 -downdb -webfrom annovar
1000g2015aug humandb/

perl annotate_variation.pl -buildver hg19 -downdb -webfrom annovar
exac03 humandb/

perl annotate_variation.pl -buildver hg19 -downdb -webfrom annovar
dbnsfp33a humandb/

perl annotate_variation.pl -buildver hg19 -downdb -webfrom annovar
clinvar_20161128 humandb/

perl annotate_variation.pl -buildver hg19 -downdb -webfrom annovar
intervar_20170202 humandb/

perl annotate_variation.pl -buildver hg19 -downdb -webfrom annovar
hrcr1 humandb/
```

The input was annotated in VCF format and output with 2 CSV files containing known and novel variant information.

```
perl table_annovar.pl \
-vcfinput 5_bdr_annotated_recal_filtered.vcf \
humandb \
-buildver hg19 \
-out 6_bdr_annotated_recal_annovar \
-remove \
-protocol
refGene,cytoBand,genomicSuperDups,esp6500siv2_all,1000g2015aug_all,ex
ac03,dbnsfp33a,clinvar_20161128,intervar_20170202,hrcr1 \
-operation g,r,r,f,f,f,f,f,f,f \
-nastring .
```

Options are:

- -vcfinput: input VCF file

- -buildver: genome build version
- -out: output filename
- -remove: remove all temporary files
- -protocol: databases to use for annotation
- -operation: which operation to use for each corresponding protocol (g=gene-based, r=region-based, f=filter-based)
- -nastring: the string to use if information is not available

Variant Effect Predictor

VEP is a powerful tool for annotation and prioritisation of multiple genomic variants simultaneously, using information from several sources. It can be used as a web interface or a Perl script executed locally on the command line, which allows larger numbers of variants to be processed. Functional consequences on protein coding and regulatory activity of variants are predicted based on the genomic location of variants. Nearest gene and transcript information is provided, along with variation type and impact on protein which takes into account SIFT and PolyPhen-2 prediction scores. Frequency of the variant in several 1000 Genomes populations can also be obtained. A local instance of VEP was run on the command line to annotate variants in the VCF output from Annovar.

```
perl variant_effect_predictor.pl \
--cache \
--merged \
--offline \
-i 6_bdr_annotated_recal_annovar.hg19_multianno.vcf \
--everything \
--format vcf \
--vcf \
--fasta ucsc.hg19.fasta \
--fields
Consequence,Codons,Amino_acids,Gene,SYMBOL,EXON,Polyphen,SIFT,Protein
_position,BIOTYPE \
-o 7_bdr_annotated_recal_annovar_vep.vcf
```

Options are:

- --cache: enable use of cache
- --merged: use merged Ensembl and RefSeq cache
- --offline: run in offline mode (do not connect to server)
- --i: input file
- --everything: shortcut flag to include multiple options (sift_b, polyphen_b, ccds, uniprot, hgvs, symbol, numbers, domains, regulatory, canonical, protein, biotype, tsl, apprix, gene_phenotype, af, af_1kg, af_esp, af_gnomad, max_af, pubmed, variant_class)
- --format: input file format
- --vcf: write output in vcf format
- --fasta: reference sequence file
- --fields: annotation fields of information to include in output
- -o: output file

3.3.19. Quality checks

Quality metrics were obtained using multiple tools to evaluate the variant data. The average ratio of nucleotide transition to transversion (Ti/Tv) for SNPs is often used as an indicator of sequencing quality. Transitions are mutations within the same type of nucleotide, either purine-purine (A↔G) or pyrimidine-pyrimidine (C↔T). Whereas a transversion mutation occurs from purine to pyrimidine or vice versa (A↔C, A↔C, C↔G, G↔T). Transitions occur at a higher rate in the human genome as Cytosine when adjacent to Guanine (CpG sites) can become easily methylated to 5-methylcytosine. This renders it more prone to transition as it can spontaneously deaminate to Thymine, causing a C↔T transition mutation. The expected Ti/Tv ratio for whole genome sequencing is 2.1 as transitions are twice as likely. This ratio should be around 2.8 or higher for on-target variants in whole-exome sequencing; a lower ratio is indicative of false positives caused by random sequencing errors (Carson *et al.,* 2014). GATK VariantEval generated information for different classes of variants.

```
$ java -Xmx128g -jar GenomeAnalysisTK.jar \
-R hg19.fasta \
-T VariantEval \
--eval 7_bdr_annotated_recal_annovar_vep.vcf \
--dbsnp dbsnp.149.sorted.vcf \
-o 7_bdr_annotated_recal_eval.grp
```

Options are:

- --eval: variant file to evaluate

VCFtools calculated Ti/Tv ratio from allele counts of only bi-allelic variants.

```
$ vcftools --vcf 7_bdr_annotated_recal_annovar_vep.vcf \
--TsTv-by-count \
--out 7_bdr_annotated_recal_filtered.tstvcount
```

Options are:

- --eval: input vcf file

Depth of coverage in sequencing is the average number of reads that covers a base. This was calculated per sample and per chromosome to determine average sequencing depth. The capture efficiency of the library kit was also computed for each sample to obtain on-targ*et al*ignment statistics against capture regions. The BED file of capture target coordinates was converted to the required format for Picard before calculating hybrid-selection metrics.

```
java -Xmx128g -jar picard/BedToIntervalList.jar \
I=SureSelect_Regions.bed \
SD=hg19.dict \
O=SureSelect_PicardIntervals.list

java -Xmx128g -jar picard/CalculateHsMetrics.jar \
R=hg19.fasta \
I=sample.sorted.rg.marked.realign.fixed.recal.bam \
O=sample_HsMetrics.txt \
TARGET_INTERVALS=SureSelect_PicardIntervals.list \
BAIT_INTERVALS=SureSelect_PicardIntervals.list
```

Options are:

- SD: reference sequence dictionary
- BAIT_INTERVALS: interval list of bait locations
- TARGET_INTERVALS: interval list of target locations

```
$ vcftools --vcf 7_bdr_annotated_recal_annovar_vep.vcf \
--depth \
--out 7_bdr_annotated_recal_vcftools.depth
```

Options are:

- --depth: calculate depth of coverage

3.3.20. Filtering variants

Filtering is essential for prioritising variants to generate a more manageable dataset. Initially, variants were filtered for only those observed within the exome capture target intervals. A bed file for these regions was obtained from the Agilent SureDesign website for the library kit used (https://earray.chem.agilent.com/suredesign/). SnpSift (Cingolani *et al.,* 2012), an auxiliary of SnpEff, is a filtering tool used to identify variants of interest. The VCF is passed to SnpSift, which then uses the intervals bed file to retain variants in the capture region.

```
$ cat 7_bdr_annotated_recal_annovar_vep.vcf | java -Xmx32g -jar
SnpSift.jar -interval S03723314_Regions.bed >
8_bdr_annotated_recal_exonv4_filtered.vcf
```

Options are:

- -interval: bed file containing regions to filter on

3.3.21. Visualisation

Integrative Genome Viewer (IGV) was used for visualisation of variants in the multi-sample VCF (Robinson *et al.,* 2011; Thorvaldsdottir *et al.,* 2013). IGV displays variants for each sample coloured by genotype and contains information provided from the annotations. The reference

sequence and coverage information for individual bases was also available.

3.4. Results

3.4.1. Sample preparation

A total of 382 DNA samples were extracted with 337 passing our quality control measures and being sent for whole-exome sequencing. Of these, 292 passed the second QC step and were successfully sequenced.

3.4.2. Quality control

FASTQ Files

Raw FASTQ files were individually run through FastQC to determine their quality based on 10 measures. A concatenated FastQC report for sample BB001 which reads in the forward orientation is shown in Figure 3.5.1. The data quality for this sample was high as no errors in sequencing were flagged. Both base and sequence quality scores were above the threshold of Q30, with the majority being greater than Q35. The length of all reads was 101bp which is the standard length of Illumina reads. Sequence duplication levels above 2 represented less than 10% of sequences which is acceptable as this will be accounted for during downstream processing steps. Although per sequence GC content distribution was highlighted as a warning, it follows the theoretical distribution relatively closely. It is also roughly normal with the central peak at approximately 50%. Sequence content across all bases indicates that an equal distribution of all bases was observed in the read data. Significant adapter content was not observed in the sample, suggesting that adaptor contamination is not a problem. The *k*-mer content was flagged as a failure and shows several *k*-mers with a length of 7 nucleotides present within the reads for this sample. The majority of *k*-mers occur at the end of the read which corresponds to library adaptor sequences. The plot also shows a spike in two of the *k*-mers in the middle of the reads, which was seen for a number of

samples in the individual FastQC reports. This could be excess adaptor sequences; however, alignment and quality filtering of the reads should remove the short adaptor sequences.

Observing trends in quality for a large number of individual files is difficult. MultiQC combines multiple FastQC reports to generate single plots for each quality measure, shown in Figure 3.5.2. This allows better comparison of samples to identify outliers. All samples had sequences of a single length (101bp) which is the standard length of Illumina reads. The average quality scores per base had a Phred score greater than Q30 for all samples, with a small reduction in quality to Q25 at the end of the read for a minority of sequences. The majority of sequence quality scores were greater than Q35, indicating high quality reads. There were a small number of reads with scores between Q20 and Q25 which is still acceptable. However, there were also some reads observed with less than Q5 quality which is poor. This can be corrected by filtering low quality information downstream. Adaptor content was present for some samples, but it is mostly observed near the ends of reads as expected and was present in less than 7% of sequences. Sequence duplication levels were relatively low, approximately 10% of reads had a duplication level greater than 10, which could correspond to adaptor content. The percentage of N-count at each position in the read was below 10% and generally less than 5%, but close to 0 for the majority of reads. Per sequence GC content indicates some skewed distributions, however they all follow a normal pattern. The average GC content is between 40-60% for reads with no extreme outliers.

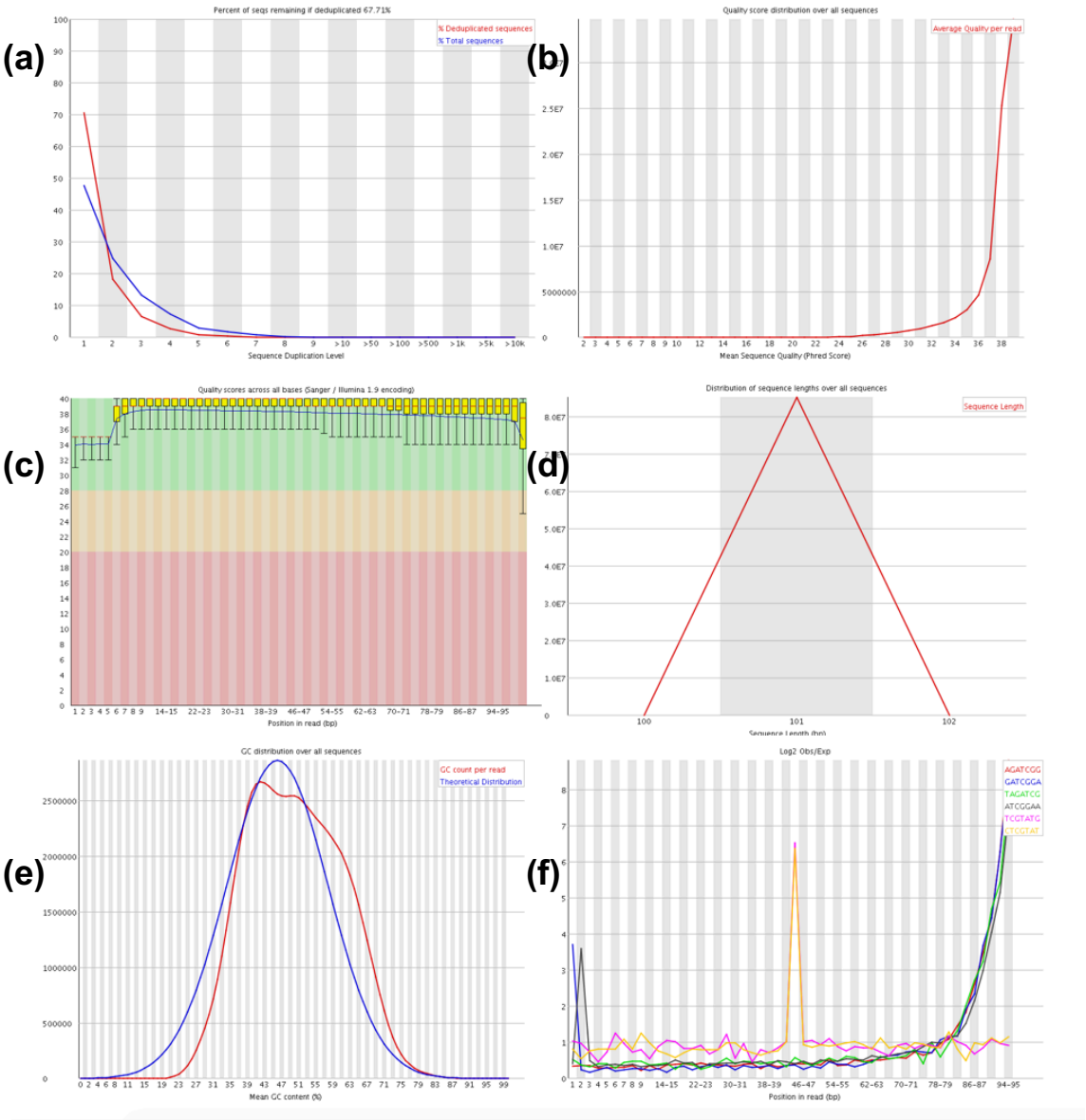
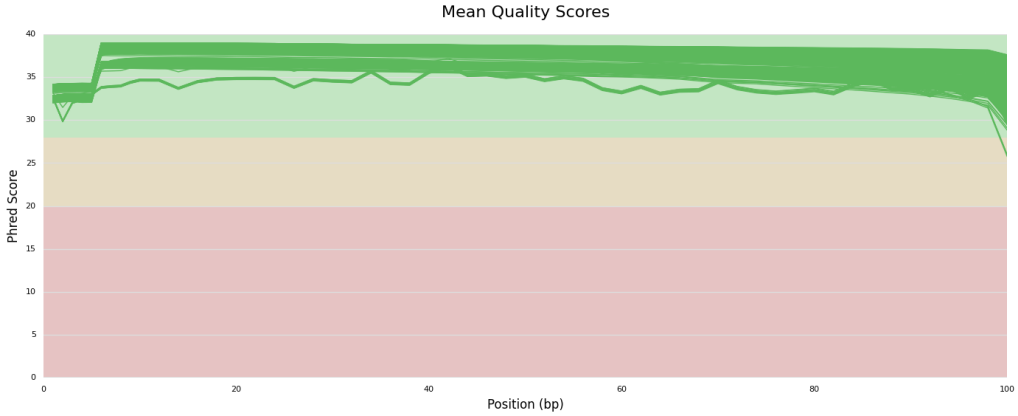


Figure 3.5.1. FastQC report on raw forward reads FASTQ data for sample BB001. The FastQC report details the quality of raw sequence data prior to alignment to determine any potential base calling errors attributed to the sequencer. (a) Percentage of sequences remaining if deduplicated; (b) quality score distribution; (c) quality scores across all bases; (d) distribution of sequence length; (e) GC distribution; (f) overrepresented *k*-mer content.

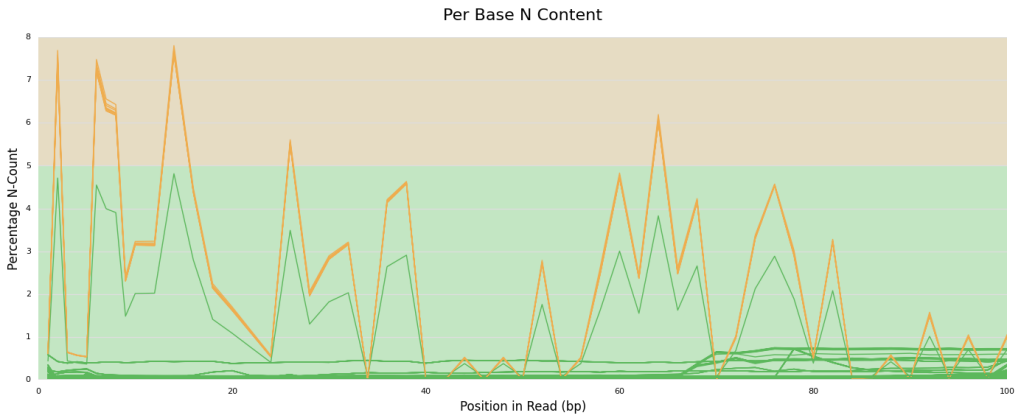
(a)


Mean Quality Scores
Phred Score
Position (bp)

(b)


Per Sequence Quality Scores
1e7
Count
Mean Sequence Quality (Phred Score)

(c)


Per Base N Content
Percentage N-Count
Position in Read (bp)

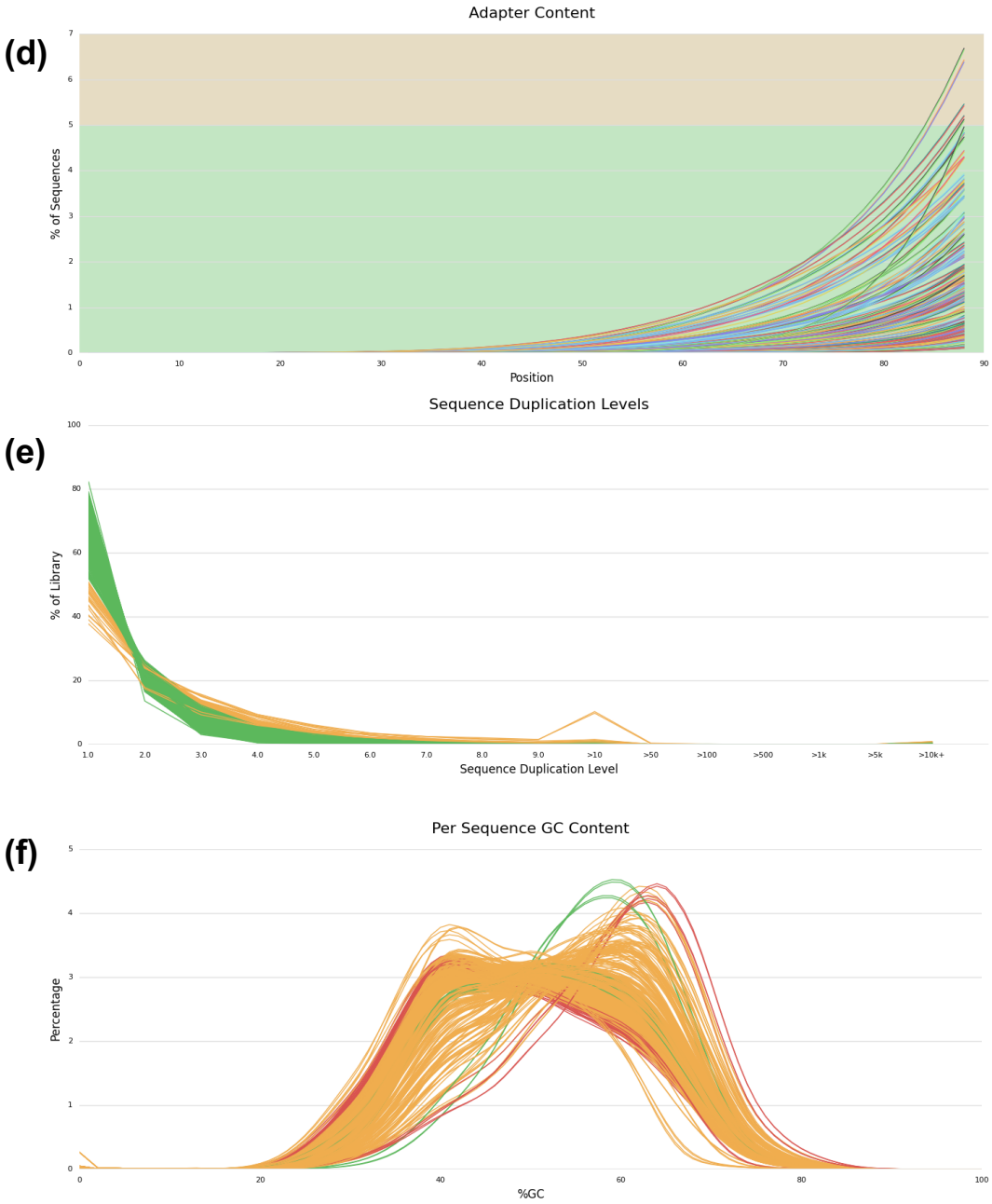


Figure 3.5.2. MultiQC report for raw FASTQ data on all samples. The MultiQC report concatenates all FastQC data to generate a combined report which makes it easier to detect outliers during the quality control process. (a) Mean quality scores; (b) per sequence quality scores; (c) sequence duplication levels; (d) adapter content; (e) per base N content; (f) per sequence percentage of GC content.

BAM Files

Following alignment to the reference and further processing of the BAM file, FastQC was performed again to check if data quality had improved. Figure 3.6.1 contains the composite results for sample BB001. Sequence duplication levels for the individual sample appear to show a small spike in duplicate reads at the 10X level, suggesting that 5-10% of sequences occur more than this. This is mirrored in the MultiQC plot for all aligned and processed BAM files (Figure 3.6.2). This is most likely due to the enrichment bias of exome-sequencing, meaning that the coverage was not uniform due to over-amplification of exonic regions. GC content and read lengths were not altered due to alignment. Although the average quality score declined, three times the number of sequences had a Phred score above Q30, suggesting that adjustment of the scores had improved overall quality. Base quality scores were recalibrated to better reflect the empirical scores and therefore these values appear lower than the original FASTQ Phred scores. However, the majority of bases are still above Q30, with only a small decline in quality at the beginning and end of reads. The distribution of overrepresented *k*-mers significantly improved after processing; this is seen in the graph, as there was no spike in *k*-mers in the middle of reads.

SamStat output in Figure 3.7 shows that 98.8% of reads were mapped to the reference with a Phred score of Q30 or greater, equalling the result from Picard alignment summary statistics. There were only a small percentage of reads that aligned with lower quality scores and no unmapped reads, which demonstrates high quality alignment.

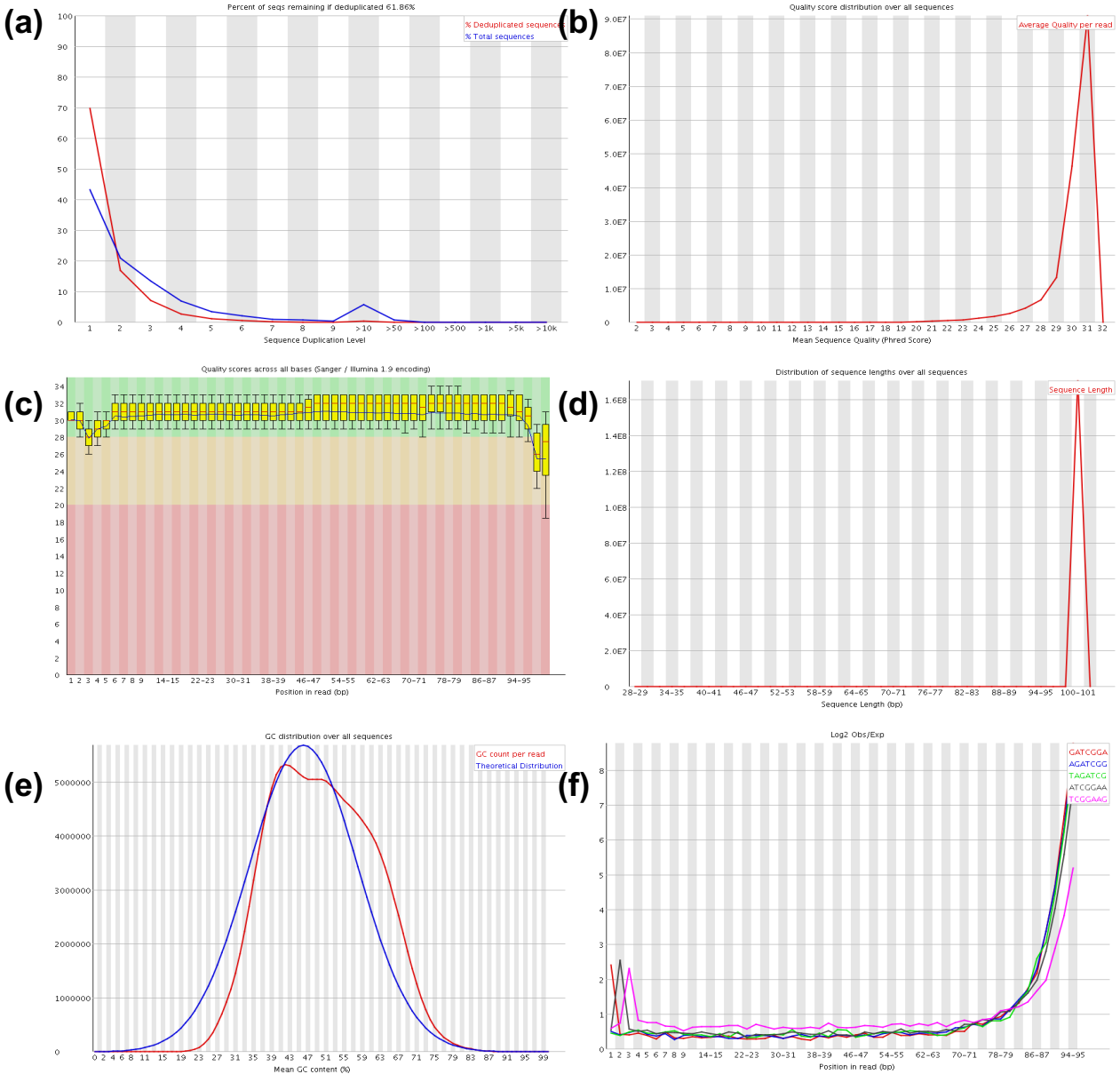
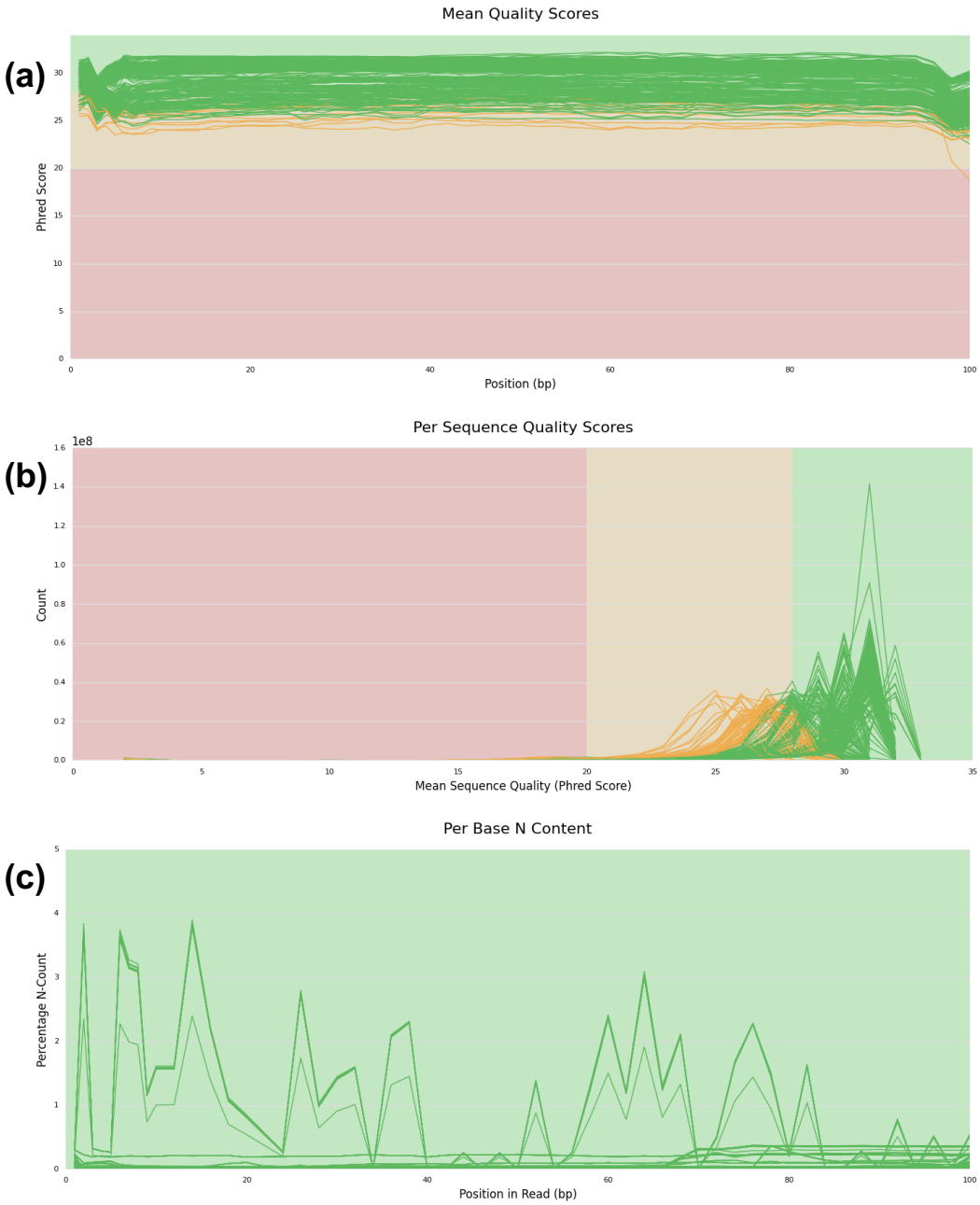


Figure 3.6.1. FastQC report for aligned BAM file BB001 after processing. The FastQC report details the quality of raw sequence data prior to alignment to determine any potential base calling errors attributed to the sequencer. (a) Percentage of sequences remaining if deduplicated; (b) quality score distribution; (c) quality scores across all bases; (d) distribution of sequence length; (e) GC distribution; (f) overrepresented *k*-mer content.


(a)
Mean Quality Scores
Phred Score
Position (bp)
(b)
Per Sequence Quality Scores
1e8
Count
Mean Sequence Quality (Phred Score)
(c)
Per Base N Content
Percentage N-Count
Position in Read (bp)

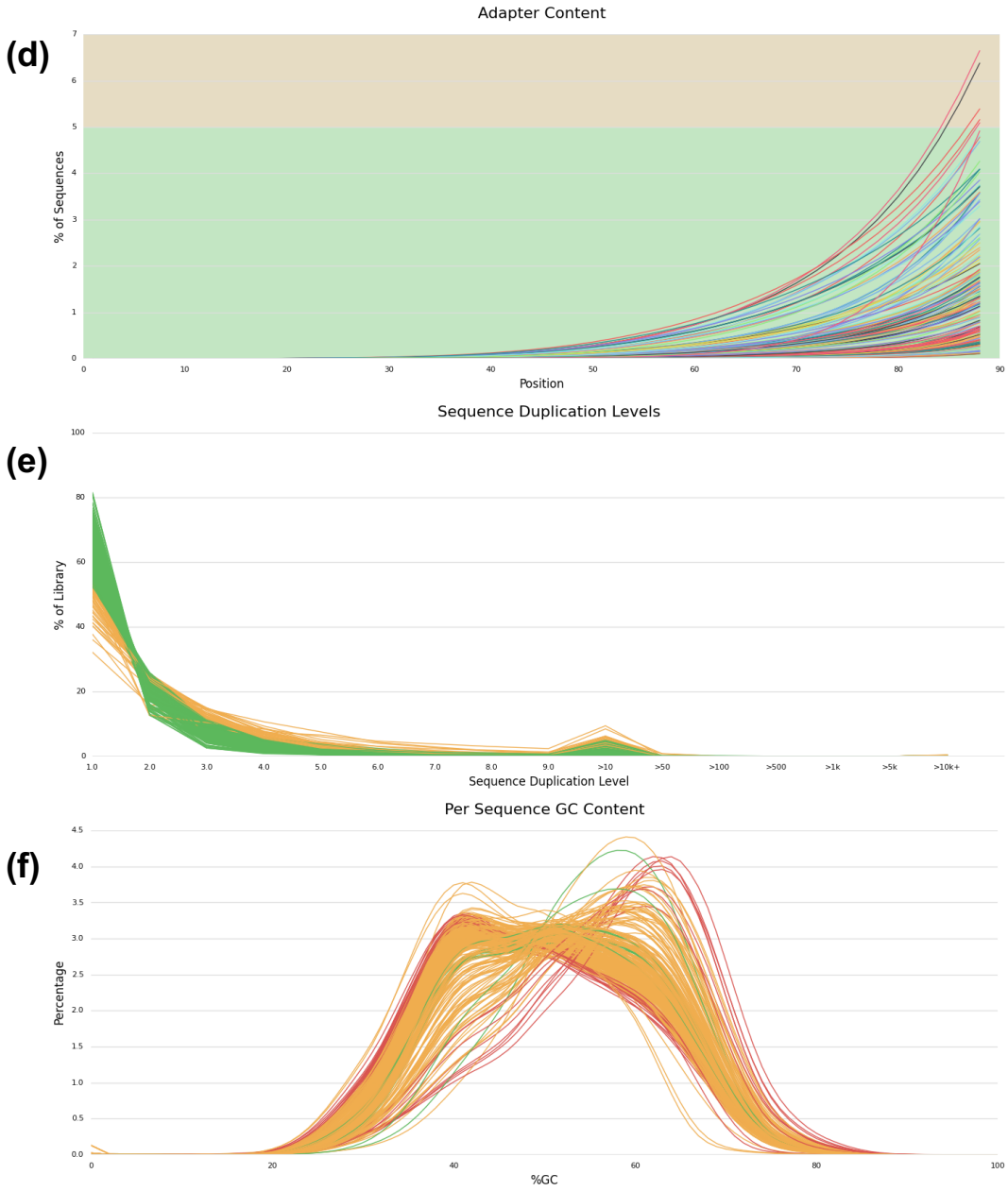


Figure 3.6.2. MultiQC report for aligned BAM files after processing. These FastQC reports highlight any problems that may have occurred during alignment and processing, such as poor mapping quality and contamination in the sequences. (a) Mean quality scores, (b) per sequence quality scores, (c) sequence duplication levels, (d) adapter content, (e) per base N content, (f) per sequence percentage of GC content.

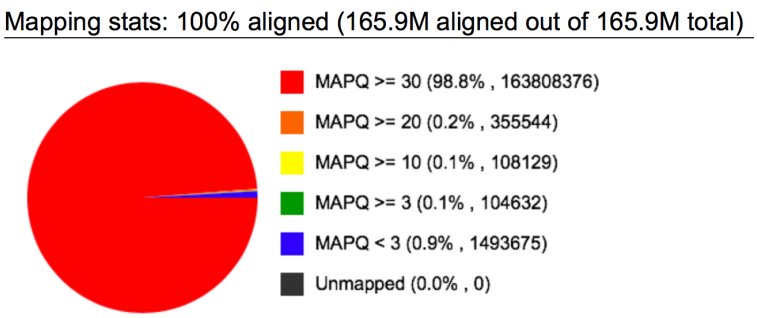


Figure 3.7. SamStats mapping statistics for sample BB001. The number of reads above each mapping quality threshold is shown in the pie chart. The majority of reads were above MAPQ of 30 which is the expected minimum quality. Further refinement steps will exclude mapped reads with lower quality alignment.

3.4.3. Base quality recalibration

Plots were generated by BQSR to compare base quality before and after recalibration. Figure 3.8 shows the plots for sample BB001 with raw scores in pink and recalibrated scores in blue for NGS paired-end reads. An improvement in the accuracy of base quality scores can be observed after applying the recalibration. Plots are separated by SNPs, insertions and deletions. The first panel shows a significant improvement for the recalibrated scores, with a stronger correlation between reported and empirical scores. The second row of plots illustrate the differences between reported and empirical scores for each machine cycle, with recalibrated scores showing improved accuracy. The bottom panel again shows the difference in reported and empirical scores, this time for each of the 16 genomic dinucleotides. All recalibrated points aggregate around $\log_{10}0$, suggesting that following the second pass of base recalibration, reported scores are very similar to empirical scores.

3.4.4. Variant quality recalibration

The tranche plots generated by VQSR in Figure 3.9 illustrate the proportions of true and false positives. The lowest tranche of 90 contains the least number of novel variants with few observed false positives. Therefore tranche 99% was selected to obtain an optimum balance without compromising the identification of true variants. The

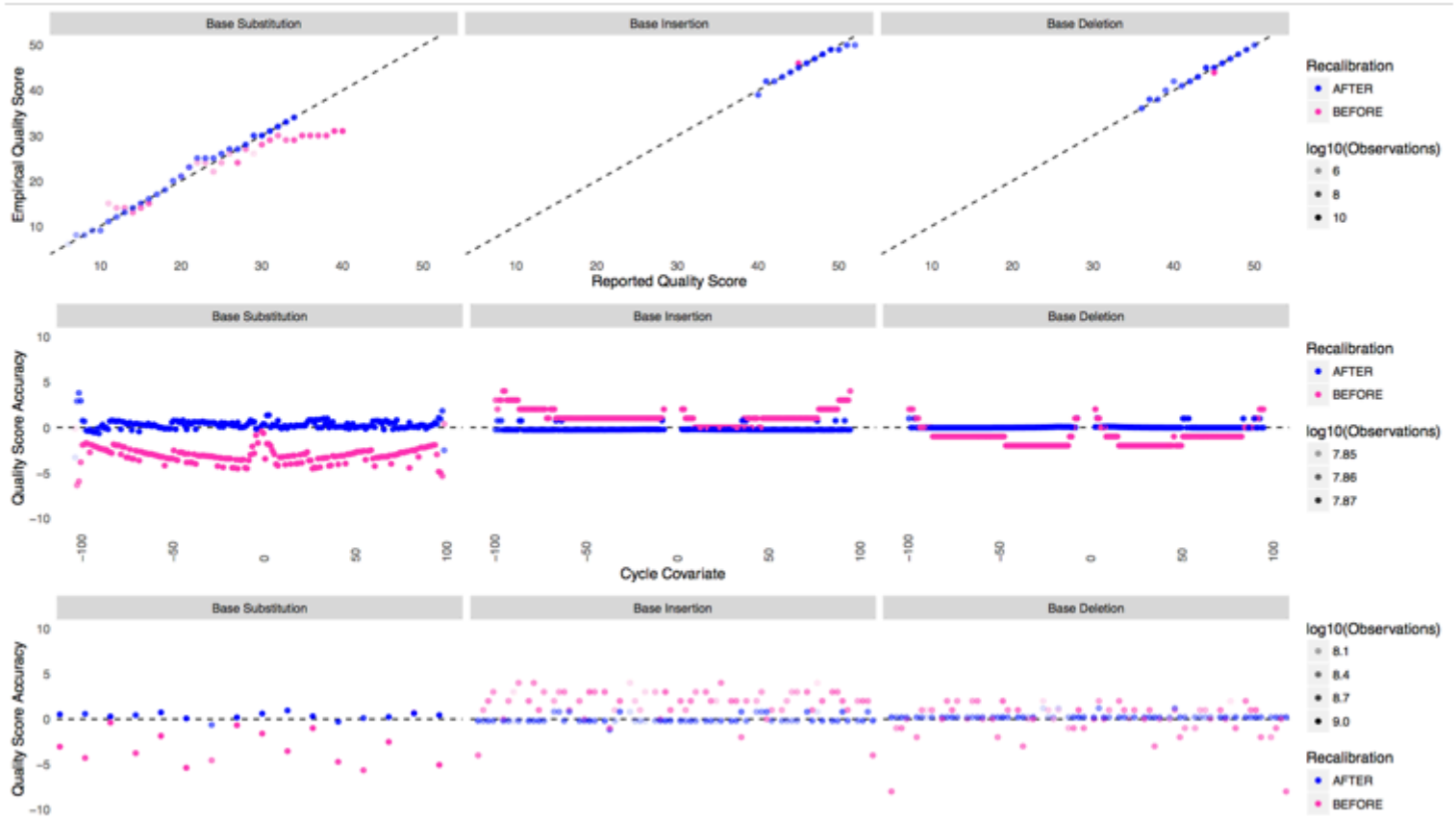


Figure 3.8. BSQR recalibration report for sample BB001. The base quality scores are shown before (pink) and after (blue) recalibration. The recalibrated scores now appear to follow the expected trends for empirical quality (top), quality score accuracy (middle) and quality score accuracy for the 16 dinucleotides (bottom).

Ti/Tv ratio follows the expected trend, with the highest specificity and low sensitivity variants having the greatest ratio at 2.56. This is close to the expected ratio for whole-exome sequencing data. Variants in the selected tranche had a Ti/Tv ratio of 1.89 prior to filtering, which is extremely low for exome sequencing. This includes singleton variants, which could consist of a high number of novel variants that are likely to be classified as false positives.

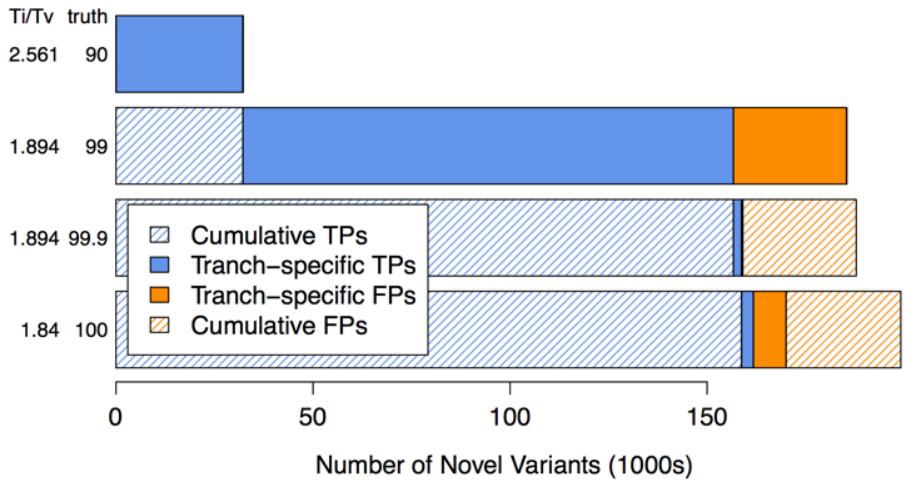


Figure 3.9. VQSR plot of the four tranches tested. With each tranche of variants, although the number of true positives increases, the number of false positives is also more prominent. A balance needs to be obtained to minimise the likelihood of false positive findings. Therefore the 99% tranche was selected.

3.4.5. Quality checks

Ti/Tv ratio for all variants was calculated as 2.09 by VCFtools after filtering, this ratio was corroborated by GATK suggesting this is reliable. Table 3.3 lists the number of known and novel variants by type and singletons identified. The ratio for known variants is higher at 2.58, which is close to the expected values for exome sequencing of 2.8-3.0, although this is still indicative of false positives. The ratio is significantly lower for the 15% of variants deemed to be novel at 0.79, which implies a high number of false positives.

Table 3.3. Variant metrics generated by the GATK VariantEval tool. Known and novel variants are classified by type. Ti/Tv ratio was calculated for known variants only and found to be closer to the expected value for exome sequencing, which indicates that there are likely to be many false positives among the novel variants.

Variants	*Number of Variants*	*SNPs*	*Insertions*	*Deletions*	*Singletons*	*Ti/Tv Ratio*
Total	325,344	301,460	7,044	12,388	131,177	2.09
Known	272,435	254,131	5,468	9,295	92,580	2.58
Novel	52,909	47,329	1,576	3,093	38,597	0.79

Average sequencing depth was calculated per sample and ranged between 39X and 126X coverage, averaging 51X. This was also calculated per chromosome for all samples and illustrated in Figure 3.10. The plot shows that coverage was on average greater than 45X across the genome, with some chromosomes exhibiting higher coverage than others. The filtered VCF file was used to compute this, therefore all base calls are above Q30, indicating quality coverage to obtain reliable variant calls. Sequencing metrics show that an average of 99.8% reads aligned to the reference sequence across samples, demonstrating high base calling accuracy from the sequencer. However, this was reduced to 85.9% of reads aligning within the 51.1Mb of SureSelect target intervals, meaning 14.1% of reads were likely to have aligned in non-target regions.

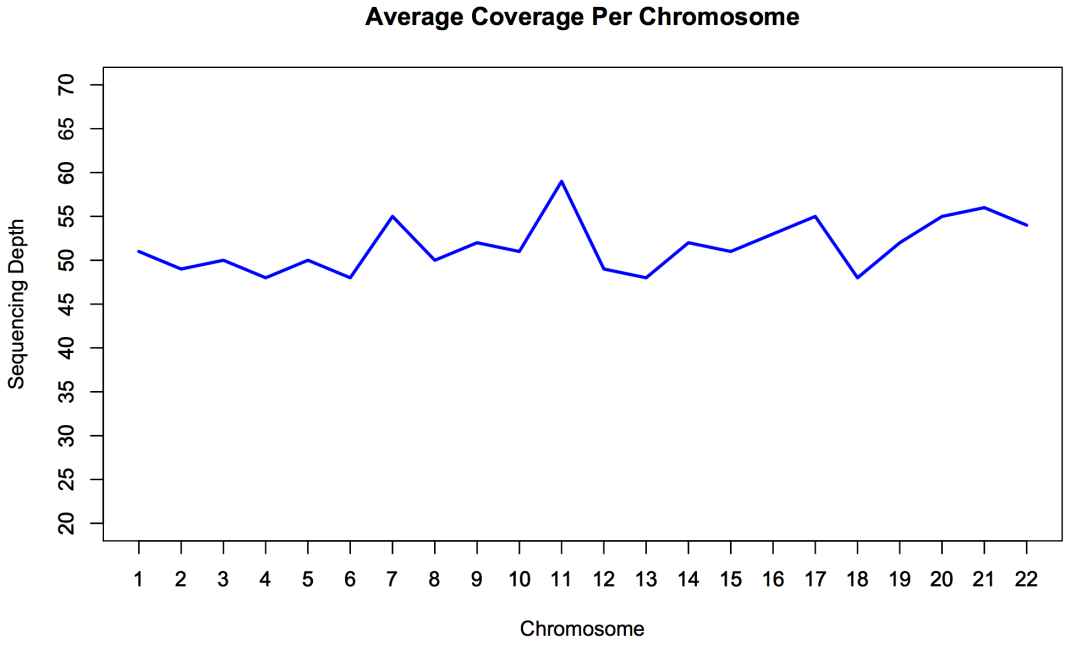
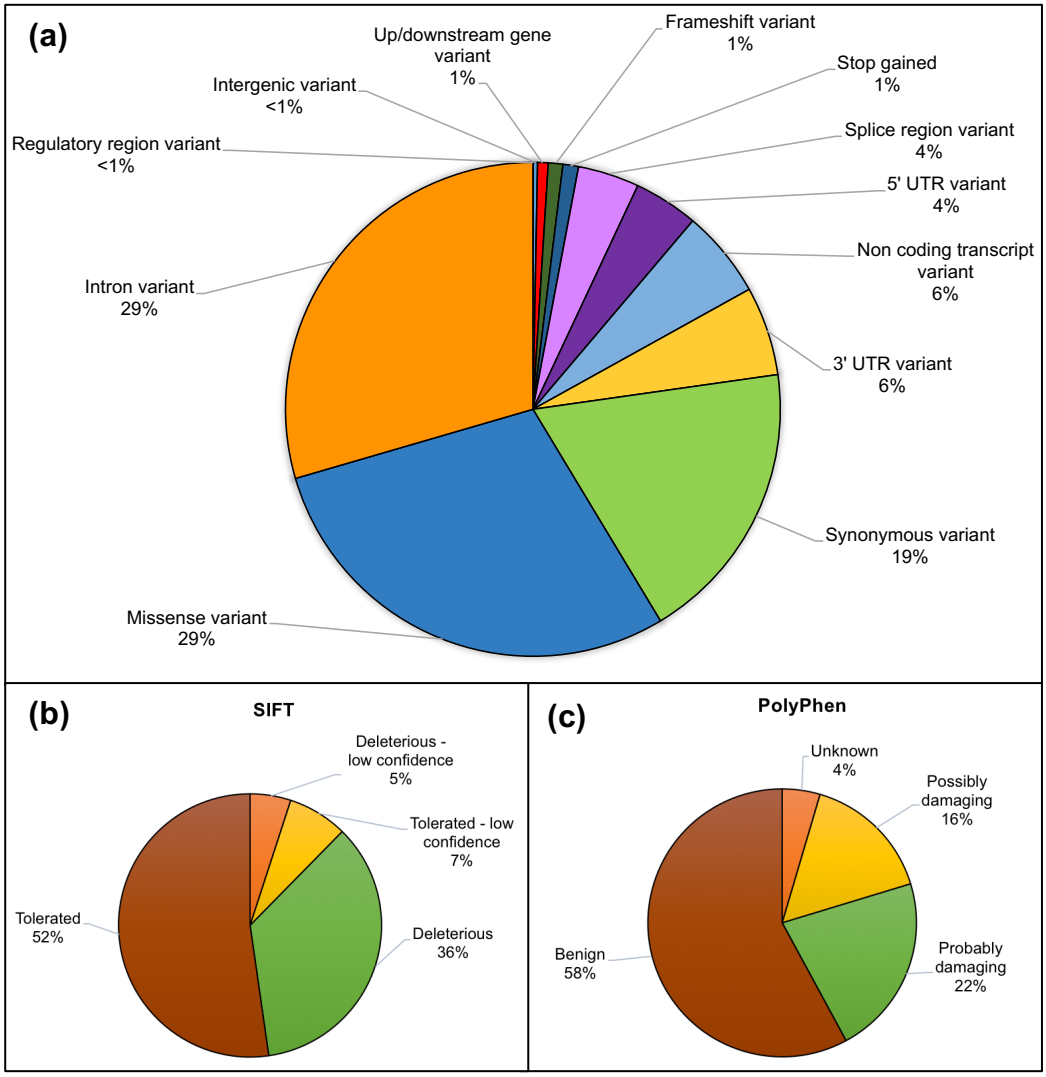


Figure 3.10. Depth of coverage plot calculated from VCFtools. The sequencing depth of coverage per chromosome was determined to ensure that there was adequate depth to make variant calls with confidence. The coverage was above the generally expected threshold of 30X for exome sequencing.

3.4.6. Functional annotation

VEP generated a summary report containing information about the variant types in the data; Figure 3.11 (a) illustrates the breakdown of these variant types. Although the majority were coding mutations, 29%



(d)

Variant Impact	Percentage
HIGH	2.2%
LOW	10.4%
MODERATE	11.1%
MODIFIER	76.2%

Figure 3.11. Summary of variant annotations from Variant Effect Predictor and SnpEff. (a) Variants were classified by consequences using curated databases from the 1000 Genomes Project and Exome Sequencing Project. (b) SIFT protein coding predictions for exonic variants indicated that most protein altering variants were tolerated with high confidence. (c) PolyPhen showed similar results, with the majority of variants predicted as benign. Damaging consequences are further separated into possibly or probably damaging, with a relatively even number of both. (d) The percentage of variants classified by impact on the protein by SnpEff. Few variants had high impact, however a significant majority were predicted as modifiers, likely to be non-coding variants.

of variants in the WES data were located in intronic regions. Other variants such as splice site and UTR variants were also annotated, possibly due to spurious capture. SIFT and PolyPhen consequence predictions (b) are quite similar to each other, with a similar number of variants predicted to be deleterious and possibly or probably damaging by SIFT and PolyPhen, respectively. The impact of variants on protein integrity (c) was also predicted by the programs. Only 2% were predicted to have a high impact on protein coding, with 11% exerting moderate impact. The largest fraction was represented by variants that showed modifier activity.

3.5. Discussion

The advent of NGS technologies has allowed large-scale sequencing of the genome at single-base resolution. Whole-exome sequencing is a time and cost-effective alternative to the whole-genome approach. It is being used to explore the role of rare variants in the heritability of complex genetic diseases (Ng *et al.,* 2010). Thus, high quality sequencing data is necessary to obtain reliable variant calls if one is to investigate the association of rare variants with a disease state.

3.5.1. Quality control

The average error rate for Illumina sequencing data ranges between 0.1-0.5% (Mardis, 2013). FastQC assumes the library should have uniform coverage across the genome, therefore it is likely that some metrics may fail due to the expected biases of targeted sequencing. Quality checks performed on the raw data indicated that the majority of measures were good or satisfactory for all samples. All quality scores were greater than Q30 suggesting high quality sequence data. A minority of reads showed a reduction in quality near the end of the sequence, which is common for Illumina sequencing and is a caveat of the sequencing method (Fuller *et al.,* 2009). Paired-end reads account for this by sequencing from the opposite end to obtain quality paired information. The GC content for some samples shows a shift in distribution ranging between 40-60%, which can indicate systematic

bias independent of base position. FastQC models a normal distribution based on the read data provided. FastQC is suitable for examining the quality of NGS data, however it is difficult to assess large numbers of samples. MultiQC concatenates this data to allow a better comparison of data quality across all samples to pinpoint outliers.

Although MultiQC did not output a merged graph of *k*-mers for all samples, this would not have been informative. However, observation of the individual FastQC reports showed that several samples failed based on the *k*-mer content. The *k*-mers likely correspond to segments of the Illumina adaptor sequences, however it is uncertain why they would be incorporated into the middle of the reads. When noting the counts of these *k*-mers, they amount to less than 10,000 occurrences in the exome, affecting only a small fraction of reads at around 85 million. Additionally, the program computes this metric using only 20% of the whole library and extrapolates to the remaining sequence for efficiency (Andrews, 2010). Some of the *k*-mers also correspond to Illumina adapter sequences, indicating adapter contamination. However, this contamination is insignificant as adapter content was very low and therefore unlikely to affect downstream analyses. Furthermore, later steps involved removal of duplicate sequences and recalibration of base quality scores, ensuring only the highest quality reads were retained.

Sequence duplication levels highlighted duplication at 10X the level, seen in around 5-10% of the library, however this was not flagged as a failure. This could be attributed to the targeted nature of the library or a result of PCR duplicates as FastQC assumes that the data has uniform coverage. The Illumina library preparation also includes several amplification steps, which can introduce sequencing errors and increase background. However, marking of duplicates by Picard can reduce these duplication levels. QC highlighted no major issues in the raw data that could inhibit downstream analyses due to sequencing errors. Generally, any low-quality information was removed when filtering for high quality bases above Q30.

3.5.2. Alignment and QC

Sequences were aligned to the hg19 reference genome as opposed to the current build hg38, which is a more accurate representation of the human genome. This was because hg19 is a complete build, whereas at the time of writing this thesis, hg38 is still undergoing frequent minor mapping updates and therefore many tools and databases do not currently support it. FastQC results indicate that some improvements were made after alignment, de-duplication and recalibration of reads. A 5% increase in sequences at 10X duplication or above was observed following processing of aligned data. This exhibits a low level of duplication as the majority of reads are unique within the library in both the raw and de-duplicated sequences. Therefore, library sampling is not near saturation and thus sequences are appropriately diverse. A large spike in duplicated sequences at a high level would mean low diversity in the original library, commonly due to PCR over-amplification. However, this is not the case here as there are not a large number of overrepresented sequences to suggest that the library is compromised.

Removal of duplicate sequences is intended to reduce noise and minimise false positive variant calls downstream, however it has little impact on the results if amplification bias is not an issue (Ebbert *et al.,* 2016). GC content was not altered after alignment and averages between 40-60% across samples, which is close to the expected GC content of 40-50% for human chromosomes (Guo *et al.,* 2014). Some samples are skewed towards a higher GC content nearing 60%; as this does not deviate more than 10% of the normal range, contamination is not assumed. Although recalibration reduced the overall base quality scores, they are more reliable as they are consistent with the empirical values as shown by BQSR results in Figure 3.8 (Bao *et al.,* 2014). Furthermore, FastQC only calculates quality metrics on a subset of the reads, extrapolating these results to the rest of the dataset. Therefore, it is necessary to look at the results in conjunction to determine whether there is an actual bias that may affect downstream analyses or low level biases that should be heeded as warnings.

The presence of *k*-mers in the middle of reads for this sample appears to be resolved, which suggests that alignment and de-duplication steps may have eliminated these reads from later steps. Some *k*-mers are still observed towards the beginning and end of reads, which could be adapter contamination. As this is not observed at high levels, it does not warrant trimming bases as this could remove important information. As most algorithms account for base quality when calling variants, an additional trimming step was not performed. However, poor mapping could increase the number of false positive and negative calls (Pfeifer, 2016). SamStat and Picard summary statistics supported an average mapping rate of 99.8% for reads with Q30 or above, demonstrating high quality alignment so this was not an issue.

3.5.3. BQSR

Recalibration of base quality scores is a necessary step to ensure base calls are accurate for variant calling. BQSR detects systematic, technical errors made by the sequencer when assigning quality scores to bases. The plots generated during BQSR clearly illustrate the improvement in base quality scores following recalibration in Figure 3.9. Accuracy is calculated from the empirical quality score minus reported quality score, therefore no difference means it is well calibrated. Variant calling algorithms rely on the quality scores assigned to individual bases in the sequence as this determines how reliable the observation of a variant call would be at a particular locus (Bao *et al.,* 2014). The plots show that the difference between the reported and empirical scores is close to zero compared to the raw base qualities, indicating significantly improved base quality scores. By adjusting for changes in quality by cycle and sequence context, true high-quality base calls can be obtained.

3.5.4. Variant calling and recalibration

Variants called by HaplotypeCaller were evaluated to distinguish the true variant calls from the false positives. The joint genotyping step is effective due to increased sensitivity in calling low frequency variants; although this can affect the number of singleton variants correctly

identified from noise. The average error rate for Illumina sequencing is between 0.1-0.4% (Quail *et al.,* 2012); increased sequencing depth could improve the reliability of such rare singleton calls and reduce random errors. Metrics showed that aligned base quality was high after processing of BAM files, essential for high quality variant calls.

Variant quality scores were recalibrated using validated databases of known variants to identify the annotation profiles of “true” variants. Establised databases include 1000 Genomes and HapMap which are curated, unlike dbSNP which allows users to add their own variants. As a result, the confidence that dbSNP represents true variants is low. A compromise between specificity and sensitivity is required to minimise the number of false positive variant calls in the dataset. The first tranche of 90% has the lowest value of truth sensitivity but has the highest value of novel Ti/Tv, making it very specific. Each subsequent tranche introduces additional true positive calls but also increases in false positives. The tranche containing 99% of the known variants was selected as the most appropriate call set, since it had a greater number of “true” variants and minimised false positives than higher tranches. The Ti/Tv ratio for the selected variants is much lower than the expected ratio for human exomes, which is between 2.8-3.0 (DePristo *et al.,* 2011). A lower ratio is indicative of false positive calls in the dataset; however, this is calculated on novel variants only which is expected to include a greater number of false positives.

3.5.5. Quality checks

Transitions occur more frequently presumably because the exome is under greater selective pressure against missense mutations. They are more tolerated as they usually affect the third base in a codon, largely responsible for the degeneracy of the genetic code. Therefore they often encode for the same amino acid and are likely to persist as ‘silent’ mutations in the genome (DePristo *et al.,* 2011). On the other hand, transversions generally have a more profound effect due to changes to the ring structure. Agilent SureSelect technology has an expected Ti/Tv ratio of around 2.26 on the target regions; this is low possibly because it

also captures some non-exonic regions such as UTRs (Chilamakuri *et al.,* 2014). When Ti/Tv ratio was calculated separately for known and novel on-target variants after filtering, the known fraction had an increased ratio of 2.58 which is greater than the expected 2.26. This implies that the known variant call set consists of good quality variants. For novel variants, Ti/Tv decreased to 0.79, indicating that the majority are false positives possibly due to random sequencing errors. The combined ratio was reliably 2.09 as determined by several tools, lower than the value calculated for Agilent technology due to the novel fraction. This comprises a high number of singleton variants (72.9%), which is likely to account for the reduced Ti/Tv ratio. Thus, it was determined that discarding singleton variants would improve this ratio as in a small dataset such as this, they will not provide any valuable association information.

SamStat showed in Figure 3.8 that almost all reads aligned to the reference with Q30 or above. Agilent technology has been shown to capture 99.8% of baits (Chilamakuri *et al.,* 2014). However, alignment statistics show that almost one sixth of reads were mapping to off-target baits in non-exonic regions. This would explain the low Ti/Tv ratio prior to filtering with target regions as whole-genome regions would exhibit a ratio closer to 2. Depth of coverage consistently averaged more than 45X across chromosomes, with the lowest per sample being 39X up to 126X. More than 91% of bases achieved at least 30X coverage. This aligned read depth was more than adequate for whole-exome sequencing data. The greater sequencing depth acquired, the more reliable variant calls can be made.

3.5.6. Functional annotation

The utility of using multiple annotation software is that consensus in annotations mean that a variant is more likely to be real, especially as not all tools use the same databases. However, there are still difficulties interpreting the data since annotation programs will only be as accurate as the databases providing the information (McCarthy *et al.,* 2014). Although WES primarily targets exons, non-coding regions can also be

sequenced as a by-product (Guo *et al.,* 2012). A large number of variants were intronic despite using a targeted exome library. Filtering against the target intervals should reduce the number of non-coding variants included. However, the Agilent SureSelect All Exon library includes flanking regions around the exonic baits. It has been demonstrated the WES can generate high quality data outside of target regions and this should not necessarily be discarded, as it could still be valuable (Guo *et al.,* 2012).

While it was traditionally believed that introns were not as important as exons, many studies have now established the functional significance of intronic regions (Guo *et al.,* 2012). It is now well known that SNPs located in *cis*-regulatory regions may influence the regulation of gene expression. Although a captured library is greatly enriched for the target exonic regions, a significant fraction of DNA fragments still fall outside these regions. The fraction varies depending on capture efficiency (Ng *et al.,* 2010). As a result, a significant fraction of intronic and intergenic regions may have been sequenced, which may include promoters, enhancers, conserved non-coding sequences, untranslated regions, miRNA sites, and other potentially functional regions. While it is common practice to ignore data outside target regions and focus only on the bases inside target regions and while justified, this approach is inefficient as it overlooks a large amount of data that would otherwise be useful in genetics studies (Guo *et al.,* 2012). It is also important to verify annotations with additional prediction tools to establish a consensus across databases prior to experimental investigation for any variants of interest as predicted functional consequences do not necessarily translate to causality.

3.5.7. Conclusion

Exome-sequencing has proven invaluable for identifying novel variants in genetic loci which are associated with Mendelian and complex diseases. It is particularly useful for investigating rare variants that can go undetected by less sensitive methods such as genotyping arrays like GWAS. Quality was assessed at several stages during the alignment

and processing of sequencing data. Adjustment for random errors in sequencing, artefacts and systematic biases and filtering on various quality metrics was performed to help reduce the number of false positive variants in the final call set. Variant annotation assigned predicted consequences using established databases. Further filtering was necessary to prioritise potential disease-causing mutations by retaining variants in coding regions. Given the large number of variants remaining, the next step will be to identify any significant associations with AD. Following multiple QC steps, we can conclude that the data is of a high quality, sufficient to proceed with further analysis. The variant data is also likely to be more accurate as a result.

Chapter 4. Searching for causal variants in Alzheimer's disease

4.1. Introduction

Some of the work in this chapter has been undertaken by MSc students and other members of the group. Kirsty Williams, George McConnell, Dr Keeley J. Brookes and Dr Tamar Guetta-Baranes performed the KASP assays to genotype the GWAS SNPs.

4.1.1. Genetics of Alzheimer's disease

To date, over 30 genetic variants have been associated with the risk of developing AD from GWAS, NGS and SNP genotyping arrays. GWAS identified 20 new susceptibility loci which provided a number of candidate genes for investigation (Harold *et al.,* 2009; Hollingworth *et al.,* 2011; Naj *et al.,* 2011; Seshadri *et al.,* 2010b). This also implicated several new pathways in AD pathogenesis, such as lipid transport, immune function and inflammation and endocytosis (Lambert *et al.,* 2013). Despite the identification of new disease pathways, the functional relevance of many of the genes also remains elusive. Although highly replicable in Caucasian groups, these effects have been difficult to replicate in other populations (Bertram *et al.,* 2010). These common risk variants only exert small effects on disease risk and therefore do not account for much of the missing heritability in AD. They are located in non-coding regions of the genome; a number of these are intergenic but the majority of risk variants lie within introns of genes. The intergenic variants are often large distances from the associated gene, therefore the level of linkage disequilibrium (LD) with the index SNP can reach as low as r^2=0.2 to incorporate the gene (Brookes and Morgan, 2017). It is likely that these highly associated index SNPs tagged by GWAS are in LD with the true causal variants, however they are too rare to be detected using this approach (Bettens *et al.,* 2013).

4.1.2. Sequencing studies in Alzheimer's disease

NGS has been employed to address the limitations of GWAS and elucidate the causes of the missing heritability in AD. It provides single-base resolution such that novel SNPs and indels can be identified. Increasing evidence suggests that low frequency, pathogenic variants could have greater effect sizes and therefore contribute more to this missing heritability (Rivas *et al.,* 2011). Complex diseases such as AD are likely to have a polygenic architecture with a combination of rare and common variants playing a role. WES is a powerful method of surveying the entire protein-coding region of the genome with the ability to uncover rare variants not detected by GWAS. The application of WES on well-characterised samples may prove very useful in delineating the genetic basis of AD by segregating variants with phenotype.

Targeted resequencing of GWAS genes has uncovered rare causative variants, highlighting the contribution of rare variants to AD risk (Tosto and Reitz, 2013). The genes *CLU*, *SORL1* and *ABCA7* have been found to contain rare variants associated with AD risk that were overlooked by GWAS. Exome sequencing has revealed a number of rare *SORL1* mutations in patients with familial EOAD and sporadic LOAD (Pottier *et al.,* 2012; Rogaeva *et al.,* 2007), however they require large sample sets to validate the observed associations. Loss of function mutations in *ABCA7* were associated with AD risk in the Icelandic population and replicated in European and Northern American groups (Steinberg *et al.,* 2015). Multiple rare variants in *CLU* were found to show significant association with AD risk independent of the common variant identified by GWAS (Bettens *et al.,* 2012). These findings suggest that both rare and common risk variants within a gene can play a role in disease susceptibility (Bettens *et al.,* 2013).

The biggest finding from NGS to date, however, is the *TREM2* variant rs75932628 which causes an arginine to histidine substitution (p.R47H). Discovered simultaneously by two independent studies using NGS, it

was shown to increase risk of developing AD by at least 3-fold (Guerreiro, Wojtas, *et al.,* 2013; Jonsson *et al.,* 2013). The missense mutation has been consistently replicated in Caucasian groups but is not observed in Asian and African American populations. However, other rare variants within *TREM2* and neighbouring *TREM* receptor genes, *TREM1* and *TREML2* have also been found to associate with LOAD risk (Benitez *et al.,* 2014; Cheng *et al.,* 2016; Jin *et al.,* 2014). This has emphasised the importance of inflammation and the immune pathway in AD pathogenesis.

4.1.3. Exome-wide association and burden

NGS studies in AD have generally focused on screening for variants in known candidate genes. WES allows rare variants to be identified and tested for association with disease traits. Detecting rare variant association with disease can be confounded by issues such as small sample size, low coverage and allele frequency. Aggregation-based tests calculate genetic burden using the cumulative frequency of all rare variants within a gene in case-control studies, providing increased power to detect associations. It can be difficult to distinguish the direction of effect when testing variants together (Sleegers *et al.,* 2016). Sequence Kernel Association Test (SKAT-O) is a combined burden test that calculates association between SNP sets and a phenotype. It accounts for the directionality of effects by weighting all variants and can adjust for covariates (Lee *et al.,* 2012).

The limitations of both GWAS and NGS are likely to ultimately be addressed by advances in custom array technology. GWAS is a tool used to explore common variants with large sample sizes and NGS can detect rare variants with modest number of samples due to cost implications (Medway and Morgan, 2014). Rarer variants discovered via sequencing are beginning to be included on custom arrays. This will provide an easier approach to summating the effects of rare and common variants on a phenotype.

4.1.4. Polygenic risk scores

Presence of the *APOE* ε4 allele remains the largest known genetic risk factor for LOAD, increasing risk by 3-fold in carriers and more than 12-fold in individuals homozygous for the ε4 allele (Farrer, 1997). Polygenic risk scores (PRS) have been increasingly used to investigate the effect of multiple genetic variants on disease traits. It is based on the notion that many variants with small effects will not be detectable at genome-wide significance, however as a collective, they may exert a strong effect (Wray *et al.,* 2008). AD is a complex disorder with a large genetic component; this polygenic architecture is likely to be made up of multiple variants with varying effect sizes. A combination of rare and common gene variants could have a more pronounced effect on disease risk (Gibson, 2012).

PRS aim to predict disease risk based on the genetic profile of individuals. Polygenic risk scoring has previously been employed in AD and proven useful for predicting individual disease risk with 78.2% accuracy obtained using 87,605 SNPs from an imputed dataset, in clinically diagnosed AD samples (Escott-Price *et al.,* 2015). This was followed up with a study to determine the upper limit of predictive accuracy that could be achieved for the same samples. Escott-Price and colleagues calculated AUC_{max} as 82% using age, sex, disease prevalence and estimates for the heritability of liability as additional predictors (Escott-Price, Shoai, *et al.,* 2017). More recently they demonstrated that using post-mortem diagnosed AD samples can significantly improve predictive ability, achieving 84% accuracy in distinguishing cases and controls when including covariates (Escott-Price, Myers, *et al.,* 2017).

The utility of using post-mortem confirmed samples was supported by another study investigating the genetic overlap between sEOAD and LOAD (Chaudhury *et al.,* 2017). A comparable predictive accuracy of 75.5% was obtained with a significantly smaller sample size and 9,434 SNPs, demonstrating the increased power that can be achieved by

using pathologically confirmed tissue over clinically defined samples. The best-fit model was able to explain 25.2% of the variance between sEOAD cases and controls, which supports a shared genetic aetiology with LOAD (Barber *et al.,* 2017). However, the number of variants common to both base and target datasets was low. It is possible that increasing SNP coverage with imputation could greatly improve prediction.

Other studies have used SNPs that have greater association with AD as a means to reduce the number of SNPs required for predicting risk (Desikan *et al.,* 2017; Lacour *et al.,* 2016; Marden *et al.,* 2016). Polygenic hazard scores (PHS) have been computed from as little as 22 SNPs, incorporating AD-associated GWAS SNPs and *APOE* ε genotypes (Marden *et al.,* 2016). Combined with *in vivo* CSF and PET biomarkers of amyloid and tau pathology, hazard scores can effectively predict longitudinal cognitive and clinical decline in cognitively normal or MCI individuals (Desikan *et al.,* 2017; Tan *et al.,* 2018). PRS have demonstrated their utility in predicting disease progression, particularly for stratifying individuals with a greater risk of conversion from MCI to AD (Lacour *et al.,* 2016). Identifying patients at high risk for developing AD could prove useful for early diagnosis and preventative measures in the future (Raghavan and Tosto, 2017).

PRSice is an R-based command line program that uses genetic information from two independent datasets to select SNPs and then test for association with a phenotype (Euesden *et al.,* 2014). The base dataset contains association data for the same or other traits of interest. These SNP effect scores are weighted to calculate each individual's risk scores in a target dataset. Logistic regression tests are performed at multiple significance thresholds to generate models of predictive accuracy. Other predictors such as age and sex are included to increase predictive power.

4.2. Aims

This chapter aims to investigate the association of variants with AD using NGS of LOAD case and control samples, neuropathologically confirmed by BDR. Based on the notion that most genetic variants leading to a severe phenotypic effect are located in the coding regions of the genome, an exome-wide approach will be employed focusing on coding variants only. We have 94% power to discover rare variants with frequency as low as 0.5% and although we achieved 83% power to detect associations of variants with MAF of 5%, they require an effect size of at least 3-fold. By combining multiple variants within genes to calculate genetic burden on AD, we can increase the power to detect associations of rarer variants. Finally, PRS will be generated on the full dataset to examine the association between multiple genetic markers and their collective effects on AD. This will include a greater proportion of coding variants compared to previous GWAS datasets that were used.

4.3. Methods

4.3.1. QC in VCFtools

WES was performed on a total of 292 samples, comprising 149 AD cases, 55 controls, 8 EOAD and 80 individuals with other dementia phenotypes; demographics are provided in Chapter 3. VCFtools was used to perform quality control checks before loading data into Plink-SEQ. The filtered VCF was assessed for missingness using the genotyping rates of individuals and variants. Individual genotyping rates were calculated to determine the optimal threshold that would retain a high call rate by excluding the least number of samples. This was selected as >97% and for SNP genotyping rate the cut-off was >98%. The Hardy-Weinberg equilibrium (HWE) states that allele and genotype frequencies in a population will remain constant over successive generations in the absence of other evolutionary influences. This test calculates the observed heterozygosity of all variants and the probability

that it significantly deviates from the expected allele frequency in controls. The significance threshold used to exclude markers failing HWE was p<0.001. Any samples or SNPs failing QC were removed as follows, with individuals selected for removal. Samples BM048, BM056 and BO118 were found to have discrepancies in their clinical information in the database and were therefore excluded from further analyses.

```
$ vcftools --vcf 8_bdr_annotated_recal_exonv4_filtered.vcf \
--missing-indv \
--out MissingnessIndividuals.txt

$ vcftools --vcf 8_bdr_annotated_recal_exonv4_filtered.vcf \
--remove-indv sample1 --remove-indv sample2 \
--max-missing 0.98 \
--hwe 0.001 \
--recode \
--out 9_bdr_annotated_recal_exonv4_filtered_vcfqc.vcf
```

Options are:

- --missing-indv: calculate genotype missingness for individuals
- --remove-indv: remove the named sample and all data from output
- --max-missing: maximum missing genotypes
- --hwe: Hardy-Weinberg equilibrium p-value cut-off
- --recode: convert output to plink map and ped format

4.3.2. Remove singletons

Singletons are variants only observed in one individual in a cohort. Due to their rarity, it can be difficult to distinguish between a true variant and artefacts due to sequencing error in small sample sizes. As they would provide little information in this study, singletons were removed from the dataset to reduce the rate of false positives. In this case, to exclude singleton variants the cut-off was a MAF = 0.002.

```
$ vcftools --vcf 9_bdr_annotated_recal_exonv4_filtered_vcfqc.vcf \
--maf 0.002 \
--recode \
--out 10_bdr_annotated_recal_exonv4_filtered_vcfqc_nosingletons.vcf
```

4.3.3. SnpEff annotation and filtering

Variants can then be filtered on numerous other annotations, such as functional consequence, genes of interest, allele frequency, chromosome and more. SnpSift requires each annotation filter individually, therefore all are passed to the tool using the "|" symbol which means "or". The variants were filtered for all known coding type mutations to investigate exonic variation.

```
$ java -Xmx256g -jar snpEff.jar -v -stats 10bdr.html GRCh37.75
10_bdr_recal_exonv4filtered_vcfqc_nosingletons.vcf >
11bdr_recal_exonv4filtered_vcfqc_nosingletons_snpeff.vcf

$ cat 11bdr_recal_exonv4filtered_vcfqc_nosingletons_snpeff.vcf | java
–Xmx256g –jar SnpSift.jar filter "(ANN[*].EFFECT has
'non_coding_exon_variant') | (ANN[*].EFFECT has 'exon_loss_variant')
| (ANN[*].EFFECT has 'frameshift_variant') | (ANN[*].EFFECT has
'stop_gained') | (ANN[*].EFFECT has 'stop_lost') | (ANN[*].EFFECT has
'start_lost') | (ANN[*].EFFECT has 'rare_amino_acid_variant') |
(ANN[*].EFFECT has 'missense_variant') | (ANN[*].EFFECT has
'inframe_insertion') | (ANN[*].EFFECT has
'disruptive_inframe_insertion') | (ANN[*].EFFECT has
'inframe_deletion') | (ANN[*].EFFECT has
'disruptive_inframe_deletion') | (ANN[*].EFFECT has
'stop_retained_variant') | (ANN[*].EFFECT has 'synonymous_variant') |
(ANN[*].EFFECT has 'stop_retained_variant') | (ANN[*].EFFECT has
'coding_sequence_variant')" >
12bdr_recal_exonv4filtered_vcfqc_nosingletons_snpeff_coding.vcf
```

Options are:

- Filter: perform filtering
- ANN: annotation

4.3.4. Convert to Plink format

The multiple sample VCF was converted to a format suitable for Plink v1.07 to enable downstream analyses using Vcftools. The VCF input was recoded to Plink map and ped files.

```
vcftools --vcf
12bdr_recal_exonv4filtered_vcfqc_nosingletons_snpeff_coding.vcf \
--plink \
--out 13_bdr_final_qc
```

Options are:

- --plink: recode output to plink format

4.3.5. Check sex

The sex of all individuals was assessed based on homozygosity rates of SNPs on chromosome X, calculated as an F-value. A male call was made if F>0.8 and a female call made if F<0.2. Comparisons were made to check for discordance with ascertained sex in the database, to highlight potential errors. If they did not match or if the SNP data was ambiguous regarding the sex, it yields a warning. The calculated sexes were validated using the BDR data genotyped on NeuroX chip version 2 in GenomeStudio. Discrepancies were resolved by manually examining the data for each discordant sample before updating the sex in either the database or excluding the sample altogether.

```
pseq bdr-exome i-stats --mask reg.req=chrX > SexCheck.pseq

plink --noweb --bfile 13_bdr_final_qc --check-sex --out
SexCheck.plink
```

Options are:

- --bfile: prefix of input binary plink files
- --sex-check: perform sex check calculations

4.3.6. PSEQ

A new project was created with the name 'bdr-exome' to load the VCF file into Plink-SEQ (PSEQ). The resources option provides the path to the resources archive, containing the necessary files for PSEQ reference databases; these include the hg19 reference sequence (SEQDB), dbSNP and 1000 Genomes VCF files (REFDB) and RefSeq gene transcripts (LOCDB).

```
pseq bdr-exome new-project --vcf
12bdr_recal_exonv4filtered_vcfqc_nosingletons_snpeff_coding.vcf
--resources /hg19/

pseq bdr-exome load-vcf
```

Options are:

- new-project: create a new project with name provided
- --resources: path to folder containing resources
- load-vcf: load the vcf file attached to project

The resource files were obtained from the appropriate sources and loaded into the database. The same hg19 reference sequence and dbSNP v149 VCF were used here, however the 1000 Genomes file used was the Phase 3 dataset as this latest version was available for PSEQ analysis.

```
pseq . ref-load --vcf dbsnp.149.vcf.gz --group dbsnp --refdb refdb --
show  XXX

pseq . ref-load --refdb refdb-g1k --group g1k --show AF --vcf
ALL.wgs.phase3_shapeit2_mvncall_integrated_v5b.20130502.sites.vcf.gz

pseq . seq-load --file hg19.fasta --seqdb seqdb --name hg19 --
description from-UCSC-20-dec-2010 --format build=hg19 repeat-
mode=lower

pseq . seq-summary --seqdb seqdb
```

Options are:

- --group: database group
- --refdb: reference database to attach file to
- --show: additional information to include from file
- --description: short description of file
- --format: genome build version and repeat masked sequence is lower case

A phenotype file comprises information about each sample including BDR ID, age, sex, number of *APOE* ε2 and ε4 alleles and other clinical information such as cognitive test scores and neuropathological staging. An example file format is shown in Figure 4.1, which was also loaded into the project.

```
pseq bdr-exome load-pheno --file AD.phe
```

4.3.7. Summary statistics in PSEQ

PSEQ was used to generate summary statistics for individual samples, single variants and also variants aggregated by gene within the dataset. A matrix of summary statistics is generated per individual, with information on number of variants called, singletons, genotyping rate, Ti/Tv ratio, mean quality, depth and more. Variant statistics include total number of variants called, mean allele counts and frequency, number of singletons, Ti/Tv ratio and number of variants passing each quality filter applied to the VCF.

When grouped by gene, similar information to the variant statistics is provided; additionally, gene size, number of exons and variants within each gene is included.

```
pseq bdr-exome i-stats > SummaryStats_samples.txt
pseq bdr-exome v-stats > SummaryStats_variants.txt
pseq bdr-exome g-stats --mask loc.group=refseq >
SummaryStats_genesRefSeq.txt
```

Column name; variable type; value if missing; description

```
##AD,Integer,-9,"AD Phenotype"
##SEX,Integer,-9,"Sex 1=male,2=female"
##Age_Onset,Integer,-9,"Age of Onset"
##Age_Death,Integer,-9,"Age of Death"
##PM_Delay,Integer,-9,"Post-Mortem Delay (Hours)"
##PM_Interval,Integer,-9,"Post-Mortem Interval"
##APOE_E4,Integer,-9,"APOE E4 Allele Count"
##DIN,Integer,-9,"DIN Score"

#ID        AD    SEX   Age_Onset  Age_Death  PM_Delay   PM_Interval      APOE_E4    DIN
BB001      1     1     -9         78         51.5       -9               0          9.6
BB003      0     1     -9         96         25.45      -9               0          9.2
BB004      1     1     -9         92         56.5       -9               1          8.6
BB005      2     2     -9         87         71.5       -9               1          8.8
```

Figure 4.1. Format of the phenotype file "AD.phe". Descriptions for each column of information is included at the beginning of the file preceded by double hashes, as shown above. The required information is as follows; column name must match that in the data table, variable type can be integer/float/string, value to denote missing data, description of the variable. This is followed by the sample ID, phenotype, sex and other covariate information similar to a covariates file for PLINK.

Options are:

- i-stats: individual statistics
- v-stats: variant statistics
- g-stats: gene statistics
- --mask: mask by selected annotation for locus group

4.3.8. Single variant association

A rudimentary single variant association with the AD phenotype was performed on all coding variants following the removal of singletons. This uses Fisher's exact test to perform allelic, dominant and recessive single variant tests, assuming a dichotomous phenotype. This is the alternative to standard Chi-squared (χ^2) for smaller sample sizes, which approximates probability whereas Fisher's test calculates empirical significance (Warner, 2013). It is a simple association test; therefore, it does not adjust for other variable factors that can affect disease risk.

```
pseq bdr-exome v-assoc --phenotype AD > SingleVariantAssociation.txt
```

Options are:

- v-assoc: variant association test
- --phenotype: the phenotype to test for association

```
plink --file 13_bdr_final_qc --assoc --ci 0.95 --out
bdr_association_qc
```

A Manhattan plot was generated in R to visualise the association results of variants with AD across the genome. The R package QQMAN v0.1.3 was used to create the plots. A large number of NA values in the association data causes errors to occur so variants for which the association could not be calculated were omitted from the plot. Manhattan plots were produced for comparison of the single variant association and logistic regression results with covariates included.

```
$ R
install.packages("knitr")
install.packages("devtools")
install_github("stephenturner/qqman")
library("knitr")
library(devtools)
library("qqman")
library("RColorBrewer")

bdr <- read.table("bdr_association_qc.assoc",header=TRUE,sep="\t")
bdr_results <- subset(bdr, select=c(SNP, CHR, BP, P))
bdr_results2 <- na.omit(bdr_results)

manhattan(bdr_results2, chr = "CHR", bp = "BP", p = "P", snp = "SNP",
col = c("steelblue", "skyblue2"), chrlabs = NULL, suggestiveline = -
log10(1e-05), genomewideline = -log10(5e-08), highlight = NULL, logp
= TRUE, annotatePval=0.00005, annotateTop=FALSE, cex=0.5,
main="Manhattan Plot of Exome Sequencing Data")
```

Options are:

- col: colour of results plot per chromosome
- chrlabs: chromosome labels
- suggestiveline: add suggestive p-value threshold
- genomewideline: add genome-wide corrected p-value threshold
- highlight: highlight specified variants
- logp: use $\log_{10}$ of p-values
- annotatePval: annotate variants above p-value threshold set
- annotateTop: annotate most significantly associated variant

4.3.9. Logistic regression analysis with covariates

Regression tests allow for single variant analyses with the inclusion of covariates; this is used to adjust for individual differences that may affect disease risk. *APOE* ε4 and age are major risk factors for AD and therefore correcting for covariates accounts for their effects on individuals so that only the association of variants with AD is calculated.

Logistic regression is the test of choice for dichotomous disease traits to compare cases and controls. For rare alleles, this test will be under powered and possibly unreliable, therefore a threshold of MAF>1% was applied. Using an additive model, this calculates the effect size of each individual variant on AD risk as an odds ratio, adjusting for the selected covariates. This accounts for other factors that influence the risk of developing AD. The covariates used were sex, number of *APOE* ε2 and ε4 alleles and calculated age at death. Other factors that may confound the data include cognitive scores on tests such as the MMSE, plus Braak, tangle and Lewy body stage and CERAD classification. Braak staging for AD ranges from 0 to 6, correlating with the spread of NFT formation in the brain during disease progression. CERAD defines disease severity as no AD, mild AD, moderate AD and severe definite AD, scoring from 0 to 3. These can be corrected for to minimise personal differences between individuals to focus on the genetics.

```
pseq bdr-exome glm --phenotype AD --covar Sex Cal_AAD
APOE_No_E4_alleles APOE_No_E2_alleles >
logisticRegression_Covariates.txt
```

Options are:

- glm: logistic regression model
- --covar: covariates to adjust for from file

4.3.10. Burden analysis

The power to detect rare variant association can be increased by aggregating sets of variants into groups and comparing their collective frequency between cases and controls, known as burden analysis. This examines the association of combinations of variants with a phenotype. The SKAT-O combined test was used to compute the burden for 14,971 genes containing 2 or more coding variants, to identify significant associations with AD. SKAT-O weights variants by effect size to

account for both protective and risk variants, allowing a truer measure of genetic burden.

PSEQ was used to generate a Set ID file for SKAT-O, containing the variants aggregated into genes. Any intergenic variants that were missed during filtering would also be excluded from burden analysis at this stage. The script Burden2SetID.py, available in Appendix 7.2, outputs information in the required format for SKAT-O, with the associated gene name and each variant listed per line.

```
pseq bdr-genes g-view --mask loc.group=refseq > PseqGenesList.txt

python Burden2SetID.py <PseqGenesList.txt> <SKATSetID.txt>
```

Options are:

- g-view: aggregate variants into genes

SKAT-O also permits adjustment for covariates, which improves precision of the results. The script BurdenAnalysisSKATO.R is provided in Appendix 7.3.

```
$ R BurdenAnalysisSKATO.R
```

4.3.11. Polygenic risk scoring

PRSice has automated the implementation of a standard polygenic scoring analysis pipeline using Plink v1.9 and R. It requires two datasets as input, a base dataset of GWAS results on a phenotype of interest and a target dataset containing full genotype information for the same or a different phenotype. The International Genomics of Alzheimer's Project (IGAP) phase I association summary statistics were used as the base dataset, collated from 17,008 LOAD cases and 37,154 controls with a total of 7,055,881 variants. The BDR WES dataset prior to removing singletons was used as the target dataset. Only variants found in both datasets were retained and ambiguous SNPs excluded. PRS are calculated as a sum of the trait-associated

alleles across variants, weighted by the effect sizes estimated from the base dataset. Scores were generated for individuals in the target dataset and correspond to the genetic risk of the base phenotype. Regression tests were then performed to test the association between PRS and the target phenotype, which can also indicate shared genetic aetiology between different phenotypes. Scores and regression models were calculated at a number of P-value thresholds to identify the best-fit model of prediction, displayed as plots generated by PRSice.

The *APOE* region is known to contain SNPs that affect the risk of developing AD and the presence of the ε4 allele significantly increases this risk. A region of 500kb around the locus was excluded from the target dataset to explore polygenic risk independent of the ε4 genotype; this was selected as chr19:45,160,844-45,660,844 (Chaudhury *et al.,* 2017). Individual genotypes for 19 GWAS SNPs were obtained through manual genotyping of SNPs using KASP and from Neurochip data for BDR samples. Additional predictors of AD risk, age, sex and number of *APOE* ε2 and ε4 alleles were included to increase predictive power of the regression models. The lower and upper significance thresholds used were 0 and 1, respectively, in increments of 0.01. Pairwise pruning of SNPs was performed with a window size of 50kb, in increments of 2kb, to remove SNPs in LD with $r^2 > 0.8$. PRSice outputs the risk scores for each individual at the best-fit significance threshold with the greatest predictive accuracy to explain the most variation (Euesden *et al.,* 2014).

```
$ plink --bfile 13_bdr_final_qc --noweb --exclude APOE_500kb.txt --
range --make-bed --out 13_bdr_final_qc_noAPOE

$ R --file=PRSice_v1.25.R --args \
base IGAP_stage_1.txt \
target 13_bdr_final_qc_noAPOE \
slower 0 \
supper 1 \
sinc 0.01 \
cleanup T \
prune.snps T \
```

```
prune.kb.r2 0.8 \
covary F \
best.threshold.on.bar T \
report.individual.scores T \
report.best.score.only F \
plink ./plink2 \
figname PRS_noAPOE
```

Options are:

- --file: PRSice R script file
- --args: arguments passed to the program
- base: base association summary dataset
- target: target dataset of genotype information
- slower: lower bound p-value threshold at with PRS is calculated
- supper: upper bound p-value threshold at with PRS is calculated
- sinc: increments to calculate PRS in between slower and supper
- cleanup: remove temporary files after analysis
- prune.snps: LD will be removed using pruning
- prune.kb.r2: pairwise LD threshold for pruning r^2
- covary: covariates file
- report.individual.scores: report all individual scores
- report.best.score.only: only report scores at the best threshold
- plink: path to plink v1.9 file
- figname: name of output figures

Multiple binary logistic regression tests were performed in SPSS using the polygenic scores at the best-fit threshold with a combination of age, sex, *APOE* ε2 and ε4 alleles and AD risk SNPs from GWAS as predictor variables. All 19 GWAS variants in the following risk genes were successfully genotyped; *ABCA7*, *BIN1*, *CASS4*, *CD2AP*, *CELF*, *CLU*, *CR1*, *EPHA1*, *FERMT2*, *HLA-DRB1*/*HLA-DRB5*, *INPP5D*, *MEF2C*, *MS4A6A*, *NME8*, *PICALM*, *PTK2B*, *SLC24A4-RIN3*, *SORL1* and *ZCWPW1*. *CD33* was excluded as it failed to replicate in the IGAP meta-analysis. Area under the curve (AUC) was computed to determine

the model with the greatest predictive accuracy, displayed using a receiver operating characteristic (ROC) curve. Nagelkerke's R^2 was calculated to compare the models for best fit, with a greater value indicating a better fit.

4.4. Results

4.4.1. Quality control

Variants were called on a total of 292 samples, which were recalibrated and filtered to ensure high quality genotypes were obtained. Further QC was performed to check samples for missing and discordant information. VCFtools excluded one individual, sample BO017, which failed to reach the individual genotyping rate threshold of 97%. A total of 15,174 variants were removed for low call rate and failing HWE. Heterozygosity rates on the X chromosome were calculated to determine the sex of each individual from the sequence data and compare against ascertained sex. The thresholds applied by PLINK were used, with the cut-off rate being $F<0.2$ for males and $F>0.8$ for females; however mismatched samples were checked individually before discarding any data. A small number of samples were found to have contradictory information regarding their sex from the sequence data. However, upon manual examination, the majority of these had homozygosity rates on the cut-off threshold and therefore the database sex was ascribed. A further three samples were excluded for disparity in their sex and also other genotyping data that did not correlate; BM048, BM056 and BO118. This resulted in 310,170 variants in 288 individuals remaining for analysis.

Singletons were removed in VCFtools by excluding all variants below a certain MAF. This was calculated to be 0.002 using the formula below, where N represents the number of samples. The overall on-target Ti/Tv ratio improved following the removal of singleton variants, increasing to 3.2, which is more reflective of the expected range for exome-sequencing data. The total number of variants remaining in coding

regions was 79,097 with a genotyping rate of 99.85%.

$$\mathrm{MAF} = {}^{1}/_{(2 \times \mathrm{N})}$$

4.4.2. Exome-wide association

Exome-wide single variant association was performed on the 79,097 non-singleton coding region variants. The final cohort consisted of 148 LOAD cases and 55 control samples for analysis. The power of this study to detect rare variant association with AD requires high effect sizes, as shown in Figure 2.5 in the General Methods. The Manhattan plot in Figure 4.2 illustrates the degree of association for coding variants with AD across the genome. Only one SNP reached the genome-wide significance threshold of $P=5x10^{-8}$ prior to adjustment for covariates. *APOE* SNP rs429358 was highly correlated with AD ($P=1.1x10^{-10}$, OR=7.4), which was expected, as this is an established risk factor. *APOE* SNP rs7412 was also observed with nominally significant association ($P=4.8x10^{-4}$, OR=0.22), exhibiting a protective effect against AD. This SNP generally indicates the presence of the ε2 allele. Table 4.1 contains information for variants associated with P<0.002. No other signals reached genome-wide significance at $P=5x10^{-8}$ or the suggestive threshold of $P=1x10^{-5}$, however with a growing sample size the power to detect significant associations is expected to increase.

Whole-exome analysis identified many tentative signals that lie just below the suggestive threshold, which could still play an important role in disease risk. Tentative single variant and burden associations were seen in several genes with developmental or signalling activity, such as *PILRA*, *ARSE*, *LRBA*, *VWC2* and *WT1*. *PILRA* SNP rs2405442 is a synonymous mutation yet exhibits a large protective effect ($P=6.3x10^{-5}$, OR=0.39). It is more frequently observed in control samples than cases. This gene has been linked to AD through *ZCWPW1*, which was previously highlighted as a risk gene in the GWAS meta-analysis. The SNP rs3764645 in known AD risk gene *ABCA7* was also associated

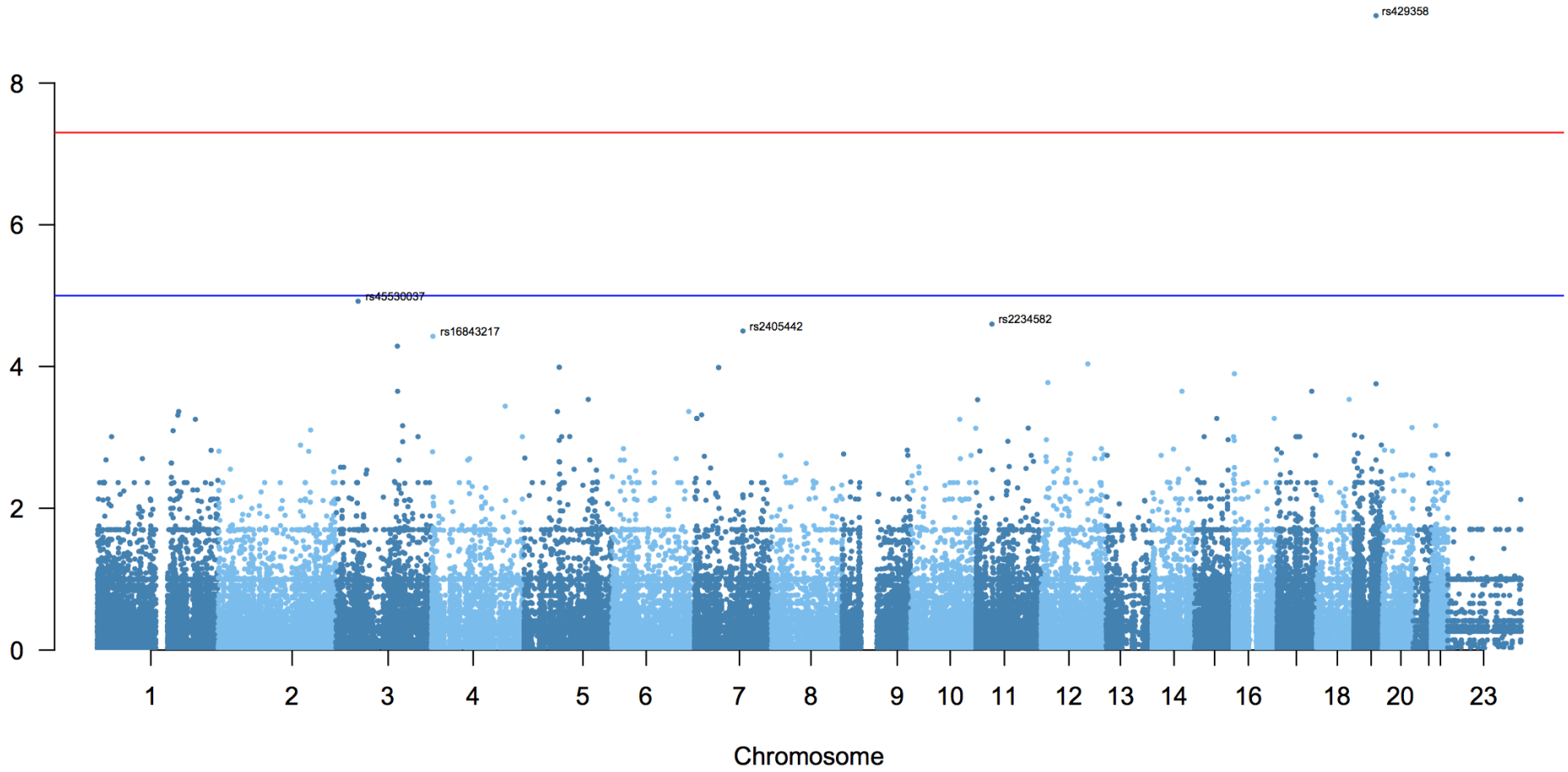


Figure 4.2. Manhattan plot of single variant association data. The genome-wide ($P=5x10^{-8}$) and suggestive significance ($P=1x10^{-5}$) thresholds are shown, indicated by the red and blue lines, respectively. Only one SNP surpasses both thresholds and this is the *APOE* risk SNP rs429358 which is known to significantly increase the risk of developing AD. Several other variants appear just below the suggestive significance threshold prior to correction for covariates.

Table 4.1. Single variant association analysis prior to correction for age, sex and other factors. Results are shown here for variants associated with AD ($P<2x10^{-3}$). Only *APOE* risk variant rs429358 reached genome-wide significance ($P=5x10^{-8}$). Other SNPs were close to the suggestive significance threshold ($P=1x10^{-5}$). However, tentative signals were present. Additionally, two variants exhibited very large effect sizes due to no minor alleles being present in control samples. It is possible that they are false positives, therefore the variants need to be considered with caution.

Variant	Gene	Position	Allele	MAF	MINU	Affected Counts (WT/Het/Mut)	Unaffected Counts (WT/Het/Mut)	P-value	Odds Ratio
rs429358	*APOE*	chr19:45411941	T>C	0.311	9	58/60/28	46/9/0	1.13E-10	7.4
rs1748	*TSC2*	chr16:2138269	T>C	0.185	7	88/52/8	48/7/0	4.90E-05	4.39
rs2234582	*WT1*	chr11:32456694	C>A	0.251	44	95/48/5	21/24/10	5.44E-05	0.37
rs2405442	*PILRA*	chr7:99971313	T>C	0.385	52	6/64/78	11/30/14	6.30E-05	3.19
rs45530037	*CCR9*	chr3:45942554	A>G	0.017	7	148/0/0	48/7/0	9.29E-05	0.02
rs17851048	*NTN4*	chr12:96077312	C>T	0.276	46	88/54/6	16/32/7	1.59E-04	0.4
rs16843217	*FAM193A*	chr4:2698230	G>A	0.042	12	143/5/0	44/10/1	1.66E-04	0.14
rs637450	*TMEM171*	chr5:72419456	C>G	0.261	44	94/46/8	17/32/6	1.99E-04	0.4
rs636926	*TMEM171*	chr5:72419617	C>A	0.261	44	94/46/8	17/32/6	1.99E-04	0.4
rs1806191	*GRIN2B*	chr12:13716638	G>A	0.52	36	39/81/28	4/28/23	2.10E-04	0.42
rs61753147	*PIK3R5*	chr17:8809025	G>A	0.062	0	123/25/0	55/0/0	3.50E-04	20.76
rs7412	*APOE*	chr19:45412079	C>T	0.057	14	139/9/0	42/12/1	4.76E-04	0.22
rs1997040	*DGKB*	chr7:14758273	C>T	0.485	41	47/74/27	8/25/22	5.28E-04	0.45
rs190533743	*SLC22A18AS*	chr11:2909513	C>G	0.108	22	126/22/0	33/22/0	5.75E-04	0.32
rs73347631	*VWC2*	chr7:49815474	C>T	0.02	7	147/1/0	48/7/0	5.75E-04	0.05
rs61390655	*VWC2*	chr7:49815493	C>G	0.02	7	147/1/0	48/7/0	5.75E-04	0.05

Variant	Gene	Position	Allele	MAF	MINU	Affected Counts (WT/Het/Mut)	Unaffected Counts (WT/Het/Mut)	P-value	Odds Ratio
rs1052067	*PMF1*	chr1:156206121	G>A	0.266	16	69/66/13	41/12/2	6.28E-04	2.65
rs1129998	*LRBA*	chr4:151719263	A>G	0.217	37	106/33/9	23/27/5	6.36E-04	0.41
rs10747057	*PWWP2B*	chr10:134218296	C>G	0.463	44	51/72/25	9/26/20	7.95E-04	0.47
rs4891392	*RTTN*	chr18:67718688	A>G	0.953	12	0/7/141	1/10/44	8.01E-04	0.2
rs40274	*SOWAHA*	chr5:132149684	G>C	0.953	12	0/7/141	0/12/43	8.01E-04	0.9
rs5761268	*MYO18B*	chr22:26239850	C>A	0.7	47	8/59/81	11/25/19	9.67E-04	0.45
rs114729571	*C3orf14*	chr3:62319048	A>G	0.054	0	127/20/1	55/0/0	1.01E-03	18.11
rs78120781	*IFLTD1*	chr12:25656706	T>C	0.058	0	127/17/3	53/0/0	1.03E-03	18.44
rs41269662	*FAM78B*	chr1:166039898	C>G	0.111	22	126/21/1	34/20/1	1.09E-03	0.34
rs3764645	*ABCA7*	chr19:1042809	A>G	0.466	44	49/75/24	9/26/20	1.14E-03	0.47
rs7830986	*LACTB2*	chr8:71550832	T>C	0.089	2	116/30/2	53/2/0	1.32E-03	7.01
rs4789623	*SDK2*	chr17:71443848	C>T	0.012	5	148/0/0	50/5/0	1.36E-03	0.03
rs35976576	*ABTB1*	chr3:127396573	C>T	0.012	5	148/0/0	50/5/0	1.36E-03	0.03
rs45488095	*CEP128*	chr14:81259336	T>C	0.012	5	148/0/0	50/5/0	1.36E-03	0.03
rs659040	*CWF19L2*	chr11:107299631	G>A	0.207	35	102/43/3	22/31/2	1.38E-03	0.43
rs34197769	*PKD1*	chr16:2144176	G>A	0.101	3	115/28/5	52/3/0	1.48E-03	5.25
rs6891545	*ANKRD32*	chr5:93987532	C>A	0.213	35	102/41/5	25/23/6	1.50E-03	0.43
rs115489792	*DZIP1L*	chr3:137787041	C>T	0.049	12	141/6/1	43/12/0	1.54E-03	0.23
rs8104361	*CYP4F11*	chr19:16034714	A>G	0.783	36	5/42/101	6/24/25	1.68E-03	0.44
rs3736038	*LRRCC1*	chr8:86027418	A>G	0.052	0	128/19/1	55/0/0	1.70E-03	17.25

Variant	Gene	Position	Allele	MAF	MINU	Affected Counts (WT/Het/Mut)	Unaffected Counts (WT/Het/Mut)	P-value	Odds Ratio
rs1193851	*PCNXL3*	chr11:65386206	C>G	0.372	55	67/66/15	14/27/14	1.73E-03	0.48
rs12237048	*KCNV2*	chr9:2718534	C>G	0.473	38	36/70/42	25/22/8	1.76E-03	2.05
rs3828267	*ASNSD1*	chr2:190531615	T>C	0.062	14	138/9/1	41/14/0	1.82E-03	0.26
rs79821910	*LGR6*	chr1:202183327	C>T	0.027	8	145/3/0	48/6/1	1.82E-03	0.13
rs113459553	*NKX2-3*	chr10:101294936	C>T	0.027	8	145/3/0	47/8/0	1.82E-03	0.13
rs12615819	*TPO*	chr2:1544617	G>A	0.389	29	42/83/23	29/23/3	1.91E-03	2.16
Novel	*ARSE*	chrX:2867704	A>G	0.017	6	147/1/0	50/4/1	1.94E-03	0.06
rs12048517	*ADCY10*	chr1:167815030	A>G	0.017	6	147/1/0	49/6/0	1.94E-03	0.06
rs111276200	*DYNLT1*	chr6:159058804	G>A	0.017	6	147/1/0	49/6/0	1.94E-03	0.06
rs10471773	*CCDC125*	chr5:68616079	G>A	0.017	6	147/1/0	49/6/0	1.94E-03	0.06

with AD, although it did not pass Bonferroni correction for multiple testing ($P=6.3x10^{-6}$). This missense variant rs3764645 (p.E188G) has been previously investigated in AD and is associated with a protective effect, independent of the GWAS SNP rs4147929. A variant in the *NTN4* gene rs17851048 also exerted an effect in the protective direction ($P=1.6x10^{-4}$, OR=0.4). *NTN4* was shown to associate with a decrease in age of onset for family-based samples with AD in a previous study (Peng *et al.,* 2015). However, there are some variants present with high effect sizes due to no minor alleles being observed in control samples and these should be treated with caution until more data is available.

4.4.3. Logistic regression

Logistic regression analysis was performed on the same variant set, correcting for age, sex and *APOE* status with number of ε2 and ε4 alleles. Again, the small sample size hinders the detection of associations to genome-wide significance. However, tentative signals were observed near the suggestive significance threshold. Many of the variants identified through single variant association were still significant following adjustment for covariates. The *APOE* SNPs rs429358 and rs7412 were not observed, as expected since they were corrected for using *APOE* ε status. The signals in AD-related genes *ABCA7*, *PILRA* and *NTN4* were still present after correction although with less significance, which was expected given the sample size. A further two SNPs in *ABCA7* were also associated with AD, rs3752237 ($P=8.8x10^{-4}$, OR=0.41) and rs4147915 ($P=1.5x10^{-3}$, OR=0.31). Although rs3764645 and rs3752237 are not in LD ($r^2=0.55$), they exhibit the same effect size of 0.41 in the protective direction and standard error. Neither SNP is in LD with rs4147915, suggesting that these signals could be independently contributing to AD risk within the *ABCA7* gene. *PILRA* and *NTN4*, which have been linked to AD in previous studies, are again associated. Although they do not reach genome-wide significance, it is possible that a larger sample size could validate these findings.

The IGAP dataset comprises a meta-analysis of four major AD GWAS in a total of 74,046 samples of European ancestry. Due to the large sample size, it is possible that some of the effects observed here could reach significance in the IGAP dataset. *PILRA* SNP rs2405442 exerted a protective effect in the meta-analysis and showed a similar level of significance despite not reaching the corrected thresholds ($P=4.2x10^{-5}$, $\beta=-0.07$). *ABCA7* variant rs3764645 surpasses the suggestive threshold for association with AD ($P=2.9x10^{-6}$, $\beta=-0.1$), the SNP rs3752237 was not significantly associated ($P>0.05$) and rs4147915 was not genotyped. *NTN4* rs17851048 was also not significant ($P>0.05$).

Gene ontological term annotation identified 7 genes that were all associated with vesicular trafficking activity. *AHSA1*, *AGKH*, *LRRC8E*, *KIF13B*, *PMF1*, *SEC31A* and *TRAPPC9* are involved in intracellular transport of vesicles from the endoplasmic reticulum (ER) to the Golgi. The endocytic pathway is enriched in this gene set, with *KIF13B* also regulating axon formation and *FCHO2* functioning in clathrin-mediated endocytosis. *SCARB1* is a receptor that mediates cholesterol uptake, which links it to the function of *APOE*, also involved in cholesterol transport. Some genes contain multiple variants exerting a tentative association with AD. There are three variants within the *LRRC8E* gene associated with a decreased risk of AD; of these, rs2042919 and rs3745382, both have the same effect size and standard error. They are in almost perfect LD with $r^2=0.99$ in European populations, indicative of potential haplotypes. The other SNP rs3745383 is in weak LD with both SNPs ($r^2=0.48$) and is therefore an independent association within the *LRRC8E* gene. Other SNPs with strong LD ($r^2>0.9$) are seen in the genes *PCNXL3*, *TMEM171*, *TRAPPC9* and *SBNO1*.

Table 4.2. Logistic regression results were adjusted for the covariates sex, age and number of APOE ε2 and ε4 alleles. Variants that were associated above P<0.002 are shown here. No variants achieved genome-wide significance or reached the suggestive threshold due to the sample size. However, tentative associations could be observed. Multiple variants in known AD risk gene *ABCA7* were present, along with a variant in *PILRA*, which has been indirectly implicated in AD. The functional consequences of the coding variants were also determined using VEP.

Variant	Gene	Position	Allele	Effect Size	P-value	Functional Prediction
rs2042919	*LRRC8E*	chr19:7963949	A>G	0.23	2.20E-05	Missense
rs3745382	*LRRC8E*	chr19:7964727	G>A	0.23	2.20E-05	Synonymous
rs1193851	*PCNXL3*	chr11:65386206	C>G	0.31	6.33E-05	Missense
rs2405442	*PILRA*	chr7:99971313	T>C	3.16	1.33E-04	Synonymous
rs1205259	*PCNXL3*	chr11:65392695	C>T	0.35	2.82E-04	Synonymous
rs5761268	*MYO18B*	chr22:26239850	C>A	0.35	3.34E-04	Missense
rs2234582	*WT1*	chr11:32456694	C>A	0.34	3.36E-04	Synonymous
rs3745383	*LRRC8E*	chr19:7965339	C>T	0.25	3.46E-04	Stop gained
rs190533743	*SLC22A18AS*	chr11:2909513	C>G	0.22	4.24E-04	Missense
rs636926	*TMEM171*	chr5:72419617	C>A	0.36	4.66E-04	Missense
rs637450	*TMEM171*	chr5:72419456	C>G	0.36	4.66E-04	Missense
rs1806191	*GRIN2B*	chr12:13716638	G>A	0.36	4.80E-04	Missense
rs1051431	*MPHOSPH9*	chr12:123645803	G>A	3.28	4.95E-04	Synonymous
rs13986	*TRAPPC12*	chr2:3483181	C>T	0.23	5.42E-04	Synonymous
rs198972	*KLK2*	chr19:51379893	C>T	0.37	5.52E-04	Synonymous
rs659040	*CWF19L2*	chr11:107299631	G>A	0.31	6.14E-04	Missense
rs41269662	*FAM78B*	chr1:166039898	C>G	0.27	6.46E-04	Synonymous
rs1997040	*DGKB*	chr7:14758273	C>T	0.4	6.54E-04	Synonymous

Variant	Gene	Position	Allele	Effect Size	P-value	Functional Prediction
rs819085	*COL6A5*	chr3:130174334	G>A	5.9	7.93E-04	Missense
rs3752237	*ABCA7*	chr19:1047161	A>G	0.41	8.78E-04	Synonymous
rs2293440	*SCARB1*	chr12:125270773	A>G	0.07	9.26E-04	Missense
rs3764645	*ABCA7*	chr19:1042809	A>G	0.41	9.34E-04	Missense
rs17804499	*RASSF6*	chr4:74442349	G>C	0.18	9.61E-04	Missense
rs7250	*AHSA1*	chr14:77935520	G>A	0.39	9.86E-04	Synonymous
rs9624909	*MYO18B*	chr22:26222454	C>T	0.34	1.05E-03	Missense
rs17559	*IGFALS*	chr16:1841033	G>A	0.29	1.12E-03	Synonymous
rs34457678	*TRAPPC9*	chr8:141461116	C>T	0.21	1.12E-03	Synonymous
rs57593250	*TRAPPC9*	chr8:141460918	C>T	0.21	1.12E-03	Synonymous
rs61739560	*TRAPPC9*	chr8:141461074	A>G	0.21	1.12E-03	Synonymous
rs8110644	*ZFP14*	chr19:36831372	A>G	0.22	1.13E-03	Synonymous
rs1056854	*HNRNPUL1*	chr19:41809995	A>G	0.32	1.20E-03	Synonymous
rs2070746	*FPR1*	chr19:52249702	G>T	2.88	1.31E-03	Synonymous
rs17851048	*NTN4*	chr12:96077312	C>T	0.38	1.37E-03	Synonymous
rs1060105	*SBNO1*	chr12:123806219	C>T	0.34	1.42E-03	Missense
rs12322888	*SBNO1*	chr12:123825559	C>T	0.34	1.42E-03	Synonymous
rs117202308	*KIF13B*	chr8:28929739	G>A	0.07	1.43E-03	Missense
rs4147915	*ABCA7*	chr19:1049305	C>A	0.31	1.46E-03	Synonymous
rs79867908	*COL6A5*	chr3:130116684	G>A	0.17	1.48E-03	Missense
rs1864312	*CRISP3*	chr6:49701439	C>A	2.54	1.51E-03	Missense
rs111475461	*SPG7*	chr16:89613073	G>A	0.06	1.61E-03	Missense

Variant	Gene	Position	Allele	Effect Size	P-value	Functional Prediction
rs10747057	*PWWP2B*	chr10:134218296	C>G	0.43	1.72E-03	Missense
rs10160116	*ZNF438*	chr10:31138817	G>A	0.42	1.74E-03	Missense
rs185435	*FCHO2*	chr5:72348272	A>G	0.35	1.77E-03	Missense
rs2227935	*BLM*	chr15:91326099	C>T	0.19	1.79E-03	Missense
rs1052067	*PMF1*	chr1:156206121	G>A	2.84	1.81E-03	Missense
rs2255336	*KLRK1*	chr12:10532326	T>C	2.85	1.81E-03	Missense
rs6488868	*SBNO1*	chr12:123799974	A>G	2.65	1.82E-03	Synonymous
rs10025654	*SEC31A*	chr4:83795806	T>C	2.39	1.84E-03	Synonymous
rs4329520	*KPRP*	chr1:152733301	T>A	0.44	1.91E-03	Missense

4.4.4. Burden analysis

Exome-wide burden was calculated for aggregated variants in 14,971 genes using a SKAT-O test. The Bonferroni-corrected threshold was $P=3.3x10^{-6}$, which was not achieved by any of the genes. This test has 2-sided directionality and therefore it can test for the presence of mixed effects of variants within the gene. Table 4.3 contains the results for genes with combined rare and common variant burden significant to $P<0.001$ and the separate results included. Several genes containing variants identified as significant by association analysis were observed to exhibit a high burden on AD risk, including *GRIN2B*, *LRRC8E*, *PILRA*, *MYO18B*, *DKGB* and *SBNO1*, amongst others. Closer observation of single variants in the logistic regression data indicates that the highly associated SNPs are likely to be driving the signals in burden analysis. Some of these contain multiple variants that were associated with AD in single variant analysis. Several variants in the genes *LRRC8E*, *MYO18B* and *SBNO1* appear to display burden on AD risk. *LRRC8E* contains 9 variants, with the common SNPs being more associated with AD, however combining all variants increased significance further. *MYO18B* contains a large number of rare and common variants that showed nominal association when tested separately, although this increased during joint testing. The burden of *SBNO1* on AD is driven solely by the common variants, as P=0.35 for rare variants tested together.

PILRA exerted the most burden on AD in joint testing ($P=3.68x10^{-5}$), however it appears that only one of the three SNPs is driving this signal. Synonymous SNP rs2405442 was highly associated with AD in single variant association, with correction for covariates. The other SNPs are both rare, missense variant rs138933650 and synonymous variant rs34266222, which were not significant individually and only showed tentative burden on AD, P=0.067. There were 7 AD cases heterozygous for rs34266222 and the minor allele was not observed in any controls; however, this was not found to be significant.

Table 4.3. Exome-wide burden analysis using SKAT-O test for nominally associated genes (P<0.001). Burden for genes was tested using rare and common variants together and individually to determine whether either group of variants exerted stronger effects on association as a collective. The number of variants tested within each gene is listed for joint, rare and common burden testing, alongside the P-value.

Gene	Joint Testing		Rare Only		Common Only	
	Tested Markers (Total)	P-value	Rare variants	P-value	Common variants	P-value
LRRC8E	9 (9)	1.42E-05	4	3.41E-02	5	5.44E-05
PILRA	3 (3)	3.68E-05	2	6.70E-02	1	6.44E-05
VWC2	5 (5)	4.61E-05	3	3.30E-05	2	9.54E-02
MYO18B	34 (34)	1.60E-04	22	4.01E-02	12	2.67E-04
GRIN2B	12 (12)	1.73E-04	5	2.26E-02	7	6.49E-04
ABTB1	2 (2)	2.97E-04	2	2.97E-04	0	NA
DHRS13	2 (2)	4.24E-04	0	NA	2	4.24E-04
ALDH9A1	3 (3)	4.97E-04	1	1.36E-03	2	2.95E-02
KRT26	4 (4)	5.60E-04	3	2.95E-03	1	1.13E-02
HNRNPUL1	3 (3)	5.62E-04	2	1.74E-01	1	6.08E-04
NTNG2	7 (8)	5.82E-04	3	3.69E-04	4	4.49E-01
MOGAT3	6 (6)	5.90E-04	5	1.56E-02	1	2.58E-03
PCNXL3	13 (14)	6.05E-04	8	7.20E-01	5	1.24E-04
DGKB	3 (3)	7.50E-04	2	5.33E-01	1	3.06E-04
SBNO1	8 (8)	9.22E-04	4	3.47E-01	4	5.51E-04
STOX2	7 (7)	9.40E-04	6	6.27E-03	1	1.05E-02
RIPK2	2 (2)	9.72E-04	0	NA	2	9.72E-04

4.4.5. Polygenic risk scores

A 500kb region around the *APOE* gene locus containing 233 variants was excluded from the analysis to identify effects independent of *APOE*. There were 49,708 variants in common between the IGAP and WES dataset. PRSice calculated scores using the base phenotype for individuals in the target dataset at multiple significance thresholds. Figure 4.3 (a) illustrates the PRS model fit across various thresholds, P_T 0.01 to 0.5 are shown in increments. Nagelkerke's pseudo R^2 indicates how well the scores fit the prediction model, with an R^2 closer to 1 representing a better fit. The table in Figure 4.3 (b) provides the number of variants that contribute to the risk scores at each threshold, with the best fit shown in red. PRSice identified the best AD risk prediction model to include variants associated where $P_T<0.05$.

As the GWAS susceptibility loci were located in intronic or intergenic regions of the genome, they were not covered in the exome sequencing. Therefore, they were genotyped manually using KASP assays or obtained from Neurochip array data incorporated as covariates. Figure 4.4 (a) displays a ROC curve which allows a comparison of the sensitivity and specificity of each risk prediction model when discriminating between cases and controls based on PRS. All models are above the reference line at 50% where sensitivity equals specificity, meaning no discrimination is possible between the groups.

The table in Figure 4.4 (b) indicates the accuracy of each prediction model, with the *APOE* ε4 alone able to distinguish cases from controls 73.1% of the time, however this is not a good fit model, as the Hosmer-Lemeshow test is significant. The protective ε2 allele increased the ability to correctly identify controls by 4%. Age and sex add little information to the model, which was expected since both case and control groups consist of elderly individuals of a similar age range and sex could be indiscriminate due to sample size. However, addition of the 19 GWAS signals increases predictive accuracy by 6%.

(a)

Threshold (P_T)	P-value	Nagelkerke's R^2	Number of SNPs
0.01	0.246	0.010	881
0.05	**0.083**	**0.022**	**3326**
0.1	0.318	0.007	6069
0.2	0.222	0.011	11257
0.3	0.217	0.011	16155
0.4	0.285	0.008	20924
0.5	0.216	0.011	25872
1.0	0.164	0.014	49708

(b)

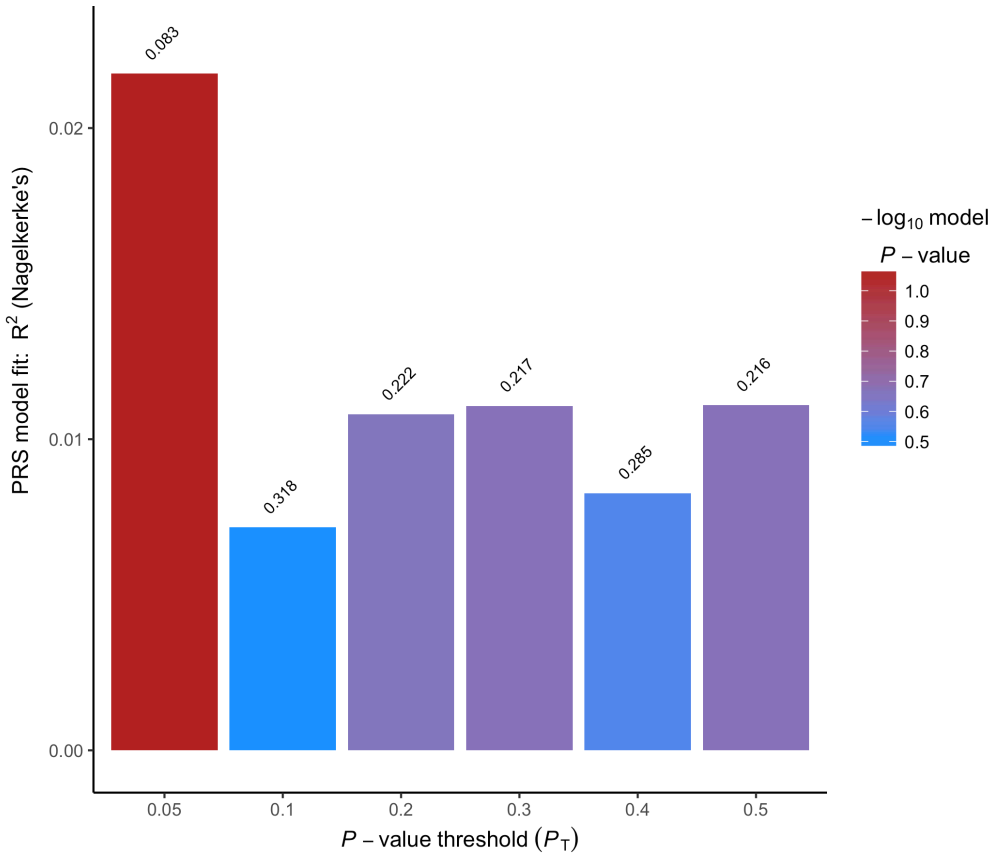


Figure 4.3. P-value thresholds tested by PRSice to determine best prediction model. (a) Raw results for various P-value thresholds (P_T) and the number of variants that contribute to scores at this level. The best threshold is shown in bold red. (b) Bar plot output by PRSice demonstrating PRS model fit across several tested thresholds, as indicated by Nagelkerke's R^2. The best fit threshold for prediction was P_T<0.05 which had a pseudo R^2 of 0.022.

Incorporating the PRS only improved the model by 1%, which is similar to what other studies have found. In total, 3,326 variants were utilised to score risk of developing AD using the best fit model, which had a predictive accuracy of 85.3% when combined with age, sex, *APOE* ε2 and ε4 status and genotypes for the 19 GWAS variants, as predictors.

There was a considerable overlap between scores for both groups as shown in Figure 4.5; however, normalised scores were higher for cases, with a mean of 0.07 compared to -0.20 for control samples. A higher score is indicative of an increased risk, whereas a lower score suggests the opposite. The median, upper and lower quartiles were overlaid on the scores for each group, along with highest and lowest scores. Although the dispersion of scores was greater for cases, on average they were higher than controls. Some outliers were present in both groups with control samples BB012, BK078 and BK105 displaying high PRS and AD case B0062 having the lowest risk score overall.

These outliers were explored further to determine whether these scores correlated with any of the clinical features and neuropathology data available. Table 4.4 contains information for samples with the lowest (PRS_{norm} < -1.50) and highest scores (PRS_{norm} > 1.50) from both groups. Control samples with high risk scores are highlighted in blue, with BB012 having the highest score overall of 2.79. Two of these patients, aged 68 and 72 at the time of death, presented with Braak tangle staging 1-3, which is consistent with normal ageing. The other individual died at a much older age of 96 and no major disease pathology aside from mild non-amyloid SVD and ageing changes; this could be a result of unknown protective variants. CERAD scores were either 0 or not available for all controls. The *APOE* ε status for all control samples, except BB012, consisted of the neutral ε3 or ε2 protective allele. BB012 was a control with a high PRS and *APOE* status of ε2ε4, with the ε2 allele likely to lessen the risk conferred by the ε4 allele. Controls with low PRS did not have AD pathology but sample BN018 did exhibit symptoms of mixed dementia, with symptoms of

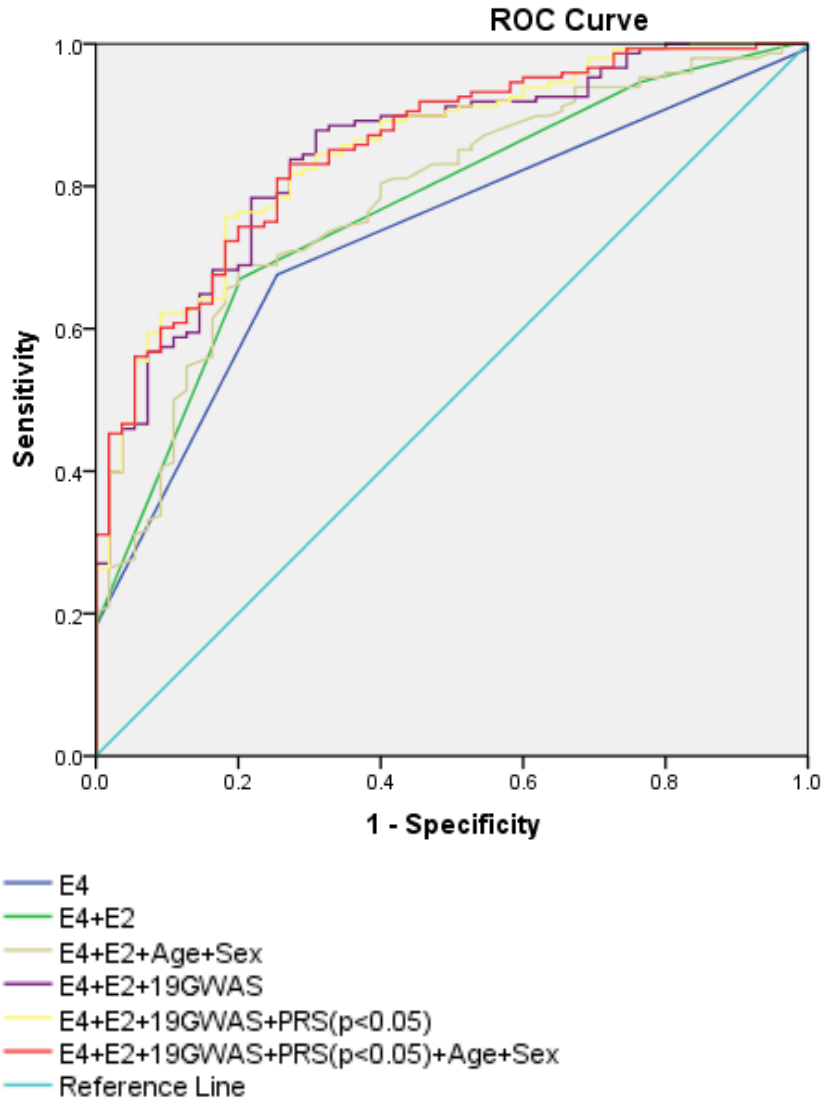


Figure 4.4. (a) ROC curve depicting the different risk prediction models for AD. Multiple prediction models were generated to determine the best combination of predictors to classify AD cases and controls with the greatest accuracy possibly. The model comprising all variables, age at death, sex, number of *APOE* ε2 and ε4 alleles and scores achieved the greatest accuracy.

(b) Predictive accuracy model for 148 AD cases versus 55 controls. Different predictors were compared for their accuracy in predicting risk of develop AD. The variables *APOE* ε4, ε2, age, sex and genotypes at the 19 GWAS risk loci were added as predictors to increase the predictive ability for the PRS model at P<0.05 threshold. Nagelkerke's R^2 expresses the proportion of variance explained by the model, with the largest R^2 value indicating the best model. Area under the receiver operating characteristic curve (AUC) is a measure of predictive accuracy, which quantifies the overall ability to discriminate between case and control individuals. The Hosmer-Lemeshow statistic is a goodness-of-fit test for risk prediction models, with a significant result indicating that the data are a poor fit to the model.

Predictive Model	Nagelkerke's R^2	Area under ROC curve (AUC)	AUC 95% CI	Hosmer-Lemeshow P-value
ε4	0.102	73.1%	0.642-0.794	0.001
ε4 + ε2	0.271	77.0%	0.660-0.809	0.347
ε4 + ε2 + Age + Sex	0.273	78.0%	0.668-0.816	0.891
ε4 + ε2 + GWAS	0.412	84.7%	0.770-0.891	0.816
ε4 + ε2 + GWAS + PRS(P<0.05)	0.424	85.1%	0.777-0.896	0.814
ε4 + ε2 + GWAS + PRS(P<0.05) + Age + Sex	0.429	85.3%	0.779-0.898	0.536

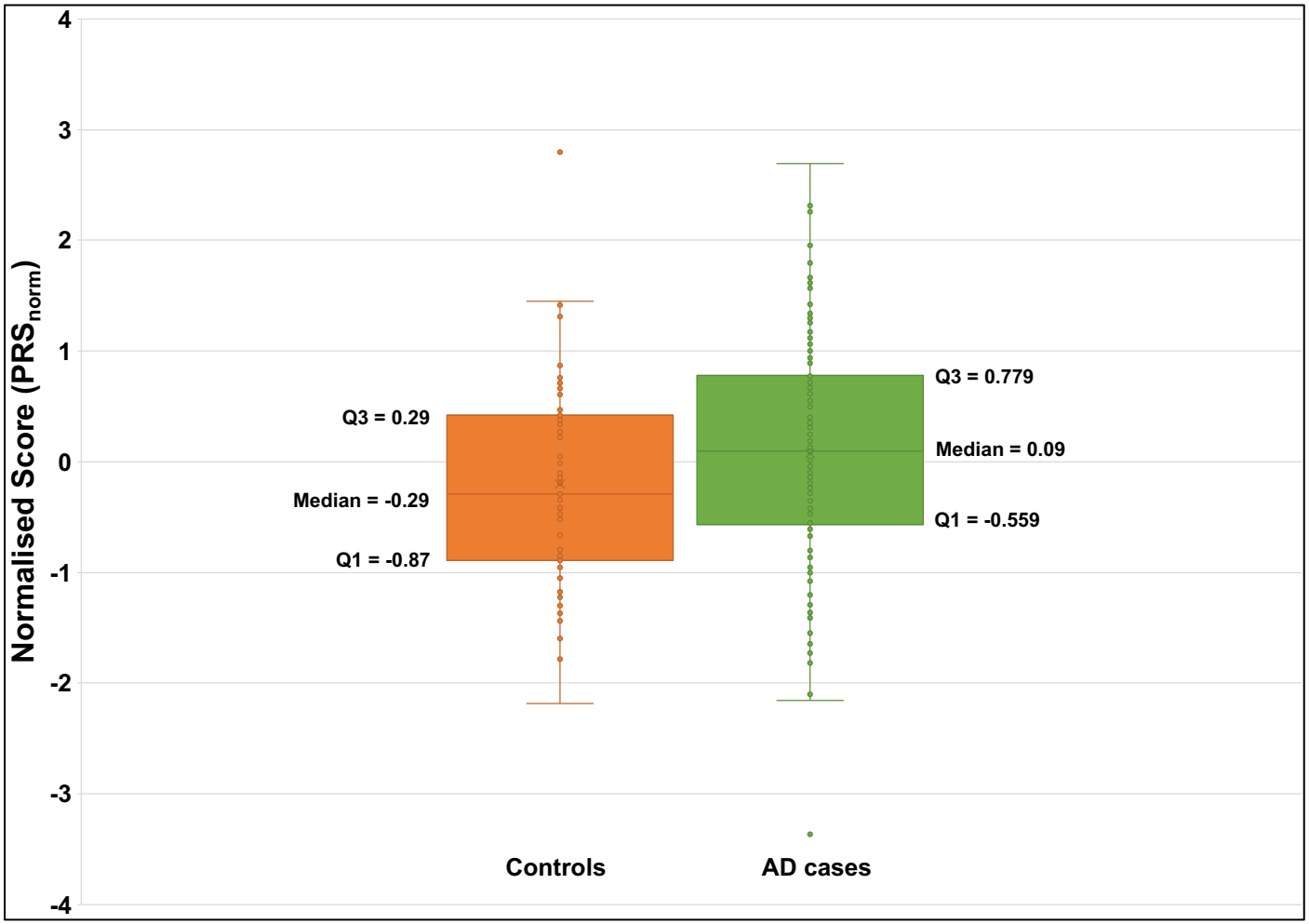


Figure 4.5. Dispersion of normalised PRS for AD cases and controls at P<0.05. The best prediction threshold was determined to be P<0.05 for controls (orange) and AD cases (green). Outliers were present in both groups, as there are controls with high PRS and cases with very low PRS. There is a large overlap, however on average the scores for cases were higher.

FTLD, CVD or PSP. This individual also possessed an ε2 allele which could have contributed to the protective effect that prevented them from developing AD. The other two samples have the more common ε3ε3 genotype.

AD cases with low PRS appeared to present with various other neuropathologies, with CVD, CAA, DLB and SVD seen among the comorbidities, highlighted in orange. Most individuals had a Braak tangle stage above 3, which is more suggestive of AD than normal ageing. A Braak stage of 5 was most common, which indicates extensive neurofibrillary tangles in neocortical regions affecting cognition. The available CERAD scores also indicate moderate to severe AD, with scores of 2 or 3. There are a mixture of *APOE* genotypes among the low risk AD group, the majority of which are neutral or contain an ε4 allele. BO062 had an *APOE* genotype of ε3ε3, high NFT involvement and definite AD as defined by CERAD score; however, they had the lowest risk score overall. Other pathologies included CVD, severe CAA, microinfarcts, atherosclerosis and arteriolarsclerosis. BB023 had a protective ε2 allele and also showed signs of mixed dementia, with signs of DLB, severe CAA, vascular pathology, infarcts and white matter disease. These other pathologies could explain the low PRS, as the genetic component of AD may be lesser due to comorbidity.

The majority of cases with high PRS possessed an ε4 allele, except BK094 which had the highest score for AD samples yet harboured an ε2 variant; although this does not guarantee protection from developing AD. Interestingly, there were no ε4ε4 homozygotes within this subset. Braak and CERAD stages indicated some signs of AD for all individuals in the high-risk AD group. All individuals were over the age of 80, with probable or definite AD as defined by CERAD if this information was available. Again, there were some comorbid phenotypes, with vascular dementia and SVD prominent and TDP-43 pathology observed in 3 of the AD cases.

Table 4.4. Table of outliers in AD case and control groups with their clinical features. AD cases with the lowest PRS (PRS < -1.5) and controls with higher PRS (PRS > 1.5) are shown here with the associated neuropathological information. There is little neuropathology data on the control individuals to indicate misdiagnosis or risk of developing AD. This highlights the likely involvement of potential protective variants that can reduce risk of AD in controls. Contrary to this, the majority of the low risk AD patients harboured a comorbidity, either in the form of another dementia or pathological features relating to another dementia. This could attribute to the lower risk of AD in these individuals as the AD genetic component may be less prominent due to the comorbid phenotype.

Group	Sample	Gender	*APOE* status	AAD	Braak Stage	CERAD	PRS	Neuropathology and Diagnosis Information
Low risk control	BN018	M	23	74	0	0	-2.19	Mixed dementia; PSP; CBD; FTLD
	BK002	M	33	86	3	NA	-1.79	Ageing changes
	BK049	F	33	87	2	NA	-1.60	Ageing changes; mild SVD, mild CAA
	BN029	M	23	88	2	0	-1.44	Ageing changes; neuritic pathology in the entorhinal stage
	BM043	F	33	95	5	NA	-1.37	Probable AD; moderate SVD; rare Lewy-like inclusions in the medullary nuclei and amygdala
	BM058	M	33	85	1	NA	-1.30	Ageing changes
	BK106	F	33	97	2	NA	-1.22	Mild AD type changes; argyrophilic grains in the limbic system
	BO011	M	33	84	-9	NA	-1.20	Ageing changes; epilepsy; bipolar disorder
High risk control	BK105	F	33	68	3	NA	1.31	Ageing changes
	BO058	M	24	85	2	2	1.42	Abnormal deposition of Aβ protein in diffuse plaques and blood vessels
	BK078	F	33	72	1/2	NA	1.43	NA
	BB010	M	34	86	2	0	1.45	No significant abnormalities
	BB012	M	24	96	2	0	2.79	Ageing changes, mild non-amyloid SVD

Group	Sample	Gender	*APOE* status	AAD	Braak Stage	CERAD	PRS	Neuropathology and Diagnosis Information
Low risk AD	BO062	M	33	73	5	3	-3.37	CVD with acute massive intracerebral haemorrhage; severe CAA; microinfarcts; atherosclerosis; arteriolarsclerosis
	BB023	M	23	71	5	3	-2.16	Mixed dementia (AD/DLB); severe CAA complicated by Aβ related angiitis; vasculopathy; thrombosis; multiple infarcts; ischaemic white matter disease
	BK088	M	33	82	5	NA	-2.10	Mild CAA; mild-moderate white matter CVD changes
	BO084	F	34	80	5	3	-1.82	DLB
	BO054	F	33	90	5	3	-1.81	Extensive subarachnoid haemorrhage
	BM011	F	34	85	3	2	-1.73	Mild AD
	BN022	M	34	91	6	3	-1.71	Moderate CAA
High risk AD	BM082	F	34	84	3	2	1.80	Moderate SVD; mild TDP-43 pathology
	BN011	M	34	87	6	3	1.82	Severe CAA; TDP-43 pathology; CVD pathology; SVD
	BB007	M	34	94	5	2	1.95	Mixed dementia; mild CAA; TDP-43 inclusions
	BM016	F	34	81	4	3	2.26	Aβ-CAA
	BB018	M	34	92	6	3	2.31	Vascular dementia; moderate-severe arteriosclerotic SVD
	BK094	F	23	85	5	NA	2.69	Vascular dementia; moderate CAA; moderate-severe white matter CVD

Although some notable differences were observed in cases and controls, there was still a large overlap between these groups, providing little evidence to distinguish low and high-risk individuals using neuropathology alone. The prediction models demonstrate that a multitude of factors are required to classify AD from controls with high accuracy.

Further investigation into the 3,326 variants that contributed to the PRS involved mapping to genes. This resulted in 2,562 genes with multiple variants being observed in 551 of these genes. Among these, there were many AD-related risk genes previously highlighted in the literature, including 10 AD susceptibility genes previously identified by GWAS. Table 4.5 lists the variants found in some of these known genes and their details. Several genes newly associated with AD were also observed, such as *TREML2* and *PLCG2*, further supporting their role in diseases. The genes neighbouring *ABCA7* also contained variants contributing to the risk scores, *CNN2* and *HMHA1*; this implicates the whole *ABCA7* region. There were 2 variants in Parkinson's risk gene *LRRK2* that made up the PRS, highlighting potential crossover. Interestingly, we also observed lesser known genes previously lacking sufficient evidence to link them to AD; *ACE*, *SPARCL1* and *VAC14*.

4.5. Discussion

GWAS have identified a number of new susceptibility loci associated with AD, addressing common variation. However, this failed to capture most of the missing heritability observed in AD. Increasing evidence suggests that low frequency variants (MAF < 5%) with greater effect sizes could contribute more to the underlying genetic architecture of complex diseases (Rivas *et al.,* 2011). The advent of high-throughput sequencing technology has allowed the discovery of multiple rare variants associated with AD, which go undetected by GWAS. WES studies fundamentally differ from GWAS as they enable unbiased v

Table 4.5. AD-associated risk genes containing risk variants that contribute to PRS. Risk variants were annotated to genes which comprised of multiple AD-related genes, including a number of lesser known genes that were previously linked to AD but with insufficient evidence. The link for each gene to AD is linked, along with the risk score attributed to the gene. Numerous GWAS genes were identified, with neighbouring genes to *ABCA7* and *ZCWPW1*. *ACE*, *VAC14* and *SPARCL1* were investigated in house. In addition, variants in *TREM1*, *TREML2* and newly identified susceptibility gene *PLCG2* were identified, as well as variants in genes related to other neurodegenerative diseases.

Gene	Number of Variants	Variant(s)	Score	Link to AD
ABCA7	8	rs4147935, rs4147934, rs3752240, rs3752229, rs4147914, rs3764645, rs12151021, rs4147910	1.47E-03	Lambert *et al*. (2013)
BIN1	5	rs76414720, rs2276575, rs1060743, rs11554586, rs11682128	1.47E-02	
CD2AP	3	rs1014001, rs9473132, rs9395287	4.71E-02	
CD33	4	rs61736475, rs35112940, rs2455069, rs12459419	3.90E-02	
CLU	1	rs7982	2.48E-17	
CR1	3	rs1752688, rs41274768, rs2296160	2.06E-02	
EPHA1	2	rs1131883, rs1131885, rs117651735	3.49E-02	
FERMT2	1	rs2357947	3.64E-04	
INPP5D	3	rs144181925, rs7346, rs1135173	2.14E-02	
MS4A locus	7	rs117801657, rs528823, rs10750931, rs6591561, rs583791, rs2289613, rs2304935	9.33E-02	
NME8	3	rs2722372, rs3213976, rs3816418	1.04E-03	
PICALM	3	rs10792821, rs592297, rs618679	8.32E-03	
PTK2B	6	rs11774417, rs1030526, rs2241652, rs1879184, rs7007145, rs4534095	4.78E-02	
SLC24A4	3	rs11623883, rs6575255, rs4900122	8.12E-02	
SORL1	6	rs1010159, rs2070045, rs146197030, rs1784933, rs17125548, rs2276412	5.56E-02	
ABCA locus	4	rs2066716, rs2066714, rs1550828, rs6501839	8.85E-02	*ABCA* gene family

Gene	Number of Variants	Variant(s)	Score	Link to AD
CNN2	3	rs2304260, rs930232, rs930231	1.01E-02	*ABCA7* flanking gene
HMHA1	3	rs3761028, rs4807506, rs10404947	2.43E-03	
ACE	5	rs4365, rs4362, rs4311, rs4309, rs3730025	8.65E-02	Braae *et al*. (2014)
C9orf72	1	rs62538126	4.26E-02	FTD risk gene
GRN	2	rs850713, rs5848	3.35E-02	Parkinson's disease risk gene
IDE	1	rs1832195	2.91E-02	Epistasis Project; Reitz *et al*. (2012)
KIF11	1	rs12256435	2.78E-02	
LRRK2	2	rs10878307, rs33962975	6.60E-02	Parkinson's disease risk gene
PILRA	1	rs2405442	4.23E-05	ZCWPW1 flanking gene; Allen *et al*. (2015); Karch *et al*. (2016)
PLCG2	5	rs11548656, rs8045964, rs4366702, rs1071644, rs117991680	6.88E-02	Sims *et al*. (2017)
PSEN1	1	rs165932	2.67E-02	FAD risk gene
SPARCL1	2	rs8342, rs72654103	6.03E-02	Vafadar-Isfahani *et al*. (2012); Richens *et al*. (2014)
TREM1	1	rs2234246	4.04E-02	*TREM* locus; Replogle *et al*. (2017)
TREML2	1	rs3747742	8.56E-03	*TREM* locus; Benitez *et al*. (2014)
VAC14	1	rs2278983	4.90E-04	C. Medway project

variant discovery and direct association between a phenotype and causal variant (Stitziel *et al.,* 2011). In addition, focusing on coding regions can help prioritise variants that are more likely to be functional or deleterious. The architecture of intronic regions is extremely complex, and while *cis*-regulatory elements likely play an important role in AD pathogenesis, the functional significance of individual non-coding mutations cannot be easily predicted and may not necessarily affect neighbouring genes. Furthermore, some regulatory elements interact with multiple genes over extremely large distances thus making a determination of the impact of a single mutation difficult (Fraser *et al.,* 2015). The implication of this is that identifying non-coding mutations may be of little use for identifying potential targets for therapeutic interventions.

4.5.1. Quality control

The final cohort consisted of data from 288 individuals; 148 AD and 55 control samples with 85 other phenotypes. Sample BO017 was removed due to a genotyping rate of 92%, which was below the threshold of 97%. Although the missingness rate was not extremely high, this cut-off was used to ensure high quality information was obtained. Filtering was performed to retain variants with a high call rate that also passed the Hardy-Weinberg test. This test excluded any markers with observed genotype frequencies that significantly deviated from the expected frequency in controls, using a cut-off of $P<0.001$. The population is expected to be in HWE, as the distribution of alleles should remain relatively steady (Turner *et al.,* 2011). Only a small proportion of markers were removed for failing to reach the thresholds, resulting in 310,170 variants within the Agilent SureSelect capture regions. Small scale sequencing projects can identify large numbers of variants; it is necessary to prioritise variants that are likely to be affecting directly impact on the phenotype. An estimated 85% of causal variants are located in exonic regions, despite only comprising 1% of the entire genome (Choi *et al.,* 2009). Focusing on coding variants

reduced the total number of variants for investigation, whilst maintaining those more likely to be causal. Singletons were also removed to further refine the callset, as they could confound the data due to the difficulties in discriminating between true variants and random sequencing errors (Carson *et al.,* 2014). Error rates for NGS are often higher compared to Sanger sequencing, at 0.1-1% for the Illumina HiSeq platform (Fox *et al.,* 2014). Therefore, an erroneous sequencing error is more likely to occur than a novel, single base substitution observed in only one individual.

The on-target Ti/Tv ratio of 3.2 was more reflective of high quality WES data, which is generally between 2.8-3 (Guo *et al.,* 2014). A lower ratio represents a higher number of transversions; however, the exome is under greater selective pressure against this type of variation. Therefore Ti/Tv ratios deviating significantly from the expected value indicate a high rate of false positives, reducing the reliability of the variant callset (DePristo *et al.,* 2011). This also implies that many of the singleton variants are likely to be false positives as the ratio increased when they were removed.

The sex of each individual was determined from their genotype data to ensure consistency in the phenotyping data. Heterozygosity rates on chromosome X were used to determine the sex with an arbitrary threshold determined by the fit of the data. Males are expected to have a heterozygosity rate close to 0 as they only possess one X chromosome, whereas this is much higher for females (Anderson *et al.,* 2011). Closer inspection of the rates resolved the majority of conflicting results; however, three samples were excluded due to large disparity in their calculated and ascertained sex and also other genotype mismatches. It is important that samples with discordant sex information were investigated to ensure that the wrong phenotype data was not connected to the genotypes (Anderson *et al.,* 2011). This can also highlight plating errors, contamination or incorrect clinical data. By comparing this, we can be more confident that the phenotypic

information is correct. Unless the sample can be identified using existing genotype data or it can be confirmed that sex was recorded incorrectly, individuals with discordant sex information should be removed from further analysis.

4.5.2. Whole-exome association

The remaining number of non-singleton variants in coding regions was 79,097 for 288 individuals, with a genotyping rate of 99.85%. Due to the sample size, this study is underpowered, so it is unlikely that variants will achieve full significance. A minimum of 80% power can only be achieved to detect rare variant associations of MAF = 5% with an effect size of 3; nevertheless, tentative associations can be made. The BDR resource is continually growing with ongoing collection of patient samples, therefore as the sample size increases it is possible that many of these tentative findings could reach significance.

Single SNP testing determines the association of individual variants with a given phenotype; it compares the frequency of the risk allele between cases and controls to identify significant differences. Single variant association was performed in PSEQ using a Fisher's exact test for 148 AD cases versus 55 controls. The Manhattan plot in Figure 4.2 shows the dispersion of variants associated with AD across the genome. The only variant to surpass genome-wide significance was *APOE* SNP rs429358 with an effect size of 7.4, which denotes the ε4 risk allele. This result was expected as the possession of an *APOE* ε4 allele remains the greatest genetic risk factor for developing AD to date (Farrer, 1997). The risk for ε4 carriers is 3-fold greater, with a 12-fold increased risk in individuals homozygous for the allele (Coon *et al.,* 2007). The other *APOE* SNP rs7412 was also observed to associate with AD, although below the suggestive significance threshold (Table 4.1). This SNP generally indicates the presence of an ε2 allele, which corresponds with the odds ratio of 0.22, supporting the protective effect of this allele against AD (Farrer, 1997).

Table 4.1 contains information for the nominally associated SNPs at the arbitrary cut-off value $P<2x10^{-3}$. The top two signals aside from *APOE* are variants rs1748 and rs2234582 in genes *TSC2* and *WT1*, respectively. Neither variant has been highlighted in AD, however rs2234582 in *WT1* has been linked to clear cell renal cell carcinoma (Li *et al.,* 2013). The *WT1* (Wilms tumour 1) synonymous variant appears more frequently in control samples, with an odds ratio below 1 indicative of a protective effect. *WT1* encodes a transcription factor involved in cellular development and survival. It was also associated with a protective effect against AD, in both single SNP and gene-based analyses, highlighting a role for neuronal apoptosis in AD pathogenesis (Chaudhry *et al.,* 2015). *TSC2* (Tuberous sclerosis complex 2) also known as tuberin, has previously been implicated in AD, with protein levels significantly downregulated in the brains of individuals with AD and Down's syndrome with Aß plaques and tangles (Ferrando-Miguel *et al.,* 2005). This missense variant was observed more frequently in cases compared to controls, with an effect size of 4.4 in the risk direction and predicted to be probably damaging by PolyPhen. There are no controls with the homozygous mutant genotype observed, which was only seen in cases.

Interestingly, *LRRCC1* variant rs3736038 (MAF = 5%) had an uncorrected effect size of 17-fold which is extremely high. The minor allele was not observed in any controls and was only present in AD cases, with one homozygote individual. It is involved in maintaining the structural integrity of centrosomes during mitosis and has not been previously linked to AD.

PILRA (Paired immunoglobulin-like type 1 receptor alpha) synonymous SNP rs2405442 appeared to exert a risk effect against AD, with an effect size of 3.2. This gene has also previously been implicated in AD through its interaction with GWAS risk gene *ZCWPW1* and paired receptor gene *PILRB*. PILRB associates with DAP12, an adaptor protein that forms an immunoreceptor signalling complex with *TREM2*

(Allen et al. 2015). The variant rs2405442 is in high LD with the *ZCWPW1* GWAS index SNP rs1476679 (r^2=0.85).

Known AD risk gene *ABCA7* missense variant rs3764645 (p.E188G) is independent of the GWAS index SNP (r^2=0.2). A study investigating the effects of several *ABCA7* variants on EOAD observed a similar protective effect for rs3764645, which was found to be in strong LD with rare missense variant rs72973581 (p.G215S; MAF=4.1%) (De Roeck *et al.,* 2017). These tentative associations have not been corrected for covariates such as age and sex, which can cause bias due to individual differences.

4.5.3. Logistic regression analyses

A logistic regression test accounts for extraneous variables, such as age, sex, cognitive abilities and clinical phenotypes. *APOE* ε status was also corrected for to identify signals independent of this known genetic risk factor. Compared to the Fisher's exact test used by PSEQ, which calculates the relationship between two variables, a regression allows the inclusion of covariates that may confound the observed effects. Without adjusting for strong risk factors such as age or major genetic risk alleles, spurious associations could occur as a result of the variance between individuals (Pirinen *et al.,* 2012). The logistic regression test adjusted for age, sex and number of *APOE* ε2 and ε4 alleles.

The *APOE* ε4 allele is significantly correlated with AD risk, which is a well-established association (Farrer, 1997). The number of ε2 and ε4 alleles was used as covariates to identify variants independent of the *APOE* ε4 effect. The SNPs rs429358 and rs7412 were not significant following adjustment due to interaction effects between the covariates and SNPs, as they are perfectly correlated (Pirinen *et al.,* 2012). Additionally, the *LRRCC1* variant rs3736038 was not observed following correction for covariates. Logistic regression computes odds ratios from allele frequencies in the dataset, however as there were no

controls with the minor allele, this could have affected the result.

Some genes were represented by multiple variants, such as *PCNXL3*, *MYO18B*, *TMEM171*, *ABCA7* and *SBNO1*, with either 2 or 3 SNPs achieving nominal significance in the logistic regression test. Most of these SNPs are in high LD with each other, which explains the presence of more than one SNP with high association.

The most associated variants were rs2042919 and rs3745382, in the gene *LRRC8E*; they are in perfect LD (r^2=1) and exert the same protective effect. Missense SNP rs2042919 is more likely to be the causative variant rather than synonymous SNP rs3745382. Another variant in this gene, rs3745383, was also nominally associated with AD; however, the protective effect was independent of the other SNPs (r^2=0.48). VEP predicted that this results in a high impact stop-gained mutation in exon 3. *LRRC8E* is a leucine-rich repeat-containing protein that encodes for a voltage-gated anion channel subunit.

Many of the variants identified prior to correction, in AD-related genes *ABCA7*, *PILRA* and *NTN4*, still emerged at P<0.002 after the addition of covariates, with lower significance. Inclusion of covariates for rare diseases increases the magnitude of the estimated genetic effects, however it can in turn reduce the power to detect associations (Pirinen *et al.,* 2012; Stringer *et al.,* 2011). *NTN4* is involved in axon migration, neurite growth and may play an important role in development (Cirulli and Yebra, 2007). It has been shown to be upregulated in response to Aß-induced injury of neurites (Shen *et al.,* 2012). A family-based study of LOAD also identified a significant association of a variant in *NTN4* with age at onset (AAO) of AD (Saad *et al.,* 2015). These studies provide some evidence to highlight a potential role for *NTN4* in AD pathogenesis.

Synonymous variants rs2405442 in *PILRA* and rs17851048 in *NTN4* were both observed more frequently in controls than cases. Despite these SNPs not directly altering the amino acid sequence, they had

large effect sizes. Although synonymous substitutions are often considered silent mutations, they can affect splicing through disruption of splicing sequence motifs within exons (Goren *et al.,* 2006). *In silico* predictions performed using Human Splicing Finder suggested that the rs2405442 mutation was likely to disrupt splicing by creation of an exon splicing silencer (ESS) site. This can have a profound effect on the resulting protein by introducing new splice sites, modifying the pattern of exon inclusion or skipping, intron retention and altering the balance of splice isoforms (Maquat, 2001; Singh and Cooper, 2012). Synonymous variants can also affect the secondary RNA structure to affect the accessibility to splice enhancer and silencer sequences (Sauna and Kimchi-Sarfaty, 2011). A recent study examined microglial transcription factor PU.1 binding sites within AD susceptibility loci. Interestingly, *PILRA* was identified as a PU.1 target. The study found that altering the expression of downstream PU.1 targets may be critical in AD pathogenesis (Huang *et al.,* 2017). It is possible that the synonymous mutation found in *PILRA* could alter the expression of the gene through alternate splicing, which could lead to increased susceptibility to AD.

The *NTN4* variant was also predicted to create a new ESS site, with three algorithms supporting this effect. However, a number of caveats exist with splicing prediction programs in that *in silico* programs use databases of SR protein binding motifs to examine whether a mutation falls within these regions (Clement *et al.,* 2016). *In silico*-derived data has to be verified experimentally before any firm conclusions can be made. Splicing activity can be investigated using Minigene assays, by creating wild type and mutant plasmids for transfection (Clement *et al.,* 2016). The splicing products are transcribed into RNA, which can then be synthesised into cDNA and sequenced to determine whether any alternative splicing has occurred.

A total of three variants in *ABCA7* were nominally associated with AD, none of which were in LD with each other or the *ABCA7* GWAS index SNP rs3764650. This suggests that they could be independently

contributing to a genetic burden on AD risk. Two of the SNPs are synonymous mutations, with rs3752237 predicted to potentially affect splicing by creating an ESS site. Common missense variant rs3764645 has previously been tested for association with EOAD and found to be significantly more prevalent in control individuals, supporting the protective effect direction observed for LOAD in this study (De Roeck *et al.,* 2017). SIFT prediction indicated that the mutation is tolerated. *ABCA7* is a known risk gene for AD, therefore these tentative associations provide further support for its involvement in AD.

Variants of interest were validated using the IGAP stage 1 GWAS meta-analysis dataset; with a sample size consisting of 17,008 case individuals and 37,154 controls, significant power was achieved to detect disease associations. Therefore, it was predicted that some of the effects observed here could achieve significance in this dataset. *PILRA* SNP rs2405442 showed a moderate significant effect in the protective direction in IGAP, correlating with the results obtained in Table 4.2.

Variants in the genes *NTN4* and *LRRC8E* were not associated with AD in the IGAP dataset, indicating that these are possibly spurious results. *ABCA7* SNP rs3764645 was highly associated with AD; although it did not pass correction for multiple testing, this exceeded the suggestive significance threshold. However, rs3752237 did not segregate with disease and rs4147915 was not genotyped.

Although the associations here do not remain significant when corrected for multiple testing, tentative associations can be made. Bonferroni correction for multiple testing is likely to be too stringent for the number of samples in this study. A larger sample size could further elucidate several of these associations or contrary, render them non-significant. This is possible due to the sample size bias, as there are almost 3 times more AD cases than controls; which is not representative of the sample population. Therefore, some of the observed signals could be spurious results or false positives.

4.5.4. Burden Analysis

Rare variants pose an analytical problem, as low frequency variants require significantly large sample sizes to detect an association with sufficient power (Lee *et al.,* 2014). Gene-based association tests group these rare and common frequency variants together to calculate burden on a trait, which in turn increases power to detect associations. The SKAT-O burden test has the ability to weight protective and risk variants when calculating burden. It can also adjust for covariates unlike the C-alpha test which is imperative in genetic association studies to control for population stratification (Wu et al. 2011; Lee et al. 2013). SKAT-O can also assess the joint and separate effects of rare and common variants within a gene remaining robust to the direction and magnitude of the effect (Neale *et al.,* 2011). Due to the low power to detect rare variant associations, we included all MAF variants in burden analysis to increase power.

Separate rare and common variant burden testing revealed that often the association is mainly driven by one set of variants, rare or common, rather than both. *MYO18B* exemplifies the notion that rare and common variants could both contribute to disease risk, as joint testing revealed a greater association with AD. The gene encodes a protein that may influence intracellular trafficking when in the cytoplasm (Salamon *et al.,* 2003). *LRRC8E* has been shown to regulate Ca^{2+} homeostasis in the endoplasmic reticulum as overexpression resulted in reduced Ca^{2+} release (Ghosh *et al.,* 2017). This implicates a possible involvement for both genes in the endocytic pathway in AD pathogenesis. Pathway analysis using the newly identified genes of interest could also highlight potential interactions with known AD risk genes.

PILRA exerted the most burden on AD, however it appears that synonymous SNP rs2405442 was contributing mostly to this signal as it was also highly associated in single variant analyses and the rare variants exerted only suggestive burden on AD. The *PILRA* gene is

located on chromosome 7, adjacent to known AD risk gene *ZCWPW1*. It has been previously implicated in AD, with GWAS index SNP rs1476679 associated with reduced *PILRA* levels (Allen *et al.,* 2015). The gene is a paired immunoglobulin-like type 1 receptor involved in regulation of the immune response (Mousseau *et al.,* 2000); it is the inhibitory counterpart to activatory receptor *PILRB*, both highly expressed on microglia. *PILRB* has also been reported to associate with signalling adaptor protein *TYROBP*, which forms an immunoreceptor signalling complex with AD risk gene *TREM2* (Karch *et al.,* 2016). This highlights a potential role for *PILRA* in the inflammatory pathway and possible contribution to AD pathogenesis.

The SNP rs2405442 is in strong LD (r^2=0.85) with the *ZCWPW1* SNP in the CEU population, however neither variant directly affects the amino acid sequence. Although it is perhaps less likely to be causative, it could contribute to AD risk by altering gene regulation. Of the two other SNPs in *PILRA*, rs138933650 is a tolerated missense variant and rs34266222 is synonymous. Further investigation is required to determine if any of these are the causal variant or whether there are more variants in LD that could be driving this association. The SNP rs1476679 produced an eQTL with *PILRB* transcripts in human brain homogenates as well as in monocytes, suggesting that this AD risk SNP may influence *PILRB* expression in microglia (Karch *et al.,* 2016). The variant rs2405442 was assessed for its effect on MCI progression to AD, but it was not found to be associated (Lacour *et al.,* 2016). The other SNPs are both rare, missense variant rs138933650 and synonymous variant rs34266222, which were not significant individually.

4.5.5. Polygenic risk scores

Clinical diagnosis of AD is accurate in 85-90% of cases yet remains confounded by factors such as comorbid dementias and other diseases (Naj and Schellenberg, 2017). A definitive diagnosis still requires post-mortem examination of the brain tissue. LOAD has a strong genetic

component, believed to be as high as 80%, yet there remains a considerable amount of heritability unaccounted for (Gatz *et al.,* 2006). It is suggested that combinations of rare and common variants with varying effect sizes may contribute to complex disease risk (Brookes and Morgan, 2017). PRS calculate the total genetic risk in an individual for a phenotype based on the variants present in their genome. PRSice generates risk scores for individuals in a target dataset based on the effect scores of variants from an independent dataset, to identify the best-fit PRS model (Euesden *et al.,* 2014). IGAP association summary data containing over 7 million variants genotyped in 17,008 AD cases and 37,154 controls was used as the base dataset. The WES dataset containing singleton variants was used to utilise the maximum number of variants possible to generate PRS. The overlap of variants was only 49,708 variants excluding 500kb around the *APOE* region, which was small when compared to other studies that use millions of SNPs from meta-GWAS (Escott-Price *et al.,* 2015; International Schizophrenia Consortium *et al.,* 2009). However, the WES dataset incorporates a greater proportion of coding variants, which could elucidate some of the causal variants contributing to AD risk.

4.5.5.1. PRS prediction models

PRSice calculated scores at multiple significance thresholds and identified the best threshold for prediction. The *APOE* region was excluded prior to pruning to identify risk variants independent of the effects of *APOE* ε4 (Chaudhury *et al.,* 2017). Pruning of variants in strong LD was performed using default parameters, removing variants with pairwise r^2>0.8 within a 50kb window in sliding increments of 2kb. LD needs to be considered to prevent multiple variants from the same LD block being regarded as independent and given scores based on the weighted effect. This could result in a higher PRS based on the effect of the index SNP on AD risk, causing bias in the scores (Chaudhury et al. 2017). A more stringent cut-off threshold to retain only independent variants in linkage equilibrium could potentially identify

variants that are more likely to be causative (Dudbridge, 2013). However, when tested using the cut-off $r^2<0.2$, the amount of variance explained by the $P_T<0.26$ was only 2.6%, an increase of 0.4%; meaning that variants in LD were not greatly affecting this value. This could be due to the WES data, as non-coding regions are excluded it is possible that many variants in LD were removed as a result. An alternate option to pruning is clumping of SNPs, which is far more stringent yet more likely to identify the affective variants. Clumping thins SNPs by LD and p-value, retaining only the most significant variant within the window and removing those in LD ($r^2>0.1$) with it. Pruning was the preferred option for this dataset, as clumping would have resulted in a considerably lower number of variants for PRS analysis.

The table in Figure 4.3 (a) showed that $P_T<0.05$ was the best-fit for the prediction model, with the highest Nagelkerke's R^2 value. This pseudo R^2 determines how much of the variance between phenotypes is explained by the model (Euesden *et al.,* 2014). In this case the scores only explain 2.2% of the variation in AD. The closer this value is to 1, the better the model is at explaining the genetic contribution to the disease. The *APOE* locus alone has been shown to contribute greatly to AD risk (Chaudhury et al. 2017). The threshold P<0.05 incorporated 3,326 coding region variants that could explain a small proportion of risk, independent of the variants effects within the *APOE* region. With the addition of other predictor variables this is likely to increase substantially, as established genetic risk factor *APOE* ε4 alone is known to contribute significantly to AD risk (Chaudhury *et al.,* 2017; Escott-Price *et al.,* 2015). These variants also did not include any of the 20 GWAS variants as they are located in non-coding regions and therefore were not captured by the target probes. Therefore, including genotypes for these susceptibility loci as covariates could increase the predictability of the risk model.

Figure 4.3 (b) provides a comparative view of the thresholds in a bar graph generated by PRSice, to highlight the best fit. P<0.05 was

evidently much greater than the other thresholds tested. PRSice provides an automated method for PRS analysis using Plink v1.9, allowing a large number of P_T to be tested simultaneously and identifying the most predictive threshold. It also has the ability to uncover shared genetic aetiology between different traits. PRSice is a useful tool to generate risk scores for large datasets, requiring minimal computational time. It has demonstrated potential for gaining insights into the genetic architecture of complex diseases and assessing the genetic overlap of traits across populations (Euesden *et al.,* 2014).

4.5.5.2. ROC curve for predictive accuracy

In order to isolate the genetic component of AD, accounting for population differences like age and sex is imperative. It is a well-known fact that individuals harbouring one or two *APOE* ε4 alleles have a significantly increased risk of developing AD, although this is not definitive (Yu *et al.,* 2014). Therefore, adjusting for risk factors such as the ε4 allele, age and sex reduces the variance as a result of these differences. The rare ε2 allele has been shown to confer a protective effect against AD and appears to reduce risk by up to 40% (Liu *et al.,* 2013). Manual genotyping of the GWAS SNPs was performed for individuals in the BDR cohort. Genotypes for these variants were also used as covariates to determine their genetic contribution to AD risk. The ROC curve in Figure 4.4 (a) illustrates the accuracy of each prediction model, which is given in the table (b) as AUC.

Each line represents a different prediction model and its ability to discriminate between the two groups. An additive approach was used to determine the increase in predictive accuracy with each covariate added. The *APOE* region alone was found to contribute significantly to AD risk, with previous estimates suggesting that approximately up to 65% of genetic risk could be explained by variants within this 500kb region (Chaudhury *et al.,* 2017). The ε4 allele of *APOE* alone can be used to classify AD cases and controls with 73.1% accuracy, however this only accounts for 10.2% of the variance observed. The ε2 variant

increases this predictive ability and also accounts for a greater proportion of the variance in combination with ε4. Age and sex appear to add little information, despite both being major risk factors for AD. Age is the greatest overall risk factor as AD is commonly known as a disease of old age. In addition, the incidence of AD is higher in women compared to men, not attributed to higher longevity (Farrer, 1997; Viña and Lloret, 2010). This could be explained by the sample distribution bias, as there are 3 times as many cases compared to controls. When incorporating the GWAS loci together, a significant increase was observed in both predictive ability and proportion of variance explaining AD risk. An improvement of 6.7% in predictive accuracy highlighted that when accumulating multiple individual variants of low effects, the combined score could potentially be a strong predictor for disease (Dudbridge, 2013). Inclusion of the PRS $P_T<0.05$ model contributed another 3,326 variants to the prediction model. Although the PRS improved the predictive accuracy by 1% with only an added 1% of the variance explained, it is hoped that an increased sample size will provide greater power to discriminate AD cases and controls (Plomin and Deary, 2015). When combining all the variables; age, sex, *APOE* ε4, ε2, 19 GWAS loci and PRS at $P<0.05$, the greatest predictive accuracy was achieved (Voyle and Hodges, 2017). This model correctly distinguished between case and control samples with 85.3% accuracy. The pseudo R^2 value indicated that by integrating all of these factors, 42.9% of the variance caused in AD versus controls was explained. However, this means that less than 50% of the genetic risk of AD has been identified within this dataset. Nagelkerke's R^2 model-fit value should ideally be closer to 1 to indicate a good-fit model.

A previous study calculating PRS in AD using large meta-scale GWAS analyses obtained an AUC of 78.2% using information from 87,605 variants combined with age and sex (Escott-Price *et al.,* 2015). Although this study acquired significant power due to the large sample sizes and number of overlapping variants, the predictive power and model fit were both lower than the results for the best fit model shown in

Figure 4.4 (b). More recently, the improvement in PRS and power was demonstrated by using pathologically confirmed samples instead of clinically diagnosed cases (Escott-Price et al. 2017). All BDR samples used in this study were classified by CERAD criteria for AD and NIA Alzheimer's Association guidelines (Costello *et al.,* 2016). This could explain why enhanced predictive ability was achieved with a notably smaller dataset compared to the 3,049 AD cases and 1,554 controls (Escott-Price *et al.,* 2015). Additionally, the 3,326 variants that contribute to the PRS are comprised of coding region variants, which are likely to have a more pronounced effect on genetic risk.

Other PRS studies in AD have been conducted using sample numbers in the thousands from meta-analyses of GWAS. The current sample size used for analysis here was small and thus the association between PRS and AD risk was not significant, however had a high predictive accuracy. Although this combined effect of variants has allowed us to tentatively explore the genetic basis of AD, hundreds of thousands of subjects may be required to estimate the gene effects (Dudbridge, 2013). To increase the proportion of genetic variation captured by PRS analysis, an increased number of SNP markers would also undoubtedly improve scores. This could be achieved by greater coverage of variants and in turn generate more accurate risk scores, providing a better predictive model. The inclusion of AD biomarker data could also markedly increase the predictive accuracy, which would be clinically beneficial for identifying at risk individuals before the onset of symptoms (Jack *et al.,* 2013; Tan *et al.,* 2018).

4.5.5.3. Comparison of normalised scores

Raw PRS were normalised and displayed in the boxplot in Figure 4.5 to show the dispersion of scores for case and control individuals. PRSice did not find a significant association between the risk scores and phenotype (P=0.83), however this could be the result of a small sample size. Nevertheless, the spread of scores evidently shows that on average AD cases had a higher PRS than controls. The presence of

controls with high PRS indicates that individuals can have a genetic disposition to a disease but not necessarily develop it. Also, it is likely that these individuals could be harbouring protective variants that have prevented them from developing the disease. It is also entirely possible that these people could have gone on to develop AD at a later stage. High PRS could also indicate risk of other neurological diseases that correlate with AD, that are either comorbid or misdiagnosed. Due to the significant overlap in clinical symptoms, the rate of misdiagnosis of dementia is also high (Beach *et al.,* 2012). These scores could prove useful in determining an individual's genetic contribution to disease. A diverse set of scores could also allow sub-diagnoses of specific groups within the disease cohort based on an individual's genetic profile (Euesden *et al.,* 2014)

Control individuals with a low PRS could be harbouring protective variants that contribute to a reduced risk of developing AD. Interestingly, there were also individuals with AD who have a much lower score than expected. It could be that the individual has a low genetic component and therefore does not possess high-risk variants, suggesting that other lifestyle or environmental factors may have contributed. Healthy controls with high PRS and cases with low PRS could possibly indicate the missing heritability or other external factors affecting AD risk, which are as yet unknown (Chaudhury *et al.,* 2017). However, more extensive genotyping, additional clinical and lifestyle information may be required to further improve predictive ability.

Previous studies with larger sample numbers identified a trend in the PRS between cases and controls. Most controls would fall in the lower deciles, negatively correlating with cases, which showed a gradual increase in number at higher deciles (Chaudhury *et al.,* 2017; Escott-Price *et al.,* 2015). It is likely that this would require more samples to achieve enough power to investigate this. Predictive ability could also be improved by increasing the number of variants in the analysis. Imputation computes the genotypes for SNPs that have not been

genotyped in the data using haplotype information. The imputed SNP may be common to the base dataset if the index SNP was not. Imputing the target data would result in a greater number of variants common to both datasets, however this could also introduce background noise and false positives (Li *et al.,* 2009). Contrary, using only a handful of variants as opposed to thousands would be more cost effective to screen individuals for AD risk markers in a clinical setting. Lacour and colleagues calculated odds ratios for the scores, which is more meaningful than using arbitrary values for the scores and should be employed in future analyses (Lacour *et al.,* 2016).

4.5.5.4. Neuropathology against score

Many individuals presented with comorbid diseases or phenotypes, which could be explained by the mixture of pathologies often observed in the brain post-mortem. It is difficult to determine a case of pure AD since other pathologies can arise during disease progression, such as Lewy bodies or abnormal TDP-43 protein deposits. The large dispersion of PRS in the case group could be attributed to this range in pathology. It is possible that diseases such as DLB or FTLD could share some aetiology with AD, potentially due to genetic pleiotropy, leading to the mixed phenotypes as a result.

There is a significant overlap in scores among cases and controls. This could be representative of the varying degree to which the individuals harbour a genetic component to AD susceptibility. Incidentally, there are controls with high scores, which is indicative of a high genetic risk and could represent controls that could have later developed AD has they lived longer. However, the lack of similarly extensive clinical data for the controls means that many symptoms of reduced cognitive ability and pathology are unknown. Conversely, the low risk AD samples could possess a lesser genetic component to their AD susceptibility. It is possible that individuals could have been exposed to other factors in the environment, for example, that can affect susceptibility to AD. Poor education in early life, loss of hearing, hypertension, obesity, smoking,

depression and diabetes have all been implicated as risk factors for dementia (Livingston *et al.,* 2017). Collecting information on individuals' lifestyle could further enhance predictive ability of the models to account for any effects of environmental factors. Therefore, more extensive phenotyping could allow more of the missing heritability in AD to be explained.

Table 4.4 examined the neuropathological features of some of the samples classed as outliers. Higher Braak staging and CERAD scores were correlated with the case groups, with the controls mostly exhibiting Braak stages 1-3, consistent with the normal ageing process. Braak stages correlate with the spread of NFT throughout the brain and are a direct indicator of early AD progression (Braak *et al.,* 2006). The BDR cohort contains extensive measures of cognition, psychometrics and pathology (Costello *et al.,* 2016). However, more definitive information from BDR will be provided as the cohort size grows. This will allow correlations of greater proportion, such as with depression, MMSE scores and more. Pathway-specific PRS are now also being tested as predictors of early changes in AD related biomarkers and cognitive function (Darst *et al.,* 2017). Although they were not able to increase predictive power over combined PRS or *APOE* alone, they could prove to be useful as more clinical and pathological data is integrated into PRS analyses.

4.5.5.5. Risk genes contributing to PRS

These risk variants were annotated to identify the risk genes potentially contributing to this effect. Many of the genes containing variants used for determining the PRS, were known AD risk genes, shown in Table 4.5. Of the 1,348 genes encompassing the coding region variants, 10 of them were GWAS risk genes. *ABCA7* harbours the most risk variants with 5 SNPs contributing to the PRS. Other genes containing multiple variants include *CD33*, *SORL1*, *CR1* and also *ABCA7* neighbouring genes *CNN2* and *HMHA1*, among others. Interestingly, lesser known genes that have previously been investigated *ACE*, *SPARCL1* and

VAC14. *ACE* has been shown to degrade the Aβ peptide and mediate AD risk through epistatic interactions with other genes (Braae *et al.,* 2015; Kehoe and Wilcock, 2007; Lehmann *et al.,* 2013). *SPARCL1* was identified in CSF along with 6 other biomarkers; they were able to discriminate AD from control samples with more than 84% accuracy (Vafadar-Isfahani *et al.,* 2012). *VAC14* is critical to the regulation of PIKfyve enzymatic activity which is involved in $PI_{3,5}P_2$ synthesis and loss of VAC14 activity results in decreased $PI_{3,5}P_2$ production. Neurons are particularly sensitive to the loss of $PI_{3,5}P_2$ as it is an essential component of intracellular vesicule transport and therefore the loss of this enzyme may lead to neurodegeneration (Currinn *et al.,* 2016; Sbrissa *et al.,* 2007).

Annotation of risk variants revealed that numerous established and tentative AD-related genes contribute to AD risk in this cohort thereby supporting previous findings linking these genes to AD susceptibility. These results indicate that currently established AD susceptibility genes from GWAS must play a role in AD pathogenesis despite the GWAS loci not being the causal variants.

Identifying the variants that increase an individual's risk for a disease, using a risk score model, could be useful for the advent of effective, symptom-specific personalised medicine. Determining the SNPs with greater effects can be useful for targeting the potentially causative variants. PRS have shown their utility in stratification of disease groups, which could help determine the most at risk groups that would benefit for recruitment into clinical trials (Marioni *et al.,* 2017). The derived PRS have demonstrated the ability to calculate individual genetic risk, allowing the prediction of disease status with 83.5% accuracy. Using sequence data brings added complexities; although single-base resolution is beneficial for rare variant inclusion, the total number of variants in the exome is significantly lower than a genotyping array with added non-coding SNPs (Curtis, 2017). Fewer variants would be

incorporated as a result, however they are more likely to exert greater effects, so this may not pose an issue.

4.5.6. Conclusion

Studies investigating rare variant association require extremely large sample sizes due to the power issues that arise with detecting low frequency variants (Zuk *et al.,* 2014). Here we achieve 83% power to detect associations for variants of MAF = 5% with an effect size of 3, which makes it difficult to investigate the role of rare variants in AD with smaller effects. Imputation provides a potential solution, as it would increase the number of rare variants available for association testing and polygenic scoring. This would be valuable for detecting rare variants with moderate effect sizes contributing to AD risk. However, whole-exome imputation can also introduce background noise, potentially confounding any rare variants that might be identified. As an alternative, genes of interest such as known risk genes, could be imputed to increase coverage of low frequency variants. Whole-exome association identified a tentative variant of interest in *PILRA*; this gene is thought to potentially interact with the AD risk genes *ZCWPW1*, *TREM2* and possibly *ABCA7*, warranting further investigation into the gene interaction network.

PRS analysis highlighted the utility of accumulating genetic variants to predict AD risk. A total of 3,326 coding region variants were able to predict almost 45% of the genetic risk in AD with 83.5% accuracy. There remains a large amount of missing heritability, however with a growing sample size the BDR cohort will hopefully be able to achieve increased power. The overlap of scores between cases and controls also shows considerable dispersion within groups. It is possible that this represents individuals with a lesser genetic component contributing to AD susceptibility.

Furthermore, the presence of controls with high PRS but no cognitive impairments highlights the possibility that there are protective variants

involved. Identifying these protective factors has positive implications for future treatments. This could also be useful for identifying cognitively healthy individuals at risk of developing AD in the future.

Many of the case samples were diagnosed with comorbid dementias, which could have common risk factors with AD. This supports the idea that AD is on a spectrum of neurodegenerative diseases. This does show that many factors influence disease risk aside from genetics and that individuals can have varying levels of genetic predisposition to AD. PRS also incorporates many known AD risk genes, including several of the GWAS signals, indicating that these genes are highly likely to contribute to AD pathogenesis.

Chapter 5: Candidate Genes

5.1 Introduction

Technological advances have allowed increasing numbers of variants to be tested in case-control studies to investigate genetic markers of AD susceptibility (Amos *et al.,* 2011). To date, over 30 genetic variants have been associated with the risk of developing AD from GWAS, NGS and SNP genotyping arrays. GWAS identified 20 new susceptibility loci which provided a number of candidate genes for investigation (Harold *et al.,* 2009; Naj *et al.,* 2011; Seshadri *et al.,* 2010b). These common SNPs only exert small effect sizes, therefore closer examination of rarer variants in the associated genes could help identify the true causal variants. Targeted NGS enables simultaneous screening of a battery of genes already implicated in disease. This approach focuses the analysis on a selection of putative gene candidates that are associated with the disease in some way. In this way candidate gene studies can have greater statistical power than GWAS, which use large numbers of variants, if sample sizes provide enough power (Amos *et al.,* 2011). Significance thresholds for such studies are also less stringent due to fewer multiple comparisons (Patnala *et al.,* 2013). However, candidate gene association studies are limited by their reliance on prior knowledge of the causative genes. They have also been criticised for poor replication of results; often caused by population stratification, which can be overcome by validating results in different populations (Zhu and Zhao, 2007).

5.1.1. Pathway analysis

New candidate genes identified by GWAS and NGS have implicated several new pathways in AD pathogenesis, such as lipid transport, immune function and inflammation and endocytosis (Lambert *et al.,* 2013). There is considerable overlap with genes being involved in multiple pathways, which highlights several potential mechanisms of AD manifestation. Pathway analysis tools are commonly used to identify

biological interactions and functional pathways that connect a number of genes. They use databases of known gene and protein-protein interactions to find potential biological links between genes (Khatri *et al.,* 2012). By highlighting common functions, we can decipher if genes of interest fit into the new or existing pathways associated with AD pathogenesis.

5.1.2. Alzheimer's risk genes

Identification of the new susceptibility loci by GWAS provided new insights into the possible biological mechanisms underlying AD pathogenesis (Lambert *et al.,* 2013). These common pathways can present new targets, enabling the potential development of more effective treatments and diagnostic tests. The genes fall into four major pathways, immune function, lipid and cholesterol metabolism, synaptic function and endocytosis and also APP processing (Jones *et al.,* 2010; Morgan, 2011). Less is known about the functional effects of some of the later genes identified by GWAS, however, they were associated with novel pathways that are also likely to be involved in AD pathogenesis, such as axonal transport and tau metabolism. There is significant crossover between the pathways, with many of the genes being involved in several different functions. These genes are discussed in more detail here.

Cholesterol metabolism and lipid transport

APOE is a key component of lipoprotein complexes, aiding cholesterol transport and metabolism in the CNS (Yu *et al.,* 2014). *In vitro* studies have demonstrated that intracellular cholesterol levels can regulate Aβ production through APP processing (Michikawa, 2003). Genetic studies have identified several genes related to cholesterol synthesis, transport or metabolism, further implicating this pathway in AD pathogenesis.

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

ABCA7 is a member of the ATP-binding cassette genes that facilitate lipid transport across cell membranes in the CNS (Karch *et al.,* 2012). It functions as a transporter in the efflux of phospholipids into APOA-I and APOE complexes by producing high-density lipoproteins (HDL) (Tanaka *et al.,* 2011). *ABCA7*-deficient APP transgenic mice exhibit greater Aβ deposition than controls (Chan *et al.,* 2008). This was confirmed by later studies, which also demonstrate that *ABCA7* expression significantly correlates with neuritic plaque burden (Kim *et al.,* 2013). However, loss or reduction of ABCA7 only exerted modest effects on cholesterol transport, suggesting that ABCA7 is not essential in lipid homeostasis (Iwamoto *et al.,* 2006; Kim *et al.,* 2013). *ABCA7* is abundantly expressed in the hippocampus, one of the earliest affected regions in AD, and also microglia, which are known to play a role in phagocytosis (Zhao, Yu, *et al.,* 2015). Knockdown of *ABCA7* in microglia resulted in reduced clearance of apoptotic cells, indicating that the gene could regulate Aβ levels through its role in immune function rather than lipid transport (Jehle *et al.,* 2006).

***CLU* (Clusterin)**

CLU, also referred to as apolipoprotein J, is a molecular chaperone that is abundantly expressed in the brain. Much like *APOE*, it has a role in cholesterol metabolism and lipid transport and has been shown to form complexes that facilitate the removal of Aβ across the blood-brain barrier (Hardy *et al.,* 2011; Nelson *et al.,* 2017). *CLU* was shown to regulate Aβ deposition in conjunction with *APOE*, demonstrated through increased Aβ load in APOE-CLU knockout PDAPP mice than single gene knockouts. APOE-CLU knockout mice also displayed elevated CSF and ISF Aβ levels (DeMattos *et al.,* 2004). AD patients were found to have elevated levels of CLU in the cortex and hippocampus compared to healthy individuals, with detectable levels found within diffuse amyloid plaques (Karch *et al.,* 2012). Elevated CLU plasma levels are also associated with hippocampal atrophy, correlating with

cognitive decline and disease progression in individuals with AD (Schrijvers *et al.,* 2011; Thambisetty *et al.,* 2010). The gene is also involved in immune function where it regulates the membrane attack complex by inhibiting the inflammatory response associated with complement activation (Nuutinen *et al.,* 2009).

***SORL1* (Sortilin related receptor L)**

SORL1 was originally identified as an AD risk gene over a decade prior to the meta-GWAS, with decreased expression associated with the disease (Scherzer *et al.,* 2004). The gene belongs to a family of sorting receptors that contain a vacuolar protein sorting (VPS-10) domain. *SORL1* has been associated with both endocytosis and Aβ metabolism (Lambert *et al.,* 2013). By binding LDL, it mediates endocytic uptake of lipoproteins, including APOE-containing particles (Rogaeva *et al.,* 2007). SORL1 determines whether APP is sorted in the retromer recycling-endosome pathway or the endosomal-lysosome pathway where it is cleaved to generate Aβ (Vardarajan *et al.,* 2015). Therefore impaired interactions between SORL1 and APP can lead to increased Aβ secretion, possibly due to the reduced ability to direct APP to the endosomal recycling pathway (Vardarajan *et al.,* 2015). SORL1-deficient mice have elevated Aβ levels (Dodson *et al.,* 2008), which is consistent with reduced expression of the gene in AD brains (Scherzer *et al.,* 2004).

Endocytosis and synaptic function

Endocytosis is central to AD pathogenesis as APP, Aβ and APOE are internalised through this pathway. Therefore, alterations in APP trafficking can directly influence proteolytic cleavage by secretase enzymes in the endosomes (Huang and Mucke, 2012). Furthermore, synaptic activity and neurotransmitter release are also disrupted in AD (Holtzman *et al.,* 2011). Several genes involved in endocytosis and synaptic function were identified by GWAS. *SORL1* was initially recognised for its role in APP metabolism, some of which occurs in the

endosomal pathway; highlighting the overlap in AD pathways (Morgan, 2011).

BIN1 **(Bridging integrator 1)**

BIN1 has been associated with multiple pathways, including endocytosis, intracellular trafficking, APP processing, inflammation, calcium homeostasis and apoptosis (Tan *et al.,* 2013). *BIN1* is a member of the Bin/amphiphysin/Rys (BAR) adaptor family that are implicated in vesicle fusion and trafficking. Isoforms expressed in the CNS are thought to be involved in clathrin-mediated endocytosis and synaptic vesicle recycling (Pant *et al.,* 2009). Amphiphysin 1 knockout in mice caused a parallel reduction of BIN1, with mice exhibiting defects in synaptic vesicle recycling (Di Paolo *et al.,* 2002). The gene may also play a role in tau processing as it is known to interact with another microtubule-associated protein, CLIP-170 (Meunier *et al.,* 2009). There is some evidence to suggest BIN1 physically interacts with tau *in vitro* and also in mouse brains (Chapuis *et al.,* 2013), however there are conflicting reports on the relationship between BIN1 expression and tau neurotoxicity. While *BIN1* knockdown causes significant suppression of tau neurotoxicity other experimental evidence suggests that a *Drosophila* orthologue of BIN1 is able to reduce tau neurotoxicity *in vitro* (Chapuis *et al.,* 2013). Similarly, elevated *BIN1* mRNA expression levels have also been linked to later age at onset of AD and shorter disease duration, which suggests a protective effect (Karch *et al.,* 2012). Therefore, the exact role of *BIN1* in AD pathogenesis is still unclear.

PICALM **(Phosphatidylinositol binding clathrin assembly protein)**

PICALM encodes a protein that is essential for clathrin assembly; it recruits clathrin and adaptor protein complex 2 (AP2) to the cell membrane where it can regulate clathrin cage size (Baig *et al.,* 2010). Predominantly expressed in neurons, it is involved in the trafficking of APP and synaptic vesicle protein VAMP2, which regulates

neurotransmission at the presynaptic membrane (Baig *et al.,* 2010; Schnetz-Boutaud *et al.,* 2012). *PICALM* is proposed to affect AD risk by modulating APP trafficking via CME which in turn affects the production, transport and clearance of Aβ (Xu *et al.,* 2015). Evidence for this was shown by disrupting *PICALM* expression which altered APP trafficking *in vitro* and overexpression in AD transgenic mice resulted in increased plaque deposition (Xiao *et al.,* 2012). Knockdown of the gene in APP/PS1 mice had the opposite effect, reducing Aβ production and plaque burden, indicating the *PICALM* may play a role in plaque formation (Xiao *et al.,* 2012). Furthermore, PICALM has been shown to co-localise with NFTs in AD cases, implicating a role in tau pathology (Ando *et al.,* 2013).

***CD2AP* (CD2-associated protein)**

CD2AP is a cytoplasmic, scaffolding molecule involved in cytoskeletal reorganisation, endocytosis and intracellular trafficking (Lynch 2003). It has been demonstrated that lysosomal function is impaired in CD2AP-deficient mouse cells, indicating that *CD2AP* is a critical regulator of vesicle trafficking to lysosomes and is integral to the formation of synapses (Dustin *et al.,* 1998; Karch *et al.,* 2012). RNA interference-mediated disruption of *CD2AP* orthologue *cindr* in *Drosophila*, enhanced tau toxicity, suggesting that like *PICALM*, it may also be involved in modulating tau (Shulman *et al.,* 2014).

***EPHA1* (Ephrin type A receptor 1)**

EPHA1 encodes a member of the ephrin family of tyrosine kinase receptors, predominantly expressed by lymphocytes and monocytes. The receptor binds to membrane-bound ephrin-A1 ligands with high affinity, facilitating bidirectional signalling on adjacent cells (Sharfe *et al.,* 2008). This interaction is thought to play a role in regulating synapse formation and synaptic plasticity (Lai and Ip, 2009). *EPHA1* is also involved in cell and axonal guidance (Martínez *et al.,* 2005), with additional roles in the regulation of apoptosis and inflammation also proposed (Ivanov and Romanovsky, 2006; Kullander and Klein, 2002).

Immune function and inflammation

Neuroinflammation is a hallmark of AD, with significant evidence suggesting that dysregulation of the immune response and activation of inflammatory pathways contribute to pathogenesis (Heneka *et al.,* 2015; Holtzman *et al.,* 2011). Microglia are a major component of this pathway, which act to clear cellular debris to reduce the accumulation of toxic products such as Aβ (Streit and Xue, 2012). Extensive evidence demonstrates that amyloid plaques are surrounded by activated immune cells (Hohsfield and Humpel, 2015). Therefore, polymorphisms in the genes linked to the immune pathway could affect the normal clearance of Aβ, leading to increased deposition and amyloid plaques.

CR1 **(Complement receptor 1)**

CR1 is a member of the complement receptor family that negatively regulates complement activation, encoding a receptor of C3b. *CR1* is expressed on phagocytes and microglia in the brain, where it acts to ingest complement-activated particles and cellular debris (Liu and Niu, 2009). It is hypothesised that $A\beta_{42}$ in amyloid plaques and tau in NFTs activate the complement system, triggering an inflammatory response (Karch *et al.,* 2012). Thus greater expression of CR1 protein indicates a higher clearance rate of immune complexes, however the clearance of plasma $A\beta_{42}$ is dependent on CR1 binding to C3b (Fonseca *et al.,* 2016). Expression of complement factors are reported to be upregulated in the affected regions of AD brains which correlates with the notion that complement activation triggers inflammation eventually leading to cell damage (Shen *et al.,* 2013). Accordingly, increased expression of *CR1* is associated with advanced cognitive decline (Karch *et al.,* 2012).

CD33 **(Sialic acid-binding immunoglobulin-like lectin)**

CD33 is a transmembrane protein expressed on cells of myeloid lineage, mainly on microglia. It belongs to a class of immune cell surface receptors that bind sialic acid, known as sialic acid-binding

immunoglobulin-like lectins (Siglecs). Gene expression profiling has revealed that the gene is also involved in innate immunity and endocytosis (Crocker *et al.,* 2007). Similar to *PICALM* and *CR1*, *CD33* could share a similar disease mechanism, by triggering immune cell interactions through clathrin-independent endocytosis (Rosenthal and Kamboh, 2014). The expression of *CD33* was increased in AD brains with a greater number of CD33-positive microglia, which correlates with increased amyloid plaque burden and disease severity (Griciuc *et al.,* 2013; Karch *et al.,* 2012). Exon 2 in *CD33* encodes the IgV domain that mediates sialic acid binding, which results in the inhibition of phagocytosis (Villegas-Llerena *et al.,* 2016). Splicing of *CD33* influences microglial activation, with loss of exon 2 associated with reduced AD risk by eliminating the inhibitory effect on Aβ phagocytosis (Malik *et al.,* 2013). Thus, CD33 is likely to play an important role in microglia-mediated Aβ clearance in the brain (Griciuc *et al.,* 2013).

***TREM2* (Triggering receptor expressed on myeloid cells 2)**

TREM2 is a myeloid cell surface receptor expressed on microglia, which has an anti-inflammatory function in the brain (Rohn, 2013). It interacts with TYROBP to trigger phagocytosis, which is critical in clearing excess Aβ peptide thereby preventing the build-up of amyloid plaques (Hickman and El Khoury, 2014). The association with TYROBP also mediates downstream signalling for suppression of inflammation reactivity (Colonna, 2003). Expression of *TREM2* in the brains of AD transgenic mice is upregulated, with the cells surrounding Aβ plaques expressing high levels of the gene (Matarin *et al.,* 2015).

Mutations in this gene are one of the most significant findings to reinforce the role of immune dysfunction in AD pathogenesis (Villegas-Llerena *et al.,* 2016). Two independent research groups sequenced AD exomes and simultaneously identified a rare variant in *TREM2* significantly associated with AD risk (Guerreiro, Wojtas, *et al.,* 2013; Jonsson *et al.,* 2013). With a frequency of 0.2% and effect size of 3-fold (Abduljaleel *et al.,* 2014), the variant was significantly associated with

increased AD risk, also being observed in only case samples. The rare missense mutation rs75932628 was found to cause an R47H substitution, potentially impairing the phagocytic function of the gene in microglia and likely to contribute to the Aβ pathology seen in AD through reduced clearance (Lue *et al.,* 2015). Previously, mutations in *TREM2* have been associated with Nasu-Hakola disease; an autosomal recessive form of dementia that presents with bone cysts and fractures at a young age (Paloneva *et al.,* 2002). Autosomal recessive mutations have also been identified in an FTD-like syndrome without bone involvement (Guerreiro, Lohmann, *et al.,* 2013).

Coding variants in other genes in the *TREM* locus have also been implicated as risk factors for AD. *TREM1* has also been shown to facilitate microglial phagocytosis of Aβ like *TREM2*, with mutations in the gene impairing this function (Saadipour, 2017). Missense variants in neighbouring gene *TREML2* have been associated with protection against AD (Benitez *et al.,* 2014). Opposing gene expression between *TREM2* and *TREML2* was found in primary mouse microglia (Zheng *et al.,* 2016), which may have further implications for modulating the inflammatory responses induced by microglia.

***MS4A* locus (Membrane-spanning 4-domains subfamily A)**

The MS4A gene cluster is a subfamily of transmembrane proteins that have been linked to Aβ generation, endocytosis and immune modulation by regulating calcium homeostasis (Giri *et al.,* 2016; Villegas-Llerena *et al.,* 2016). Several genes within the *MS4A* cluster have been implicated in AD, including *MS4A4A*, *MS4A4E* and *MS4A6A* (Hollingworth *et al.,* 2011; Naj *et al.,* 2011). Elevated levels of *MS4A4A* and *MS4A6A* have both been linked to increased risk of AD (Giri *et al.,* 2016; Proitsi *et al.,* 2014). Greater levels of *MS4A6A* in AD brains were also associated with increased Braak tangle and plaque scores, indicating advanced disease pathology (Karch *et al.,* 2012).

INPP5D (Inositol polyphosphate-5-phosphatase D)

INPP5D is a member of the inositol polyphosphate-5-phosphatase family and is primarily expressed in macrophages and microglia in the brain along with CD2AP (Efthymiou and Goate, 2017). The gene hydrolyses the 5' phosphate from phosphatidylinositol-3,4,5-triphosphate and inositol-1,3,4,5-tetrakisphosphate, thereby participating in various signalling pathways, where it is involved in a number of inflammatory responses. *INPP5D* in particular is a key regulator of cytokine signalling and the PI3K-driven oncogenic pathway (Leung *et al.,* 2009; Metzner *et al.,* 2009). INPP5D was found to inhibit TREM2 signalling by binding to the DAP12 adaptor protein and preventing PI3K recruitment to DAP12 and its activation (Peng *et al.,* 2010). An increase in blood INPP5D levels was linked to a greater risk of haemorrhagic transformation in ischemic stroke patients, which is associated with increased blood-brain barrier permeability (Jickling *et al.,* 2013). This mechanism provides additional insight into the role of *INPP5D* in AD pathogenesis.

HLA-DRB1/HLA-DRB5 (Major histocompatibility complex, class II DR Beta 1/5)

HLA-DRB1/HLA-DRB5 are members of the major histocompatibility complex II and play an important role in the CNS immune response (Villegas-Llerena *et al.,* 2016). Highly expressed on reactive microglia, the proteins are implicated in antigen presentation, inflammation and the complement cascade. The HLA locus has been associated with both multiple sclerosis (Sawcer *et al.,* 2011) and PD (Ahmed *et al.,* 2012; Saiki *et al.,* 2010). Knockout of MHC II in mice uncovered a protective mechanism against α-synuclein induced neurodegeneration (Harms *et al.,* 2013). This locus may play a similar role in AD by regulating inflammatory responses.

Newer susceptibility genes

Some of the newer genes identified by the GWAS meta-analysis have relatively unknown function. Many of the genes fit into the aforementioned pathways that are altered in AD.

CELF1 (CUGBP Elav-like family member 1), *CASS4* (Cas scaffolding protein family member 4) and *NME8* (NME/NM23 family member 8) have been implicated in Aβ metabolism, cytoskeletal function and axonal transport (Reitz, 2014). *NME8* has also been linked to proliferation and differentiation of neuronal cells (Liu *et al.,* 2014). *CELF1* is thought to regulate pre-mRNA splicing, mRNA editing and translation (Beisang *et al.,* 2014). In addition, loss of function in *CELF1* homolog *aret* enhanced tau retinal toxicity in *Drosophila*, with gain-of-function mutations suppressing this effect (Shulman *et al.,* 2014).

CASS4 is also involved in tau metabolism and has been associated with neuritic plaques burden and neurofibrillary tangles in AD brains (Beecham *et al.,* 2014). *FERMT2* (Kindlin 2) is another gene that has been linked to the modulation of tau toxicity in *Drosophila* (Shulman *et al.,* 2014).

MEF2C (MADS box transcription enhancer factor 2) is involved in neurogenesis and regulation of microglial development and function (Lavin *et al.,* 2014). It has been linked to the innate immune response in *Drosophila* where it acts as a central transcriptional component (Clark *et al.,* 2013). Additionally, *MEF2C* expression was associated with cognitive decline in Japanese AD patients, alluding to its use as a potential biomarker (Sao *et al.,* 2017).

PTK2B (Protein tyrosine kinase 2 beta) is involved in calcium-induced regulation of ion channels and activates the MAPK signalling pathway (Pandey *et al.,* 1999). It is required for synaptic function in the CNS and also modulates long-term potentiation in hippocampal neurons (Huang *et al.,* 2001). A recent study observed strong interactions between

PTK2B and tau, with PTK2B increasingly accumulating in tau-positive NFTs with AD progression. This suggests that it is a modulator of tau pathology and could be an early marker of disease (Dourlen *et al.,* 2017).

SLC24A4 (Solute carrier 24 member A4) encodes a potassium-dependent sodium/calcium exchanger which has been linked to AD through its roles in DNA methylation and lipid metabolism (Yu *et al.,* 2015). It has also been implicated in neural development (Larsson *et al.,* 2011). Neighbouring gene *RIN3* interacts with *BIN1* to play an important role in the early endocytic pathway (Kajiho *et al.,* 2003).

The function of *ZCWPW1* (Zinc finger CW type with PWWP domain 1) is unknown, however it may be involved in epigenetic regulation by histone modifications (He *et al.,* 2010). *ZCWPW1* GWAS variant forms an eQTL with *GATS*, *PILRB* and *TRIM14*, affecting the binding of RFX3 and CTCF (Beecham *et al.,* 2014; Karch *et al.,* 2016).

5.1.3. Screening for causative mutations

Mendelian forms of neurodegenerative disorders result from autosomal dominant or recessive mutations in a single gene (Mahdieh and Rabbani, 2013). The familial form of AD commonly aggregates within families with an autosomal dominant pattern of inheritance. It is linked to mutations in one of three genes *APP*, *PSEN1* and *PSEN2*, which are rare but highly penetrant (Tanzi and Bertram, 2001). Several hundred variants have been identified, with the majority of causative variants located in *PSEN1*. Individuals with these mutations are almost certain to develop fEOAD (Bertram *et al.,* 2010). This makes NGS an ideal tool for screening candidate genes as the rare, singleton variants can be detected due to single base resolution. The AD & FTD mutation database and PD mutation database document known causative mutations identified in AD, FTD, ALS and PD (Lindquist *et al.,* 2009). Prion disease mutations were gathered from the HPPDB (http://www.mad-cow.org/prion_point_mutations.html). Screening for

causative mutations in genes linked to other disorders could be useful in classifying potentially misdiagnosed cases. Genes in PD (*GBA*, *LRRK2*, *MAPT*, *PARK2*, *PARK7*, *PINK1*, *SNCA*), FTD (*C9orf72*, *CHMP2B*, *GRN*, *VCP*), ALS (*FUS*, *SOD1*, *TARDBP*) and Prion disease (*PRNP*) have been found to harbour predicted pathogenic mutations linked to disease susceptibility.

5.1.4. Genetic pleiotropy

There are several overlapping clinical and neuropathological features across different neurodegenerative diseases. Misdiagnosis between AD and other dementias occurs in 17-30% of cases (Beach *et al.,* 2012), raising the question of whether genetic risk factors may play a role in more than one disease. Interestingly, studies have demonstrated that numerous neurodegenerative disorders can share aetiology via common biological pathways (Armstrong, 2012). This highlights the potential role of genetic pleiotropy, where one gene can influence multiple seemingly unrelated phenotypes. Therefore, mutations in a pleiotropic gene can have an impact on several disease traits due to the heterogenic nature of neurodegenerative disorders.

DLB shares clinical features with AD and PD, often resulting in misdiagnosis. PD and DLB are both synucleinopathies that present with alpha-synuclein deposits in the brain. On the other hand, major AD risk factor *APOE* ε4 also increases risk of developing DLB (Bras *et al.,* 2014), demonstrating that genetic risk factors can contribute to more than one disease.

FTD and ALS also share common neuropathology and genetic features, despite being two distinct diseases. This is evidenced by pleiotropic effect of multiple genes such as *C9orf72*, *FUS*, *MAPT* and *TARDBP*; mutations in all of these genes have been linked to familial forms of the diseases (Ferrari *et al.,* 2011). Implications are that they share common biological pathways. The overlap in genotypes and phenotypes across multiple disorders also indicates that they are likely to be part of a

spectrum of neurodegeneration, as a ‘pure’ form of any of the diseases is very unlikely.

5.2 Aims

Following on from using an exome-wide approach to identify putative candidate genes, we performed pathway analysis to search for biological relationships with known risk genes. We also explored the role of pleiotropy in AD by screening for mutations in familial genes linked to other neurodegenerative diseases. Testing for association and burden of variants within known risk genes may also provide more useful insight as the genes have already been implicated in neurodegenerative disease.

5.3 Methods

5.3.1. Pathway analysis

Pathway analysis was performed using STRING (Szklarczyk *et al.,* 2017) to explore the interactions of candidate genes and also identify biological pathways enriched for genes associated with AD. This method is useful to identify related genes within pathways and visualise the interactions using information from curated databases. The tools construct interaction networks from databases of protein-protein networks and biological function annotations. A gene list was created from exome-wide association, burden analysis results and PRS risk genes, comprising the most associated genes. These were uploaded to the STRING tool, along with known AD risk genes, to search for interactions and enrichment in AD-related pathways.

STRING integrates several different annotations to identify protein-protein interactions with high confidence. Data mining is performed through the curated databases Biocarta, BioCyc, GO, KEGG and Reactome. Experimental evidence is taken from available protein interaction datasets obtained from Protein-Protein Interaction databases and co-expression datasets. A list of genes one per line was uploaded

to the STRING online tool at https://string-db.org. This output a network of protein interactions with different coloured lines to denote the evidence supporting the link, explained as follows:

- Red line – indicates presence of fusion evidence
- Green line – neighbourhood evidence
- Blue line – co-occurrence
- Purple line – experimental evidence
- Yellow line – text mining evidence
- Light blue line – curated database evidence
- Black line – co-expression evidence

5.3.2. Screening for causative mutations

Screening was carried out to determine whether any individuals harboured causative mutations in genes associated with a range of familial neurodegenerative disorders. A total of 17 genes were examined from fEOAD (*APP*, *PSEN1* and *PSEN2*), FTD and ALS (*C9orf72*, *CHMP2B*, *FUS*, *GRN*, *MAPT*, *TARDBP* and *VCP*), PD (*GBA*, *LRRK2*, *PARK2*, *PARK7*, *PINK1* and *SNCA*) and Prion disease (*PRNP*). Known and predicted pathogenic variants were catalogued from various sources, such as the AD, FTD and PD mutation databases (Cruts *et al.,* 2012; Nuytemans *et al.,* 2010). Variants in these genes were extracted using PLINK, which required a file comprising the chromosomal range for each gene; these are listed in Appendix 7.4. The dataset prior to removal of singletons and filtering for coding consequence variants was used to ensure rare and UTR mutations were not overlooked.

```
plink --bfile 13_bdr_final_qc \
--extract FamilialGenes.txt \
--range \
--make-bed \
--out 14_bdr_FamilialGenes
```

Options are:

- --bfile: prefix of the binary plink files for input

- --extract: extract variants using the file provided
- --range: chromosomal coordinates within which to extract variants

5.3.3. Association analysis

Variants within candidate genes were extracted from the WES data for single variant analysis to increase the statistical power to detect associations. This included both familial neurodegenerative genes listed in Section 5.3.1, *APOE* and 23 known AD risk genes identified through GWAS (*ABCA7*, *BIN1*, *CASS4*, *CD2AP*, *CD33*, *CELF1*, *CLU*, *CR1*, *EPHA1*, *FERMT2*, *HLA-DRB1*, *HLA-DRB5*, *INPP5D*, *MEF2C*, *MS4A6A*, *NME8*, *PICALM*, *PTK2B*, *SLC24A4*, *RIN3*, *SORL1*, *ZCWPW1*) and NGS (*TREM2*). Variants were extracted using the chromosomal coordinates in Appendix 7.4. The dataset containing only coding variants without singletons was used for association and burden testing of candidate genes.

```
$ plink --bfile 13_bdr_final_qc \
--extract AllKnownGenes.txt \
--range \
--make-bed \
--out 15_bdr_AllCandidateGenes
```

A logistic regression association test was performed in PLINK with adjustment for the covariates age at death, sex and number of *APOE* ε2 and ε4 alleles. The upper and lower 95% confidence intervals for the odds ratio were also calculated.

```
$ plink --bfile bdr_AllCandidateGenes \
--logistic \
--ci 0.95 \
--covar BDR_covar.txt \
--covar-name Sex,cal_AAD,APOEe4,APOEe2,Braak \
--hide-covar \
--out bdr_logistic_allKnownGenes
```

Options are:

- --logistic: perform logistic regression test
- --ci: include confidence intervals (95%)

- --covar: covariates file
- --covar-name: name of covariates to adjust for
- --hide-covar: hide covariate association results for each variant

5.3.4. Burden analysis

A SKAT-O burden test was applied for all candidate genes, using the SKAT package in R. The SKAT-O algorithm calculates the burden of multiple variants in a SNP set; this is weighted by effect size to account for both protective and risk variants within a gene. SKAT-O required a gene:marker set file containing variant sets for each gene. This was generated with the script Burden2SetID.py (Appendix 7.2). The resulting file comprised 419 variant markers binned into a total of 38 genes containing at least 1 variant.

```
$ python Burden2SetID.py <PseqGenesList.txt> <SKATSetID.txt>
```

SKAT-O was performed whilst correcting for sex, age and *APOE* ε2 and ε4 dosage. A minimum of 2 variants passing quality control were required to effectively test genetic burden on AD. The Bonferroni-corrected significance threshold was calculated as P=0.001; P-values were also adjusted for false discovery rate (FDR) to reduce the stringency.

```
$ Rscript BurdenAnalysisSkato.R
```

Options are:

- BurdenAnalysisSKATO.R: SKATO script

5.3.5. Sequencing the *TREM2* R47H mutation

Individuals were also screened for the rare *TREM2* variant rs75932628, which causes an R47H substitution. This mutation has previously been associated with LOAD, however it is too rare to detect through single variant association analyses. Samples were screened for this SNP in the sequencing data and verified using Sanger sequencing as

described in Section 2.1.5. Primers were designed using the approach in Section 2.1.3, based on the reference sequence in NCBI (Gene ID 54209) with approximately 200bp each side around the SNP locus. Primer-BLAST was used to determine binding specificity by identifying all potential PCR products. Primers were only selected if they did not generate any non-specific products. The forward and reverse primer sequences are given in Table 5.1. DNA was amplified using PCR for any individuals harbouring the mutation as well as a number of wild type samples, all samples were Sanger sequenced to confirm the genotypes.

5.4 Results

5.4.1. Pathway analysis

Pathway analysis was performed using STRING with the genes of interest identified from association and burden analysis in Chapter 4. Many of the new tentatively associated genes connect with known AD risk genes (Figure 5.1). Several genes from the AD GWAS were connected through co-expression or experimental evidence. *ABCA7*, *CD2AP*, *SORL1*, *PICALM*, *BIN1*, *RIN3*, *FERMT2*, *EPHA1* and *INPP5D* were directly linked to each other either experimentally or via information from curated databases. Additionally, the network appears to segregate into five distinct clusters of interactions, denoted by the coloured circles in the figure. The central cluster of genes, including numerous known AD risk genes, bridges with the other clusters. Many of these genes have also been implicated in APP processing and Aβ metabolism, such as *APOE*, *ABCA7*, *SORL1* and *CLU*. Another common biological pathway is endocytosis which involves *BIN1*, *PICALM*, *CD2AP*, *CD33* and *SORL1* again. *APOE* was found to interact metabolism, such as *APOE*, *ABCA7*, *SORL1* and *CLU*. Another common biological pathway is endocytosis which involves *BIN1*, *PICALM*, *CD2AP*, *CD33* and *SORL1* again. *APOE* was found to interact directly with *SCARB1* both *in vivo* and *in vitro*. *SCARB1* is known to

Table 5.1. Primer sequences for *TREM2* R47H mutation. PCR amplification was performed using the primer sequences above as described in Section 2.1.4.

Primer	Sequence	T_m (°C)	GC content (%)	Amplicon size (bp)
R47H_Forward	CTCTGGCACTGGTAGAGACCC	63.7	61.9	445
R47H_Reverse	CACAAACACATGCTGTGCCA	57.3	50.0	

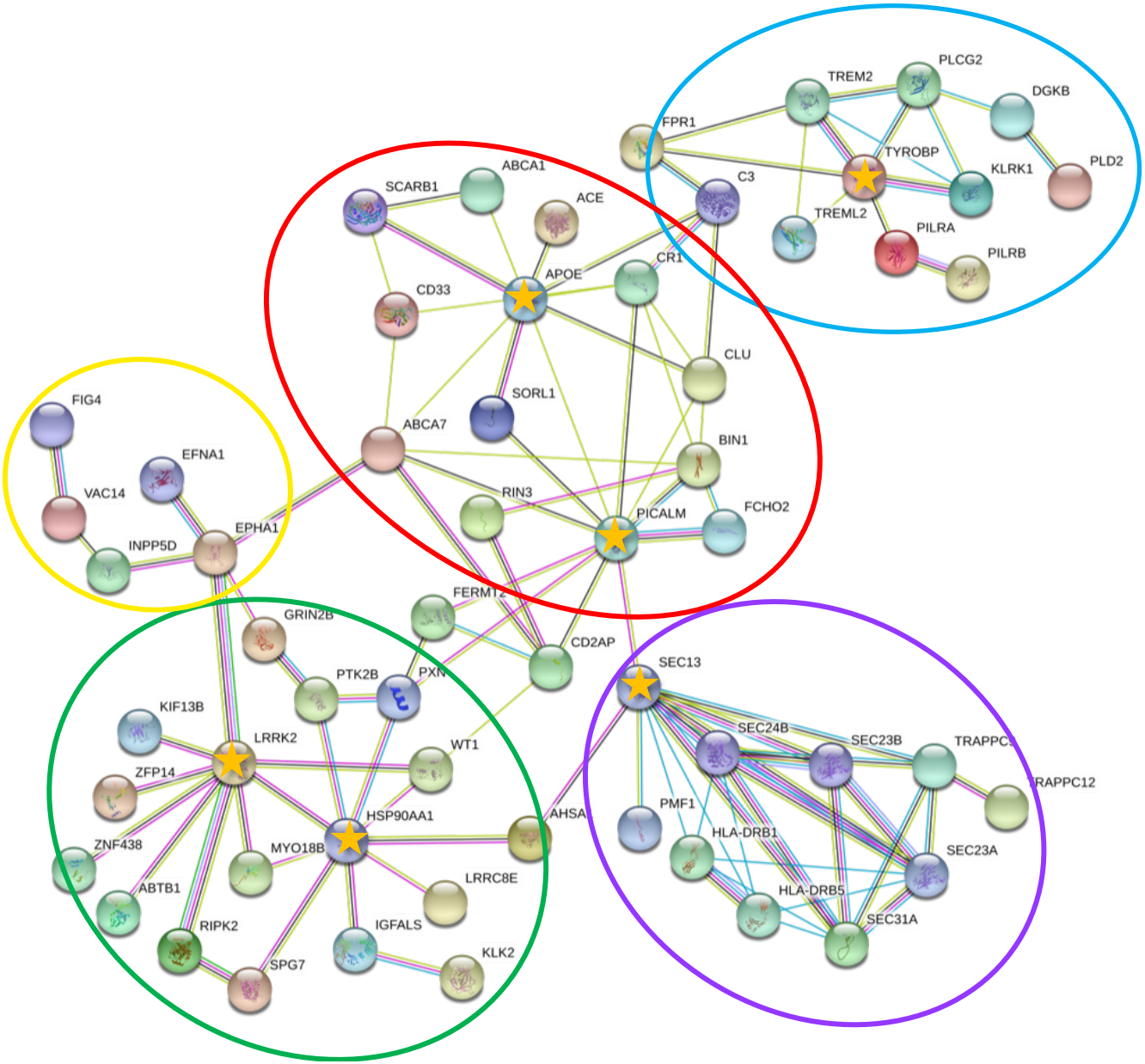


Figure 5.1. Predicted protein-protein interactions using the gene candidates of interest in STRING. Colours of edges refer to the type of evidence linking the corresponding proteins: red, gene fusion; dark blue, co-occurrence; black, co-expression; magenta, experiments; cyan, curated databases; light green, text mining; mauve, homology. There appear to be possible clusters of genes with similar biological functions, highlighted by the coloured circles. Blue, immune; red, cholesterol and lipid transport; yellow, endocytosis; green, tau microtubules and unknown functions; purple, immune signalling. Hub genes within the network clusters are denoted by a yellow star. The *TREM2-TYROBP* network incorporates some known immune-related genes. *PICALM*, *BIN1*, *SORL1*, *ABCA7*, *CLU* and *APOE* function in endocytosis and lipid transport within the cell. There also appears to be a cluster around PD-related gene *LRRK2* with genes of lesser known functions. Unconnected nodes were not shown here.

facilitate the uptake of cholesteryl esters from high-density lipoproteins in the liver. *ACE* has previously been implicated in APP and Aβ processing, here it was found to co-express with *APOE*.

The network of genes containing *TREM2* included newly published risk gene *PLCG2* and *KLRK1*, which was nominally associated through single variant analysis. *PILRA* interacted with the *TREM2* network via adaptor protein TYROBP through co-expression and previous literature which was exploring the role of *ZCWPW1* in AD. The interaction between *PILRA* and *PILRB* has been experimentally determined as they are paired receptors. Many of the genes in this cluster have been linked to the inflammatory pathway.

PD risk gene *LRRK2* forms the core of another cluster, with many of the interactions being strongly supported by experimental evidence. A convincing link with AD gene *EPHA1* was observed, which could suggest a potential mechanism of effect in AD. Variants in *KIF13B*, *MYO18B* and *WT1* were associated with AD and appear to be linked via *LRRK2*. *KIF13B* and *MYO18B* are predicted to play a role intracellular transport of molecules, whereas *WT1* is a transcription factor believed to be involved in normal development of the urogenital system. *GRIN2B* encodes an NMDA receptor and interacts with both *EPHA1* and *PTK2B*, possibly via the endocytic pathway. *VAC14* is co-expressed with *INPP5D*, which also directly interacts with *EPHA1*, likely through the same pathway.

The final cluster of genes contains several protein transport genes which are involved in vesicle trafficking and biogenesis from the endoplasmic reticulum, namely *SEC13*, *SEC23A*, *SEC23B*, *SEC24B* and *SEC31A*. *TRAPPC9* and *TRAPPC12* share a similar role to the *SEC* genes in vesicular transport from the endoplasmic reticulum to the Golgi. This network of genes links the GWAS genes *HLA-DRB1* and *HLA-DRB5* with *TRAPPC9* and *TRAPPC12*, which contained variants nominally associated with AD. However, the HLA genes are involved in

signalling in the immune system, therefore the functional link within this cluster is less clear.

5.4.2. Screening of familial AD genes

Familial AD genes, *APP, PSEN1* and *PSEN2* were screened for coding mutations to identify potentially pathogenic variants linked to EOAD. A total of 13 coding variants were identified, with 1 SNP present in *APP*, 2 in *PSEN1* and 10 in *PSEN2*, documented in Table 5.2. There were 8 samples that harboured a singleton variant in either the *PSEN1* or *PSEN2* gene. In addition, all AD individuals harbouring variants in the familial genes were heterozygous and had disease onset after the age of 65, indicating that these samples are representative of LOAD.

The majority of mutations were very rare, with all missense variants displaying MAFs less than 1.5%. Synonymous variants exhibited no direct functional effects on the protein sequence and had not previously been associated with AD. Six of the missense variants were present in the AlzGene database, with only rs200598249 (p.N24T/S) not documented. The rs17125721 SNP (p.E318G) was present in 1 control sample and 5 AD cases (mean age of death = 84.2), indicating that this mutation is not a definitive marker of fEOAD. This substitution mutation is predicted to modify disease risk with increased risk of AD observed in *APOE* ε4 carriers (Benitez *et al.,* 2013). This was also observed here as the control individual had a ε3ε3 genotype, whereas all case samples possessed at least 1 ε4 allele. There were 5 samples with other neurodegenerative disorders carrying the mutation, including LBD, VaD, PD and FTLD; only the individual with PD was ε4 positive.

The variant rs150400387 results in a p.R62C or p.R62G substitution depending on the base mutation. Pathogenicity for an arginine to cysteine change is unclear as *in silico* predictions indicate the mutation is probably damaging, whereas p.R62G is not pathogenic. However, it was only present in a control sample and therefore likely to be benign.

Table 5.2. *In silico* functional predictions of the coding variants found in *APP*, *PSEN1* and *PSEN2*. There were no known causative variants identified in the BDR cohort of LOAD samples. Although a mutation in *PSEN2* (rs150400387) was predicted to be probably damaging, it was only observed in a single case therefore it was not considered a pathogenic mutation in this study. Some variants had unclear pathogenicity due to conflicting information about their prevalence in case and control samples. As a result, it was concluded that no pathogenic FAD mutations were identified in this cohort.

Genes	SNP	Position	Ref>Alt	MAF (Count)		ExAC MAF	Protein Change	Consequence	Pathogenicity
				Case	Control				
APP	rs148888161	21:27264121	G>A	0.003 (1)	0.009 (1)	0.002		Synonymous	No functional change
PSEN1	rs200598249	14:73614798	A>G	0.003 (1)	0 (0)	0.000009	p.N24T/S	Missense	Benign
	rs17125721	14:73673178	A>G	0.0169 (5)	0.009 (1)	0.0153	p.E318G	Missense	Risk modifier
PSEN2	rs11405	1:227069677	C>T	0.186 (55)	0.2 (22)	0.2397		Synonymous	No functional change
	rs150400387	1:227071448	C>T	0 (0)	0.009 (1)	0.0002	p.R62C/H	Missense	Probably damaging
	rs140501902	1:227071475	C>T	0.003 (1)	0 (0)	0.0034	p.R71W	Missense	Unclear
	rs1046240	1:227071525	T>C	0.432 (128)	0.409 (45)	0.4991		Synonymous	No functional change
	rs200610057	1:227071600	C>T	0.003 (1)	0 (0)	0.0003		Synonymous	No functional change
	rs63750197	1:227073271	C>T	0.003 (1)	0.009 (1)	0.0006	p.S130L	Missense	Unclear
	rs61757781	1:227075813	A>G	0 (0)	0.009 (1)	0.0006	p.M174V	Missense	Not pathogenic
	rs144629723	1:227076551	T>C	0 (0)	0.009 (1)	0.00007		Synonymous	No functional change
	rs61730652	1:227076671	T>C	0.003 (1)	0 (0)	0.013		Synonymous	No functional change
	rs63750110	1:227083249	A>C	0 (0)	0.009 (1)	0.00004	p.D439A	Missense	Unclear

Conversely, rs140501902 (p.R71W) was found in a LOAD individual with an age at death of 91. It is predicted to be damaging *in silico* and has also been detected in familial cases of EOAD and LOAD. Although it has been observed in a handful of AD samples, healthy controls have also harboured the variant, indicating that it is not pathogenic but could modify disease risk. The SNP rs63750197 (p.S130L) has been reported in EOAD, LOAD and control samples, therefore its pathogenicity is unclear. In this dataset, it occurs in one case and control sample and is likely to be benign as a result. The variants rs61757781 (p.M174V) and rs63750110 (p.D439A) have been found to be benign as they did not segregate with disease in either early or late-onset patients. In this study, they were observed as singleton variants in healthy control samples with no AD pathology.

Although some predicted pathogenic mutations were observed, all case samples were confirmed to be LOAD due to the age of onset. No known pathogenic mutations were identified in AD samples; however, a number of the SNPs were detected in both case and control samples. This could be indicative of potential risk modifying variants or a potential role in delaying the onset of AD.

5.4.3. Screening of other neurodegenerative disease genes

Genes associated with other familial neurodegenerative diseases were also screened for potential pathogenic mutations to identify genetic overlap with sporadic AD. A total of 105 variants were extracted from 13 of the 15 familial genes, as no coding variants were present in *CHMP2B* and *SNCA*. These were filtered to retain only protein coding variants, which left 39 variants in 10 genes, listed in Table 5.3. The AD, FTD and PD mutation databases were used to catalogue mutations and define their pathogenicity.

The majority of mutations in PD-related genes were classified as non-pathogenic; unclear pathogenicity was assigned if variants could not be

Table 5.3. Coding mutations identified in genes implicated in Mendelian forms of neurodegeneration. Synonymous variants were excluded to retain only functional mutations. The predicted pathogenicity was determined by functional predictions from SIFT and Polyphen2 and clinical significance information from mutation databases.

Disease	Gene	SNP	Position	Ref/Alt	ExAC MAF	MAF (Count)		Protein change	Consequence	Functional Prediction
						Case	Control			
Parkinson's disease	*GBA*	rs76763715	1:155205634	T>C	0.002	0 (0)	0.009 (1)	N409S	Missense	Benign
		rs2230288	1:155206167	C>T	0.010	0.020 (6)	0.009 (1)	E365K	Missense	Benign
	LRRK2	rs33995463	12:40629436	T>C	0.001	0.003 (1)	0 (0)	L119P	Missense	Benign
		rs10878245	12:40631791	C>T	0.474	0.412 (122)	0.389 (42)	L153M	Missense	Benign
		rs56108242	12:40634396	G>C	0.0001	0 (0)	0.009 (1)	C228S	Missense	Benign
		rs7308720	12:40657700	C>G	0.086	0.082 (24)	0.046 (5)	N551K	Missense	Benign
		rs10878307	12:40671989	A>G	0.059	0.071 (21)	0.055 (6)	I723V	Missense	Benign
		rs35173587	12:40677813	G>T	0.001	0.003 (1)	0 (0)	R793M	Missense	Probably damaging
		rs72546338	12:40702283	G>A	0.0002	0.003 (1)	0 (0)	R1325Q	Missense	Probably damaging
		rs17466213	12:40702420	A>G	0.001	0.003 (1)	0 (0)	I1371V	Missense	Unclear
		rs7133914	12:40702911	G>A	0.084	0.081 (24)	0.046 (5)	R1398H	Missense	Benign
		rs35507033	12:40707778	G>A	0.004	0.014 (4)	0.018 (2)	R1514Q	Missense	Benign
		rs33958906	12:40707861	C>T	0.030	0.024 (7)	0.018 (2)	P1542S	Missense	Benign
		rs35303786	12:40713899	T>C	0.009	0.007 (2)	0.046 (5)	M1646T	Missense	Benign
		rs11564148	12:40713901	T>A	0.298	0.276 (81)	0.291 (32)	S1647T	Missense	Benign
		rs200143418	12:40713905	A>G	0.0001	0.003 (1)	0 (0)	Q1648R	Missense	Probably damaging
		rs3761863	12:40758652	C>T	0.376	0.331 (98)	0.391 (43)	M2397T	Missense	Benign
	MAPT	rs143624519	17:44068850	G>A	0.001	0.011 (3)	0 (0)	A469T	Missense	Benign
	PARK2	rs149953814	6:161771219	G>A	0.002	0.003 (1)	0.009 (1)	P437L	Missense	Unclear
		rs1801334	6:161781225	C>T	0.026	0.057 (17)	0.018 (2)	D394N	Missense	Benign
		rs1801582	6:161807855	C>G	0.165	0.179 (53)	0.2 (22)	V380L	Missense	Benign

		rs34424986	6:162206852	G>A	0.002	0.003 (1)	0.009 (1)	R275W	Missense	Possibly damaging
		rs55774500	6:162683724	G>T	0.005	0.003 (1)	0.009 (1)	A82V	Missense	Unclear
		rs746646126	6:162683747	C>CCA	0.00003	0.003 (1)	0 (0)	W74C/X	Frameshift	Pathogenic
		rs55777503	6:162864410	CCT>C	0.0003	0.003 (1)	0 (0)	Q34X	Frameshift	Pathogenic
	PARK7	rs71653619	1:8030994	G>A	0.008	0.003 (1)	0 (0)	R98Q	Missense	Benign
	PINK1	rs148871409	1:20960385	A>T	0.064	0.054 (16)	0.046 (5)	Q115P	Missense	Benign
		rs781425640	1:20964413	T>C	0.00001	0 (0)	0.009 (1)	Y156H	Missense	Probably damaging
		rs3738136	1:20972111	G>A	0.092	0.051 (15)	0.027 (3)	A340T	Missense	Benign
		rs777895210	1:20975488	G>C	0.00002	0 (0)	0.009 (1)	V418M	Missense	Probably damaging
		rs1043424	1:20977000	A>C	0.297	0.318 (94)	0.255 (28)	N521T	Missense	Benign
FTD	*C9orf72*	rs17769294	9:27561628	T>C	0.104	0.139 (41)	0.118 (13)	N207S	Missense	Benign
		rs201250307	9:27566973	G>C	0.0002	0 (0)	0.009 (1)	T49R	Missense	Benign
	FUS	rs778725612	16:31196420	0>CGGTGGT	0.0001	0.003 (1)	0 (0)	-/GGG228-229	In-frame insertion	
	GRN	rs63750742	17:42426631	C>T	0.003	0.007 (2)	0 (0)	D33E	Missense	Unclear
		rs63750412	17:42429500	C>T	0.004	0.003 (1)	0 (0)	R433W	Missense	Benign
		rs373138049	17:42430120	G>A	0.0001	0.003 (1)	0 (0)	R579H	Missense	Benign
Prion disease	*PRNP*	rs138688873	20:4680094	CCATGG>C	0.019	0.003 (1)	0.009 (1)	HGGGWGQP/-77-84	In-frame deletion	
		rs1799990	20:4680251	A>G	0.308	0.348 (103)	0.382 (42)	M129V	Missense	Benign, pathogenic allele

segregated by disease in multiple studies. Two frameshift mutations, rs746646126 and rs55777503, were identified in *PARK2* exons 2 and 3, both observed in a case sample. They were predicted to be pathogenic in the PD mutation database and by SIFT and PolyPhen2 scoring, however they have not been directly linked to PD. The SNPs rs781425640 and rs777895210 in *PINK1* were not documented; they were only present in cognitively normal control samples. *In silico* functional prediction with SIFT and PolyPhen2 affirmed that both mutations were probably damaging or deleterious. *LRRK2* contained several singleton variants that were determined to be non-pathogenic. There was an undocumented variant, rs200142418, present in a single AD sample and predicted to be deleterious. A further three SNPs, rs35173587, rs72546338 and rs17466213, followed the same pattern in different AD samples, yet had unclear pathogenicity. *MAPT* variant rs143624519 was present in 2 AD cases and no controls, with one of the samples being homozygous for the mutation. It has been reported to increase susceptibility to neurodegenerative conditions by over 2-fold, including AD.

An in-frame insertion in *FUS* was only present in an AD sample; rs778725612 has not been previously investigated in FTD or AD. *PRNP* common SNP rs1799990 has previously been implicated as a risk factor for prion disease; however, it does not appear to segregate with disease in AD. A rare in-frame deletion in the *PRNP* gene was present in one case and a healthy control individual. Other studies have also identified this variant in both AD cases and controls, suggesting that it is potentially a risk factor.

5.4.4. Single variant association in candidate genes

Single variant analysis was performed on the coding variants in candidate genes using a logistic regression test in PLINK. This was corrected for sex, age and number of *APOE* ε2 and ε4 alleles. Fewer multiple comparisons could increase the power to detect associations

that were previously below the threshold. Table 5.4 shows the most associated SNPs at the nominal cut-off P<0.05. Using a Bonferroni-adjusted threshold of P=0.0001, there were no significantly associated SNPs following correction for multiple testing.

Several variants in *ABCA7* were nominally associated at P<0.005, further supporting the link with AD identified by GWAS. Three of the 4 variants were identified in the exome-wide association analysis, which are rs3764645, rs3752237 and rs4147915. All 4 variants had a common MAF and only rs3764645 was not a silent mutation. This variant causes a p.E188G substitution and has previously been implicated in AD. Two of the SNPs exerted a two-fold increase in AD risk, whereas rs3764645 and rs4147915 appeared to be protective with effect sizes below 0.5.

Interestingly, rare missense variant rs35303786 (p.M1646T) in PD risk gene *LRRK2* also showed tentative association with AD (P=0.019). This was the only non-AD gene to appear in this list and occurred more frequently in controls. Functional predictions indicate that the mutation is tolerated and likely to be benign in PD, however it has not been investigated in AD. The synonymous variants in *PTK2B* and *SLC24A4* were only just below the P<0.05 cut-off. They were observed more frequently in cases, indicating an increased risk of AD. Although synonymous variants do not affect the amino acid sequence, they can have an impact on splicing and thus mRNA levels of the gene as a result (Sauna and Kimchi-Sarfaty, 2011).

5.4.5. Burden analysis

The SKAT-O combined test was implemented to increase statistical power for calculating gene-based burden. The joint test was performed to determine the combined effects of common and rare variants; it was also conducted for rare and common variants separately to identify their burden. The results are shown in Table 5.5, with the number of SNP markers tested within each group. The SNPs were weighted according to their MAF, with low frequency variants assigned a greater weight

Table 5.4. Logistic regression analysis of known risk genes in AD and other neurodegenerative diseases. The covariates sex, *APOE* status and age were adjusted for. Results are only shown for variants associated below P<0.05. The Bonferroni-corrected P-value was 0.0002, however no variants reached this significance threshold, likely due to insufficient power. The three *ABCA7* variants that were associated with altered AD risk in exome-wide analysis appear here, as does an additional variant.

Gene	Variant	Position	Allele	ExAC MAF	MAF (Counts)		Odds Ratio (95% CI)	P-value	Consequence	Protein Change	Functional prediction
					Cases	Controls					
PTK2B	rs1030526	8:27277537	G>A	0.40	0.44 (129)	0.38 (42)	1.70 (1.02-2.82)	0.0415	Synonymous	T110	No functional change
LRRK2	rs35303786	12:40713899	T>C	0.009	0.007 (2)	0.05 (5)	0.12 (0.02-0.71)	0.0194	Missense	M1646T	Tolerated
SLC24A4	rs7144273	14:92920371	T>C	0.50	0.51 (150)	0.39 (43)	1.69 (1.03-2.77)	0.0389	Synonymous	P319	No functional change
ABCA7	rs3764645	19:1042809	A>G	0.48	0.42 (123)	0.60 (66)	0.41 (0.24-0.70)	0.0009	Missense	E188G	Tolerated
ABCA7	rs3752234	19:1047002	G>A	0.55	0.51 (151)	0.39 (43)	2.11 (1.26-3.55)	0.0047	Synonymous	A608	No functional change
ABCA7	rs3752237	19:1047161	G>A	0.68	0.46 (136)	0.29 (32)	2.44 (1.44-4.14)	0.0009	Synonymous	G617	No functional change
ABCA7	rs4147915	19:1049305	C>A	0.18	0.11 (31)	0.20 (22)	0.31 (0.15-0.63)	0.0015	Synonymous	V807	No functional change

compared to common variants. For some genes only 1 or no SNPs remained for testing after quality control. This prevented burden testing for the genes *APP*, *FUS*, *VCP* and *MEF2C* as more than one marker is required for calculating a burden effect. There were no genes that exhibited significant burden on AD after adjustment for multiple testing, Bonferroni-corrected P=0.001.

Tentative associations were observed in several genes, when collectively testing either rare or common variants. Common variants in *ABCA7* and *PTK2B* demonstrated a burden effect, P=0.04 and P=0.05, respectively. On the contrary, rare variants in *CLU* and PD risk gene *LRRK2* exerted some burden, P=0.03 and P=0.07, respectively. However, when variants were combined, this effect was not observed for *PTK2B*, *CLU* and *LRRK2*; altogether this suggests that the other variants are not likely to be contributing to this signal. There are 8 rare variants in *CLU* and 22 in *LRRK2* that could be contributing to the signals and therefore these variants should be examined further. For *ABCA7*, the combined test was suggestive of burden with P=0.07, yet rare SNPs in the gene do not associate with AD (P=0.2).

5.4.6. *TREM2* R47H mutation

Screening of the *TREM2* gene in the WES data revealed that three samples were carrying the rs75932628 SNP, which has been linked to an increased risk of AD. All three samples heterozygous for rs75932628 were AD cases (BN037, BM007 and BO100). This mutation is extremely rare yet the MAF in the BDR dataset was elevated at 0.5%. These samples were confirmed using Sanger sequencing to verify the WES data. Chromatograms are shown in Figure 5.2 for the 3 heterozygous samples and a wild type sequenced as a control. The C/T mutation was observed in all 3 samples as overlapping peaks, confirming the sequence data and demonstrating that this was indeed a true finding. The control sample BK081 has a C-allele, which validates that it is homozygous wild type, as seen in the WES. The intensity of the sequencing signal for sample BO100 was considerably lower than

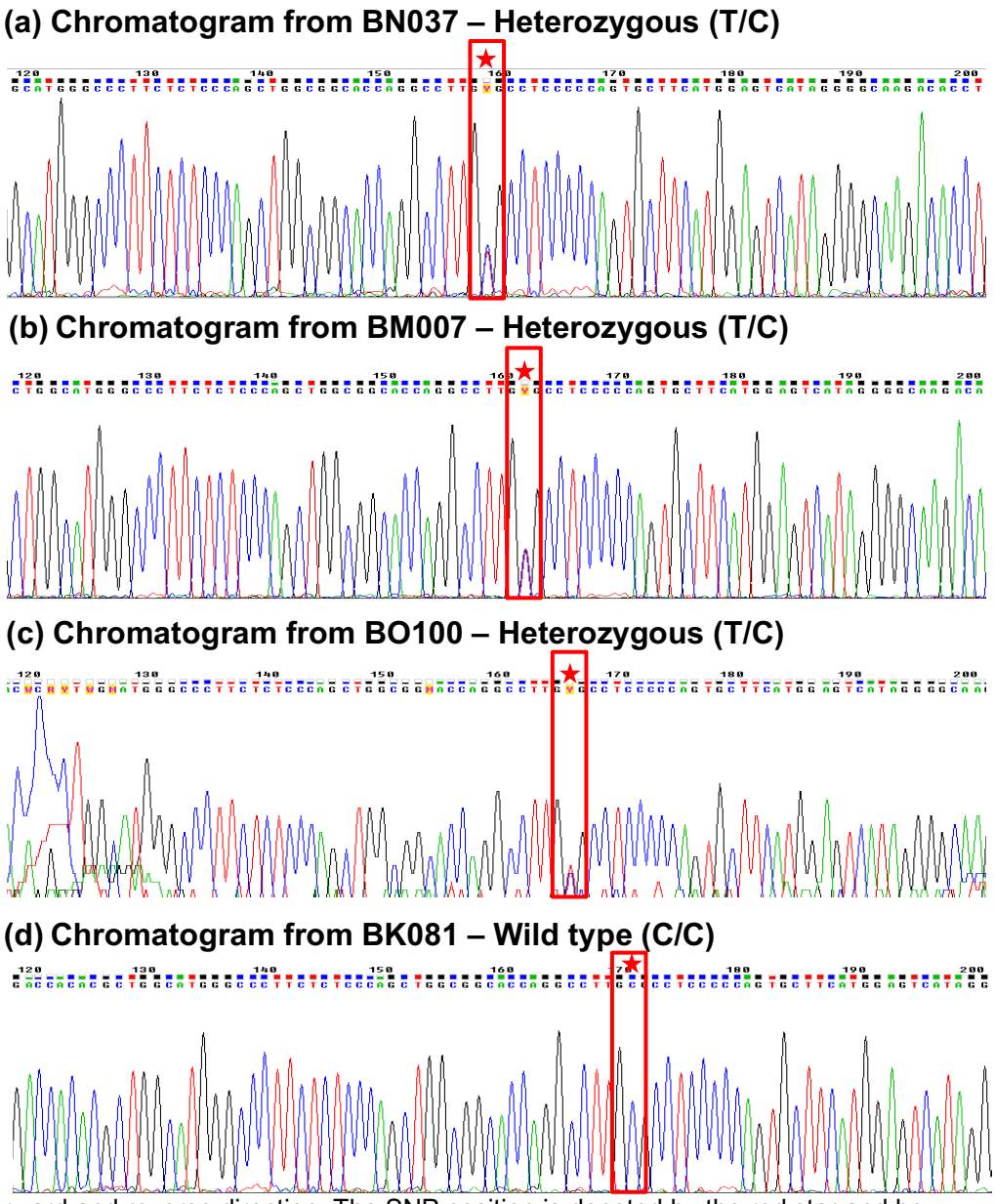
Table 5.5. Burden association for all selected risk genes containing a minimum of 2 variants. SKAT-O was used to perform burden analysis jointly and individually on rare and common variants to identify which contribute greater to risk of developing AD. *CLU* contained 8 rare variants that were collectively associated with AD risk P<0.05. However, when the common variant was included, the association became non-significant.

Disease	Gene	Joint Testing		Rare Only		Common Only	
		Markers Tested (Total)	P-value	Rare variants	P-value	Common variants	P-value
Familial AD	*APP*	1 (1)	0.1	1	0.1	0	0
	PSEN1	2 (2)	1	2	1	0	0
	PSEN2	11 (11)	0.42	9	0.2	2	1
FTD	*C9orf72*	6 (6)	0.57	4	0.57	2	0.36
	CHMP2B	4 (6)	0.75	2	0.58	2	0.51
	FUS	2 (3)	0.4	1	0.15	1	0.8
	GRN	8 (8)	0.75	8	0.75	0	0
Parkinson's disease	*GBA*	3 (3)	0.51	3	0.51	0	0
	LRRK2	35 (46)	0.46	22	0.07	13	0.27
	PARK2	9 (9)	0.49	7	0.36	2	0.6
	PARK7	2 (2)	0.63	2	0.63	0	0
	PINK1	7 (11)	0.53	3	0.31	4	0.59
	SNCA	2 (2)	0.34	2	0.34	0	0
Prion disease	*PRNP*	4 (4)	0.56	3	0.37	1	0.67
ALS	*VCP*	1 (1)	0.35	1	0.35	0	0

Late-onset Alzheimer's disease	*ABCA7*	43 (49)	0.07	21	0.2	22	0.04
	APOE	6 (6)	0.38	3	0.18	3	0.28
	APP	1 (1)	0.1	1	0.1	0	0
	BIN1	6 (7)	0.14	5	0.12	1	0.36
	CASS4	16 (16)	0.93	10	0.93	6	0.71
	CD2AP	8 (8)	0.25	8	0.25	0	0
	CD33	5 (5)	0.27	2	0.22	3	0.17
	CELF1	3 (3)	0.83	3	0.83	0	0
	CLU	9 (10)	0.17	8	0.03	1	0.34
	CR1	16 (17)	0.38	11	0.3	5	0.22
	EPHA1	16 (19)	0.83	12	0.75	4	0.57
	FERMT2	7 (9)	0.56	6	0.41	1	0.36
	INPP5D	3 (3)	0.28	2	0.85	1	0.16
	MEF2C	0 (3)	1	0	1	0	1
	MS4A6A	6 (8)	0.31	4	0.21	2	0.75
	NME8	15 (15)	0.56	8	0.38	7	1
	PICALM	4 (6)	0.65	2	0.6	2	0.41
	PSEN1	2 (2)	1	2	1	0	0
	PSEN2	11 (11)	0.42	9	0.2	2	1
	PTK2B	21 (22)	0.11	11	0.41	10	0.05
	RIN3	19 (25)	0.46	12	0.77	7	0.28
	SLC24A4	9 (10)	0.31	4	0.19	5	0.23
	SORL1	25 (35)	0.65	19	0.18	6	0.49
	TREM2	9 (9)	0.42	9	0.42	0	0
	ZCWPW1	9 (9)	0.87	7	0.68	2	1

the other samples, however the sequence calling software was able to make the prediction indicated.

Figure 5.2. Chromatograms for samples sequenced for the *TREM2* R47H variant. The three heterozygous samples were confirmed using Sanger sequencing in the forward and reverse direction. The SNP position is denoted by the red star and box.



5.5 Discussion

A number of candidate genes have now been implicated in AD pathogenesis through GWAS and NGS studies. Moving away from data-driven approaches, exploration of rarer variants within these risk genes could be the key to finding causal variants with greater effect sizes. However, both methods will face drawbacks when modest sample sizes are used or if the effects are small (Amos *et al.,* 2011). Candidate gene studies are based on prior knowledge of risk genes; therefore, identification of putative causal variants is limited by the genes selected. Significance thresholds for such studies are less stringent due to fewer multiple comparisons (Patnala *et al.,* 2013).

5.5.1. Pathway analysis

Pathway analysis predicted some interesting connections between known AD risk genes and the new genes of interest we highlighted from exome-wide association and burden testing. Figure 5.1 revealed that there was experimental evidence indicating that many of the GWAS genes may physically or functionally interact. Although there were a large number of connections based on literature and curated database information alone, there was also evidence from co-expression and experimental studies. This was particularly true in the *LRRK2* centred gene network cluster. However, *in silico* pathway analyses can only be as accurate and informative as the databases used to generate the predictions (Khatri *et al.,* 2012).

Gene expression networks are of particular biological interest since co-expressed genes could be controlled by the same transcriptional regulators, functionally related, or involved in the same pathway or protein complex (Weirauch, 2011). As a number of the genes put forward from the analyses in Chapter 4 were predicted to interact with known risk genes, we can assert more confidence that they could play a role in AD and warrant further investigation. Clusters of genetic interactions were evident from the network, with many of the genes

within clusters sharing similar biological functions. Nevertheless, there is a significant crossover in the functions of existing AD genes (Smith *et al.,* 2016), as observed in the pathways implicated by GWAS in Figure 1.6.

LRRK2 forms a cluster of many direct connections with experimental evidence supporting the interaction. Mutations in this PD-related gene have been associated with AD-like pathology, suggesting a potential overlap in the underlying neurodegenerative pathways in both AD and PD (Zimprich *et al.,* 2004). Although *LRRK2* mutations are not common in AD, we have uncovered some tentative evidence through polygenic risk contribution and interactions with other AD risk genes. This highlights the potential pleiotropic effects of this gene which warrants further investigation in AD.

The biological functions of many of the genes remains unclear. However, using predictions from pathway analyses we can infer the potential relevance to disease. This can help form hypotheses to experimentally investigate the role of new genes in AD. This analysis also provides further support for the role of *PILRA* in the immune pathway via interactions with *TREM2* and *TYROBP*. However, the databases used to generate such predictions require more information before we can assign greater confidence to the interactions identified.

We also potentially identified new classes of genes involved in the endocytic and vesicular trafficking pathway. Multiple genes in the *SEC* and *TRAPPC* gene families were involved in transporting vesicles from the ER to the Golgi. *SEC31A*, *TRAPPC9* and *TRAPPC12* were nominally associated with AD in single variant analysis. Although there is no known physical interaction with HLA genes, there is sufficient evidence to suggest functional links between the SEC and TRAPPC gene groups.

STRING has highlighted some interesting connections, however due to the nature of *in silico* analyses, these interactions would need to be

verified experimentally before conclusions can be made (Hu *et al.,* 2017). It has been suggested that using data from AD-specific curated databases could provide valuable information for exploring mechanisms relating specifically to AD (Hu *et al.,* 2017).

5.5.2. Screening familial AD genes for known causative mutations

Causal mutations in the familial AD genes *APP*, *PSEN1* and *PSEN2* are extremely rare but highly penetrant. Harbouring a pathogenic mutation in one of these genes almost guarantees that an individual will develop the autosomal dominant form of AD. *PSEN1* mutations are the most common cause of familial EOAD, whereas *APP* and *PSEN2* mutations are rarer. Samples were screened for potentially pathogenic variants in these genes to ensure they were representative of sporadic LOAD (Table 5.2). Screening revealed a greater number of mutations in *PSEN2* rather than *PSEN1*, which is known to harbour the most causative variants. Singleton variants were previously excluded due to possible sequencing artefacts as a result of only being observed in one sample. However, due to the low frequency of these variants, singletons were also checked for familial mutations. The majority of SNPs identified were silent mutations and therefore had no direct impact on the protein sequence. However, these synonymous variants have been implicated in mRNA splicing and secondary structure, which can affect its stability (Sauna and Kimchi-Sarfaty, 2011).

The rs17125721 variant (p.E318G) in *PSEN1* was identified in 5 AD samples and 1 control. The variant did not segregate with disease in several studies (Hippen *et al.,* 2016; Lindquist *et al.,* 2009); however it was associated with a 2-fold greater risk of developing LOAD for *APOE* ε4 carriers harbouring the mutation than individuals without the variant (Benitez *et al.,* 2013; Nho *et al.,* 2016). The pattern was also observed in this dataset as the 5 AD individuals had either 1 or 2 ε4 alleles, whereas the control sample had the ε3ε3 genotype. Benitez and colleagues demonstrated that the effect of this variant is dependent on an interaction with *APOE* ε4; in the presence of at least one ε4 allele

p.E318G was associated with increased Aß deposition and faster cognitive decline.

All AD individuals had an age at death > 70 years with no clinical diagnosis of EOAD; therefore, the samples were correctly classified as LOAD. This also indicates that these variants are likely to be risk factors for LOAD rather than causative mutations. Another explanation is that some individuals could harbour protective variants that can delay the onset of AD.

5.5.3. Screening genes linked to other neurodegenerative diseases for familial mutations

Genes associated with familial forms of other neurodegenerative disorders can also provide insight into the biological mechanisms underlying neurodegeneration. There is considerable overlap in both clinical and neuropathological features shared by dementias, which commonly lead to misdiagnosis. Genetic pleiotropy has also been shown to play a role in the genetic basis of these diseases, whereby one gene can influence multiple traits. Screening for previously documented familial mutations within the genes linked to Mendelian forms of diseases like PD and FTD revealed no known causative mutations. However, there were some variants present only in AD cases within PD genes *PARK2* and *LRRK2* with unclear pathogenicity or predicted probably damaging that may warrant further investigation. Two heterozygous frameshift mutations were identified in only AD cases in *PARK2*, predicted to impact protein structure and is therefore likely to be pathogenic according to the ClinVar database. Both rs746646126 and rs55777503 variants result in a premature translation termination codon, causing truncation of the protein. In these samples they are heterozygous, although the homozygous variants cause loss-of-function mutations in the *PARK2* gene, associated with early-onset PD (Abbas *et al.,* 1999). The variants were present in LOAD cases with age of death over 80 and no PD pathology, however they exhibited some TDP-43 pathology and had high numbers of NFTs.

Known pathogenic *LRRK2* mutations are not common in clinical AD patients (Toft *et al.,* 2005). Two variants rs33995463 and rs200143418 were predicted to be damaging and were only identified in a single case sample each. The former variant was indicated as non-pathogenic in the PD mutation database, but the latter had not been documented.

An in-frame insertion in the FUS gene was noted in an AD sample but no controls. This rs778725612 variant is not documented and did not have a functional prediction in VEP. The AD patient carrying this mutation also had a comorbid case of Lewy body disease and poorer cognitive function. FUS mutations are commonly linked to ALS and FTLD, which show little to no overlap with DLB pathology.

Although we have not identified any known pathogenic familial mutations from other neurodegenerative diseases in this sample of AD patients, this does not exclude a possible role for these genes in AD. Such findings require further investigation by Sanger sequencing to verify the genotypes. It would also be beneficial to validate these results in a larger sample, either when the BDR collection increases in size or using an independent cohort.

5.5.4. Association analysis

The exome-wide analyses performed in Section 4 were underpowered to detect associations of magnitude greater than 2 for low frequency variants. Gene-based approaches are used to investigate subsets of variants in genes of interest as they can provide increased statistical power due to smaller corrections for multiple testing (Amos *et al.,* 2011). A logistic regression test was used with correction for covariates to identify associations independent of age, sex and *APOE* ε status of individuals.

Multiple common variants in *ABCA7* were highly associated with AD, despite non-significance following multiple correction. All were silent mutations apart from missense variant rs3764645, which was associated with protection from AD. This previously showed nominal

protection against AD in a Belgian cohort (Cuyvers *et al.,* 2015). This finding was also replicated in a study of EOAD, where it was in strong LD with rs72973581 (D'=0.96), another missense variant in *ABCA7* (De Roeck *et al.,* 2017). This missense variant was predicted to be damaging by PolyPhen PSIC score (Nuchnoi *et al.,* 2011). However, in this case both SIFT and PolyPhen2 functional prediction tools indicated that it was tolerated, also supported by the aforementioned studies. Sassi and colleagues found rs3752234 to be highly associated with protection against LOAD, however the opposite was observed in this dataset, although the association was not significant (Sassi *et al.,* 2016). Closer examination of the minor allele counts indicates that the ExAC database MAF is more congruent with that of AD samples rather than the control population. This observation is true of several variants in Table 5.4. Erroneous results like this could be due to a smaller sample size; additionally the number of controls is significantly less than cases, which could skew the results (Zuk *et al.,* 2014). Synonymous variant rs4147915 was not associated with AD and rs3752237 had not been genotyped in any previous studies. The results appear to support rs3764645 as a possible protective variant against AD, however its functional impact on the protein requires further verification.

The rare variant rs35303786 in *LRRK2* (p. M1646T) was associated with decreased risk of developing AD and predicted to be tolerated. This variant is not considered a known pathogenic mutation, although it has been associated with an increased risk of PD in Caucasian individuals (Ross *et al.,* 2011). Both tools SIFT and Polyphen and the PD mutation database predict the substitution to be a tolerated mutation. However, *in vitro* cell survival assays showed significant reduction in survival in cells harbouring the mutant allele (Refai *et al.,* 2015). This suggests that carrying the SNP is likely to increase risk of developing PD. *LRRK2* mutations have not been extensively documented in AD, although they were shown to correlate with Lewy body pathology in Lewy body variant AD (Linnertz *et al.,* 2014).

Synonymous variants in *PTK2B* and *SLC24A4* were tentatively associated with increased risk of AD, though not as significantly as the other variants. It is possible that an increased sample size could push some of these associations into significance, or to the contrary, cause them to fall below the threshold if they are spurious results. There are a large number of synonymous mutations presenting with high associations with AD and relatively large effect sizes. Although silent, it has been well documented that silent mutations can potentially alter splicing and mRNA levels to indirectly affect protein functionality. By elucidating the mechanisms by which synonymous mutations work it may be possible to assess if synonymous mutations play any role in susceptibility to AD.

Imputation of variants within the candidate genes could significantly improve the number of markers available for testing. This uses known haplotypes to infer genotypes for individuals in the given dataset from large curated reference panels like 1000 Genomes (1000 Genomes Project Consortium, 2015). GWAS has significant coverage of non-coding variants, whereas WES is able to encompass more coding variants. As a result, neither can directly genotype all genomic variation without using WGS. Imputation of the genotypes can boost the coverage of genomic variation, allowing more variants to be assessed without additional sequencing or genotyping (Edwards *et al.,* 2013).

5.5.5. Burden analysis

GWAS investigate common variant association with disease; in contrast, individual SNP analysis in rare variant studies is extremely underpowered. A significantly larger number of rare variants and samples would be required to detect associations. Burden tests aggregate or collapse variant sites into gene or region-based sets to obtain a single association value for the set. This can also improve the power to detect larger and more cumulative effects. SKAT-O is a combined test that can achieve robust power by merging burden and variance-component tests (Lee *et al.,* 2014). Gene-based burden

analysis in SKAT-O allows correction for covariates such as sex, age and *APOE* ε status. It can also adjust for other factors such as MAF of SNPs. SKAT-O has the ability to weight variants using their effect sizes or other external information, allowing for a truer measure of genetic burden. This also accounts for effects in both risk and protective directions as not to overlook variants with smaller effects.

Burden testing was performed on rare and common variants together and individually to determine which grouping had the greater effect on AD risk. The Bonferroni threshold was too stringent to detect any significant associations beyond multiple testing correction. A tentative amount of burden was detected in *ABCA7* when rare and common variants were tested together. However, upon closer examination it appears that only the common variants were having an effect. This incorporates the 4 highly associated SNPs identified from single variant analysis and suggests that they could be driving this signal. *PTK2B* also contained common variants which exhibited tentative association with AD, however the combined effect was not significant. This could again be the result of an insufficient sample size, which can be overcome with increased sample collection. When testing rare genetic variants alone, *CLU* and *LRRK2* were the only genes exhibiting tentative burden. Fine mapping of the gene loci would allow more in-depth exploration of the potential causal variants that are behind this association (Edwards *et al.,* 2013). SNPs in *CLU* were not previously identified as risk variants in the single SNP analysis, suggesting that the effects might only be seen when they are inherited collectively. PD risk gene *LRRK2* has once again been implicated in AD risk. Since comorbid diseases are common in dementias, AD and PD can co-occur. This highlights the possibility of shared biological mechanisms in which genes can have pleiotropic effects. Although some of the genes exhibit a weak burden on AD, again these signals could potentially reach significance with a larger dataset.

5.5.6. Sanger sequencing confirmed heterozygous *TREM2* R47H mutation

Homozygous loss-of-function mutations in the *TREM2* gene have been associated with Nasu-Hakola disease; an autosomal recessive form of early-onset dementia that presents with bone cysts (Bianchin *et al.,* 2004). Heterozygous rare variants within this gene significantly predispose individuals to AD, in particular rs75932628, the R47H substitution confers a 3-fold greater risk of developing the disease (Guerreiro, Wojtas, *et al.,* 2013; Jonsson *et al.,* 2013). This mutation occurs with a frequency of 0.2% in the general population and 0.3% in non-Finnish Europeans. The mutation was observed in 3 samples from the total cohort of 288, with MAF of 0.5%. It was heterozygous in 3 AD samples and did not occur in any controls. This increased frequency could be due to the sample size consisting of 148 AD cases and 55 controls. None of the 85 samples with other phenotypes harboured the variant. Although the genotypes of these samples were confirmed by Sanger sequencing, the MAF needs to be verified in the full cohort when the sample size and detection power increases.

The same variant was also found to confer a significant risk in EOAD patients below age 65, highlighting its pleiotropic properties (Pottier *et al.,* 2013). The carriers of this mutation exhibited significant AD pathology, with BNE stages of 5 or 6 and CERAD scores of 2 and 3. Interestingly, these individuals had average PRS, indicating that this mutation did not significantly increase their genetic risk of developing AD.

5.5.7. Conclusion

We have investigated the relationship between genes involved in familial forms of aging disorders and AD, demonstrating tentative associations as a result. Some of these genetic links have been previously published, which suggests that our findings are likely to be reliable. The growing body of evidence surrounding *LRRK2* in AD is

encouraging as there is substantial overlap between AD and other neurodegenerative diseases. Only 1 *LRRK2* mutation has been previously associated with AD. Therefore, the fact that it has been identified through PRS and now single variant and burden tests indicate an avenue worth pursuing. Although the power to detect significant associations remains small, the BDR sample collection process is ongoing and eventually a larger dataset will improve this. Another resolution for identifying risk variants in the candidate genes would be imputation, which uses large curated datasets, such as 1000 Genomes, to infer genotypes for missing SNPs based on haplotype information. This would significantly improve the number of variants available within the genes and therefore would be useful to test, as it would afford greater power to detect potential associations.

Chapter 6: General Discussion

6.1 Summary of research aims and relevant results

The objective of this project was to investigate genetic variation in AD using NGS of the BDR cohort. BDR is a recently established initiative that aims to generate a well characterised resource for dementia research. The main aim was to identify genes harbouring coding variants associated with AD risk independently and in burden, both exome-wide and in known risk genes. We also sought to generate PRS and correlated these scores with clinical and neuropathological information to distinguish between high and low risk individuals. Additionally, screening for mutations in genes associated with other neurodegenerative diseases could highlight a potential role for genetic pleiotropy in AD.

WES was performed on 148 neuropathology confirmed AD samples and 55 controls, with the goal of identifying new genes of interest associated with the risk of developing AD. Sequencing data was obtained and processed for BDR samples using an optimised pipeline **(Chapter 3)**. Variants were assessed for their association with AD and PRS were also calculated to determine the genetic risk profiles for individuals **(Chapter 4)**. Using a candidate-gene approach, known AD-related and familial genes linked to other neurodegenerative diseases were also assessed for association with AD **(Chapter 5)**.

The GATK Best Practices pipeline was employed for data processing in **Chapter 3**. This is now considered the “gold standard” for variant calling in human DNA sequence data (Warden *et al.,* 2014). It is a robust pipeline used to generate high quality variant call data for downstream analyses. There are multiple sources of errors in NGS, therefore GATK comprises several processing steps through which the data was refined to effectively reduce the probability of false positives as a result (Van der Auwera *et al.,* 2013). To investigate coding variation, functional

annotation and filtering were implemented to prioritise variants of interest from the vast number of variants detected.

Exome-wide single variant association analysis in **Chapter 4** identified SNPs in several genes that were highly associated with AD risk, however no variants achieved genome-wide significance after Bonferroni correction. Tentative associations were identified in the known risk gene *ABCA7* and also *PILRA* and *NTN4*, which have previously been implicated in AD but not studied extensively. PRS could distinguish cases and controls with 84% accuracy, however this only accounts for 45% of the genetic risk in the present dataset. Comparison of neuropathological features between samples at the extremes of PRS suggested that individuals with lower risk scores exhibited more mixed pathologies. Conversely, high risk individuals were generally classified with higher Braak staging and CERAD scores. Exploration of the risk variants that contributed to the scores, highlighted multiple variants in several GWAS genes and a number of previously implicated genes.

Pathway analysis predicted that several of the genes highlighted by exome-wide analysis had potential interactions with known AD risk genes. Some were also suggested to be involved in existing biological pathways linked to AD, such as inflammation and endocytosis in particular. Screening for known, familial mutations in genes related to EOFAD and other neurodegenerative diseases revealed no causative variants. An additional common *ABCA7* variant was tentatively associated with AD risk in the candidate gene association analysis, along with rare *LRRK2* missense variant rs35303786. The *LRRK2* variant is predicated as benign in PD, however its suggestive association with AD could be an indicator of genetic pleiotropy across neurodegenerative phenotypes. Furthermore, burden analysis found rare variants in *CLU* to exert a burden effect on AD, although none of the variants were significant individually. Despite only tentative associations with AD, a number of variants were deemed interesting candidates for further investigation, listed in Table 6.1. It is possible that with increased power, many of these variants could reach significance.

Table 6.1. Variants selected for follow up investigations. Variants in the genes *PILRA*, *ABCA7*, *LRRK2* and *CLU* were selected as potential candidates for further investigation. Initially, exploration of the same variants in larger datasets would be ideal, followed by genotyping to confirm that the associations observed here are replicable. MAFs were obtained from the ExAC database for known variants and calculated by Plink for the novel variant in *CLU*.

Gene	Variant	ExAC MAF	Common or Rare	Functional Consequence	Follow up experiments
PILRA	rs2405442	0.33	Common	Synonymous	Genotype
ABCA7	rs3764645	0.48	Common	Missense	Genotype
	rs3752234	0.45	Common	Synonymous	Genotype
	rs3752237	0.32	Common	Synonymous	Genotype
	rs4147915	0.18	Common	Synonymous	Genotype
LRRK2	rs35303786	0.009	Rare	Synonymous	Check in larger dataset
CLU	rs9331939	0.009	Rare	Synonymous	Knockdown experiments of *CLU*
	rs374233511	0.00002	Rare	Synonymous	
	rs149705964	0.0007	Rare	Missense	
	rs41276297	0.0017	Rare	Missense	
	rs201670453	0.005	Rare	Missense	
	rs372043736	0.00002	Rare	Missense	
	rs149859119	0.002	Rare	Missense	
	chr8:27468073	0.002 (calculated)	Rare	Novel	

6.2 High quality data is required to obtain reliable variant calls

Chapter 3 describes the application of the WES data processing pipeline. Quantification and testing the DNA for degradation prior to sequencing reduced the initial potential for poor quality DNA to be sequenced. The quality of NGS data is crucial for accurate and reliable variant detection (Pfeifer, 2016). Therefore, quality control measures were employed at several stages to ensure the most accurate and reliable data was obtained (Guo *et al.,* 2014). Additionally, this monitoring allows potential issues to be pinpointed at source, making it easier to resolve them when they arise. Generating QC metrics for the raw sequence data was an additional step to identify any problems with the starting material or the sequencer (Andrews, 2010). There are several different tools commonly used for determining NGS data quality, with each having its own merits. Therefore, multiple tools were used here to ensure that these quality metrics were reliable.

Compared to other popular variant calling algorithms, the GATK pipeline has been shown to yield higher quality variant calls (Pirooznia *et al.,* 2014). The processing steps allow further refinement of variant calls, resulting in a high-quality dataset. Whilst there are still false positives, downstream filtering steps can help to identify most of these to reduce their impact on the results. Additionally, these false positives can be enriched for novel variants (Warden *et al.,* 2014). Nevertheless, GATK remains one of the most robust methods for processing WES data and generating variant calls, meaning the rest of the analysis can continue with the best possible data quality. Annotations are also imperative to the quality of data as different tools can use different databases, causing discrepancies. Although commercial tools integrate data across the annotation categories, there is often low concordance between the annotations. Therefore, combining tools can help to identify common annotations and prioritise variants (McCarthy *et al.,* 2014; Oliver *et al.,* 2015). Using this combinatorial approach we can

assert a greater degree of confidence in the annotation of the data.

The notion that most genetic variants leading to a severe phenotypic outcome are located in the coding region of the genome and the low costs in comparison to WGS make WES a more suitable choice (Guo *et al.,* 2014). Despite the exome only representing around 1% of the human genome, our understanding of the functionality of non-coding regions is still very poor. GWAS generally capture more known non-coding variants and in most instances it is proving to be extremely difficult to link these signals to the causative variants (Edwards *et al.,* 2013). One advantage of WES is that it can identify coding mutations with the ability to directly affect the protein.

Nevertheless, non-coding variants have been implicated in gene regulation and could play an important role in the disease causation as a result. It has been established that more than 80% of the human genome has an active function contrary to the traditional belief that it comprises 'junk' DNA (ENCODE Project Consortium *et al.,* 2012). WES misses these regions and GWAS targets known variants, therefore, alone they are unable to fully address the complex genetic basis of AD. WGS would cover the entire genome without bias, however this approach is less feasible due to the vast amount of data produced. Nevertheless, with the decreasing cost of sequencing, eventually WGS will become more cost effective without having to sacrifice the level of information for data storage.

6.3 The BDR cohort provides a well-characterised resource for genetics research in dementia

The BDR initiative collects post-mortem brains from donors at 6 centres across the UK, to provide high-quality brain tissue for dementia research studies. The network of brain banks has established standard procedures for documenting donor information, diagnosis and tissue collection to ensure data is consistent across centres. Volunteers undergo regular neuropsychological tests assessing thinking, memory and behaviour. Individual information includes two independent clinical

diagnoses, age at recruitment, onset if applicable and death, sex, ethnicity, MMSE scores, Braak staging, CERAD score and other neuropathological assessments. Two independent neuropathological diagnoses are also provided upon post-mortem examination to obtain the most definitive diagnosis.

Although the current sample size is underpowered for identifying significant associations, the BDR is an ongoing initiative with aims to collect over 2000 brains for researchers in the future. Patients can have a range of diagnoses, including MCI, other dementias and also neurodegenerative diseases such as PD and Huntington's disease. Healthy control participants are also recruited to facilitate case-control studies. As the library of samples grows, increased numbers of other phenotypes will also be available for cross-trait analysis across dementias. Researchers are provided with a complete medical history which makes the BDR unique from other brain banks, as it allows correlation of genetics with symptoms as well as disease severity at death which also has implications for identifying potential protective factors.

The error of diagnosis in dementias is very high, owing to the crossover in symptoms and pathology. Many diseases such as AD can only be accurately confirmed at autopsy. The BDR cohort has extensive clinical information with two independent neuropathological diagnoses to ensure it is as diagnostically accurate as possible. This makes it one of the most well characterised collections available. It also has the potential to be useful in identifying potential protective factors against AD through comparison of cognitive abilities, neuropathological features, risk scores and diagnosis.

6.4 Exome-wide analysis reveals tentatively associated variants in *PILRA* and *ABCA7* genes

Previous WES studies in AD have been effective in uncovering potential causal variants associated with altered disease risk (Guerreiro and Hardy, 2013). Although exploring the 'Common Disease – Rare

Variants' hypothesis is considerably easier using NGS compared to GWAS, large sample sizes are still required to ensure there is enough power to detect them reliably (Zuk *et al.,* 2014). A prime example is the *TREM2* variant, which was significant only after testing in cohorts of greater than 1000 samples in each group (Guerreiro, Wojtas, *et al.,* 2013; Jonsson *et al.,* 2013). The exome-wide analyses in **Chapter 4** did not have enough power to detect associations above the genome-wide or suggestive significance thresholds. This is due to the small sample size in comparison to larger GWAS studies. However, the BDR is an emerging resource meaning the number of samples will increase with time, which should continually improve the power. The advantage of using NGS data for association testing over GWAS is that lower frequency variants can be included. In addition, WES targets coding variants which can directly affect the protein. Nevertheless, in small sample studies it is difficult to detect rare variants unless they exert large effects, which is relatively uncommon in complex diseases (Zuk *et al.,* 2014). In the current dataset, rare variants with 5% MAF would require an effect size of almost 3-fold to be detected which sufficient statistical power.

This analysis identified suggestive associations between synonymous variants in *ABCA7* and *PILRA* with AD. A large number of the associated variants coded for synonymous mutations and were not in LD with a non-synonymous variant, begging the question of their role in disease if they did not alter the protein sequence. Although previously referred to as 'silent' mutations, synonymous variants are now widely recognised to cause changes in protein expression, conformation and their functions (Sauna and Kimchi-Sarfaty, 2011). Four *ABCA7* variants were tentatively associated, strengthening the evidence for this genes role in AD. The only non-synonymous variant was rs3764645, found to be in high LD with two non-coding SNPs that are predicted to affect transcription factor binding, one of which was linked to expression of a target gene. Interestingly, this could indicate that we may not necessarily be looking for rare, causative mutations in LD with the index

SNPs, but it is possible that an accumulative effect of coding and regulatory mutations might affect the regulation or functionality of a gene product. The other variants were not in high LD with any coding mutations, further suggesting that synonymous variants must be exerting some sort of effect on genes to result in their associations. The protective effect of the synonymous *PILRA* SNP rs2405442 was supported by the IGAP data. It is in complete LD with a nearby missense variant rs1859788, which could be contributing to this signal.

The driving hypothesis behind WES studies of complex diseases is that rare coding variants are involved and these variants have low population frequencies (Ng *et al.,* 2010). Therefore, the major hindrance in bioinformatic analyses remains the sample size which determines the statistical power to detect associations with a phenotype. The number of case and control samples will also be affecting the effects observed in the dataset as there are 3 times more cases, potentially skewing the results. Therefore, all findings are tentative and could change as more power is achieved as the collection expands.

6.5 Pathway analysis provides further evidence for immune function in AD pathogenesis

For many of the genes identified through GWAS and NGS, their function has not been fully elucidated. Pathway analysis is an effective tool to predict interactions between genes and assess which are the most relevant biological pathways for further investigation. In this regard, *in silico* analysis has proven to be incredibly useful for directing experimental work into understanding the role of genes and testing functionality. Interestingly, the genes of interest identified in **Chapter 4** were predicted to interact with known AD risk genes; they also linked to the existing pathways from GWAS. Pathway analysis revealed that *PILRA* was co-expressed with *ABCA7*, *CD33*, *SLC24A4*, *TYROBP*, *TREM1* and *TREM2*; all genes except *SLC24A4* have been connected to the immune function pathway. TREM2 has an anti-inflammatory function in the brain, interacting with TYROBP in microglia to trigger

phagocytosis of Aβ and apoptotic neurons. PILRA is paired with PILRB which can directly interact with TYROBP, potentially being involved in the same pathway as *TREM2*. This finding provides even more support for the role of immunity and microglial function in AD pathogenesis, which makes it a viable target for future treatments. However, the functionality or biological mechanism through which these genes interact remains unknown and therefore, this gene warrants further investigation to establish its function and whether it plays a role in AD pathogenesis. New compounds targeting this immune pathway should also be explored as alternate avenues for treatment compared to the traditional Aβ-related drugs.

PILRA was also recently associated with AD in a much larger case-control study by the Alzheimer's Disease Sequencing Project (ADSP), comprising more than 10,000 individuals. This unpublished data provides significant support for the findings in this study and further implicates *PILRA* in AD risk. The gene has been previously implicated in AD via regulation by GWAS risk gene *ZCWPW1* which was nominally associated with reduced *PILRA* levels (Allen *et al.,* 2015); although it has not been directly linked to the risk of developing AD through association testing until now. The interactions between *PILRA* and AD-related genes predicted by pathway analysis, however, should be fully investigated prior to making any conclusions. Gene ontology tools should not be taken at face value because, while they can implicate different pathways and functions, discrepancies between tools can affect the results (Khatri *et al.,* 2012). At present, *in silico* analyses can only inform an association; however, as the tools advance and more databases are generated these predictions will significantly improve.

6.6 Rare and common variants in known GWAS genes could be contributing to AD risk

The Common Disease – Rare Variant hypothesis indicates that rare variants with larger effect sizes are significantly contributing to common disease risk. It has been proposed that such variants are likely to be the

true causative variants, either in LD with the GWAS tag SNP or in close proximity (Rivas *et al.,* 2011).

An additional common variant in *ABCA7* was highly associated with AD; burden analysis revealed that common variants were nominally associated when tested together. *ABCA7* is already known to be involved in AD, however, identifying more variants that may contribute to AD susceptibility helps catalogue potential causative mutations.

Multiple rare, missense and synonymous variants in *CLU* exerted a burden effect on AD, whereas individually they were not associated. Rare coding variants, independent of the common variants identified by GWAS, in this gene have previously been implicated in increased AD risk (Bettens *et al.,* 2012). These findings suggest that both rare and common risk variants within a gene can play a role in disease susceptibility (Bettens *et al.,* 2013). Burden analysis can increase statistical power to detect rare variant associations, as was observed in this case. However, imputing genotypes for other variants within these candidate genes could increase the coverage of genomic variation (Lee *et al.,* 2014).

6.7 Genetic pleiotropy can cause misdiagnosis across different dementias

Samples were screened for familial mutations relating to EOFAD and other neurodegenerative diseases, however none were observed. This suggests that the observed phenotypes were correctly classified as AD. However, misdiagnosis can occur still due to the multitude of overlapping clinical and neuropathological features and genetic pleiotropy across the different diseases (Gratten and Visscher, 2016). Some pleiotropic genes are involved in several neurodegenerative diseases, for example *MAPT* and *APOE* which can influence multiple phenotypes (Desikan *et al.,* 2015). PRS could eventually be implemented to help differentiate AD from other dementias using genetic risk profiles and neuropathology to make more definitive diagnoses.

A rare variant in the PD-associated gene *LRRK2* was tentatively associated with protection against AD, although it is benign in PD; it was only present in 5 controls and 2 cases. However, the pathological information supplied by the BDR on these samples was available and it was confirmed that no 'PD-like' pathology was observed in any of these samples. This is one of the significant benefits of the BDR resource as neuropathological information can be compared with the genetic data.

6.8 PRS can be used for stratifying high and low risk patients for clinical trials

PRS quantify an individual's genetic profile to determine their level of risk of developing a particular disease or phenotype. A correct diagnosis of AD can now be determined with almost 85-90% accuracy (Escott-Price, Shoai, *et al.,* 2017; Patel *et al.,* 2018). PRS are being increasingly utilised in both research studies and clinical trials to stratify patients by disease risk. The recent drug trial testing Solanezumab was performed in patients with mild to moderate AD. The drug was able to clear Aβ plaques effectively but not reduce cognitive decline in moderate AD patients, due to significant neuron loss during disease progression. However, it appeared to show some benefit for those with mild AD as they experienced slower decline in symptoms, although results from a subsequent trial targeting just mild AD patients were not significant. Nevertheless, this highlights the necessity of identifying the at-risk individuals prior to symptom onset for early drug intervention, such as treatments that would aim to ensure amyloid build-up is cleared at the earliest possible stage to reduce Aβ deposition (Marioni *et al.,* 2017). PRS can also quantify the genetic component of AD for individuals to predict whether they could be good candidates for drug trials that target certain pathways or genes. In the future, PRS could be particularly useful for personalised medicines for AD, particularly when treatments targeting the underlying molecular cause become available.

There is significant crossover in the clinical features of different dementias, which makes misdiagnosis quite common. Additionally,

some genetic risk factors associated with the diseases can also overlap. As an example, *APOE*, the biggest genetic AD risk factor has also been implicated in DLB (Bras *et al.,* 2014). *MAPT* is also linked to several diseases aside from AD, including PD, FTD and ALS (Ferrari *et al.,* 2017). Low risk individuals generally exhibited more mixed dementia pathologies and comorbidities than those with higher scores, which could suggest that these diseases have a distinct genetic background from AD. By closely examining the clinical and neuropathological information alongside PRS, we can improve disease classification and also better identify misdiagnoses. For example, if high risk controls have AD pathology, it is more than likely to be misdiagnosis. To facilitate this, however, more extensive clinical and phenotypic information is required for control samples.

There are multiple competing theories for AD pathogenesis, with no single one able to fully explain the phenotype. Complex disorders like this are rarely due to a single cause, instead it is more likely to be a combination of factors that cause susceptibility and potentially lead to developing the disease. Although speculative, it is also probable that the primary cause could be different for individuals, depending on the genetic and environmental risk factors they are exposed to. Elucidating the variants that contributed to the risk profile revealed variants in several known AD genes, supporting their role as susceptibility genes. It also implicated some genes that have been investigated previously but did not have sufficient evidence, including *ACE*, *SPARCL1* and *VAC14*. Furthermore, the genes are involved in multiple pathways, indicating that these risk variants in combination could dysregulate the pathways or initiate the cascade of events that leads to AD, providing a more complete picture of the pathogenesis.

Using either GWAS or WES data to calculate PRS for individuals can be limiting as they do not capture both coding and non-coding variation. The BDR resource has been genotyped on the NeuroChip v2 which comprises ~250,000 genome wide variants with an extra ~25,000

neurodegeneration specific variants. Combining this data with the WES would essentially be targeting variants across the genome, albeit with less coverage than WGS, but with reduced time and expense. The utility of identifying the risk variants used to generate PRS could potentially lead to a small genotyping chip that can be used to test people for risk in a clinical setting.

6.9 Future work

NGS can have a notoriously high rate of false positive variant calls, due to the nature of the technology used, therefore experimental validation of variants is imperative (Lord *et al.,* 2012). Table 6.1 highlights several variants that were deemed worthy of further investigation. All variants should be validated through larger studies prior to any experimental work, as our sample size did not achieve enough power to rule out some of the rarer variants being potentially spurious. Genotyping of samples to confirm the associations with AD would also be necessary before proceeding.

Additionally, it would be interesting to explore the effects of the highly associated synonymous variants on gene and protein levels, which can be tested in cell lines. Synonymous *PILRA* variant rs2405442 was highly associated with reduced AD risk in IGAP, therefore it should certainly be investigated further for its role in AD and interactions with *PILRB* and other AD-related genes *TREM2*, *TYROBP* and *ZCWPW1*. Overexpression in neuronal or microglial cell lines could help elucidate some of the functions of *PILRA* and related genes.

RNA sequencing of the BDR cohort is the next envisaged step in characterising the resource and would allow us to associate genetic variants with gene expression.

6.10 Future of Alzheimer's genetics

LOAD is a multifactorial disease with complex aetiology; interactions between genetic and environmental factors may all contribute. The most recent meta-GWAS in 2013 identified susceptibility loci associated with AD risk, which gave us 20 new genes and novel pathways to explore (Lambert *et al.,* 2013). However, the roles of all of these genes remain to be fully elucidated. The advent of NGS has allowed investigation into the "Common Disease - Rare Variant" hypothesis, to determine the contribution of rare variants in AD. Sequencing studies have identified a number of genes harbouring rare variants associated with disease. However, sample sizes can be limited by the cost and time of sequencing, affecting the power to detect rare variants. A recent case-control study using a whole-exome genotyping array targeted known exonic variants in 85,133 samples identified new variants in three immune-related genes, *PLCG2*, *ABI3* and *TREM2* (Sims *et al.,* 2017). However, it is becoming increasingly apparent that combinations of common and rare variants are more likely to impact disease risk (Gibson, 2012).

The polygenic architecture of AD has been investigated using PRS analysis. The utility of PRS have been demonstrated in AD, with studies now achieving nearly 90% predictive accuracy to discriminate between cases and controls (Escott-Price, Shoai, *et al.,* 2017) especially at the extremes of the score ranges. This method has also shown that individuals stratified by genetic risk factors can also be distinguished by clinical features. AD patients with high PRS showed cognitive decrements compared to the low risk group (Marioni *et al.,* 2017). This could prove useful for selecting participants for clinical trials as those at high risk could be identified at preclinical stages and benefit to a greater extent from treatments and early intervention.

We may have apparently exploited much of the genetics of AD; however, a large proportion of the heritability remains unaccounted for. The focus now appears to be shifting towards investigating gene

expression and the epigenetic architecture of the disease. Integrated analysis using DNA-sequencing, gene expression and epigenetic data is critical to understanding many diseases and it would provide the BDR with a significant advantage. This also has implications for improvements in diagnosis and refining administered treatments.

6.11 Conclusion

In this project, we aimed to explore the BDR resource using a gold standard workflow to identify exonic variants associated with the risk of developing AD. We have uncovered tentative associations both in previous and newly identified loci; highlighting potential new genes of interest for further investigation. We also demonstrated the potential utility of the extensive phenotyping available for the BDR resource in predicting AD patients using PRS. This has also provided further evidence for the usefulness of PRS, which will continue to be an invaluable tool for research and it is hoped that it also could aid in clinical diagnostics in the near future. Furthermore, we hope that as the BDR collection grows in size, we will be able to increase the power of this study to detect significant associations with AD.

References

1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature 526, 68–74.

Abbas, N., Lücking, C.B., Ricard, S., Dürr, A., Bonifati, V., De Michele, G., Bouley, S., *et al.,* 1999. A wide variety of mutations in the parkin gene are responsible for autosomal recessive parkinsonism in Europe. Hum. Mol. Genet. 8, 567–574.

Abduljaleel, Z., Al-Allaf, F.A., Khan, W., Athar, M., Shahzad, N., Taher, M.M., Elrobh, M., *et al.,* 2014. Evidence of TREM2 variant associated with triple risk of Alzheimer's disease. PLoS One 9.

Adzhubei, I., Jordan, D.M., Sunyaev, S.R., 2013. Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. 7, 7–20.

Ahmed, I., Tamouza, R., Delord, M., Krishnamoorthy, R., Tzourio, C., Mulot, C., Nacfer, M., *et al.,* 2012. Association between Parkinson's disease and the HLA-DRB1 locus. Mov. Disord. 27, 1104–1110.

Alam, M.M., Lee, J., Lee, S.Y., 2017. Recent progress in the development of TSPO PET ligands for neuroinflammation imaging in neurological diseases. Nucl. Med. Mol. Imaging. 51, 283-296.

Allen, M., Kachadoorian, M., Carrasquillo, M.M.M., Karhade, A., Manly, L., Burgess, J.D., Wang, C., *et al.,* 2015. Late-onset Alzheimer disease risk variants mark brain regulatory loci. Neurol. Genet. 1, e15.

Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E.B., Müller-Myhsok, B., 2012. A beginner's guide to SNP calling from high-throughput DNA-sequencing data. Hum. Genet. 131, 1541-1554.

Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., *et al.,* 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65.

Alzheimer's Association, 2014. 2014 Alzheimer's Disease Facts and Figures. Alzheimer's Dement.

Amos, W., Driscoll, E., Hoffman, J.I., 2011. Candidate genes versus genome-wide associations: which are better for detecting genetic susceptibility to infectious disease? Proceedings. Biol. Sci. 278, 1183–1188.

Andel, R., Crowe, M., Pedersen, N.L., Mortimer, J., Crimmins, E., Johansson, B., Gatz, M., 2005. Complexity of work and risk of Alzheimer's disease: A population-based study of Swedish twins. J. Gerontol. B. Psychol. Sci. Soc. Sci. 60, 251-258.

Anderson, C. a, Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, P., Zondervan, K.T., 2011. Data quality control in genetic case-control association studies. Nat. Protoc. 5, 1564–1573.

Ando, K., Brion, J.P., Stygelbout, V., Suain, V., Authelet, M., Dedecker, R., Chanut, A., *et al.,* 2013. Clathrin adaptor CALM/PICALM is associated with neurofibrillary tangles and is cleaved in Alzheimer's brains. Acta Neuropathol. 125, 861-878.

Andrews, S., 2010. FastQC: A quality control tool for high throughput sequence data. Babraham Bioinforma. http://www.bioinformatics.babraham.ac.uk/projects/.

Annese, A., Manzari, C., Lionetti, C., Picardi, E., Horner, D.S., Chiara, M., Caratozzolo, M.F., *et al.,* 2018. Whole transcriptome profiling of Late-Onset

Alzheimer's Disease patients provides insights into the molecular changes involved in the disease. Sci. Rep. 8, 4282.

Antonell, A., Lladó, A., Altirriba, J., Botta-Orfila, T., Balasa, M., Fernández, M., Ferrer, I., *et al.,* 2013. A preliminary study of the whole-genome expression profile of sporadic and monogenic early-onset Alzheimer's disease. Neurobiol. Aging. 24, 1772-1778.

Armstrong, R., 2012. On the 'classification' of neurodegenerative disorders: Discrete entities, overlap or continuum? Folia Neuropathol. 3, 201–218.

Baig, S., Joseph, S.A., Tayler, H., Abraham, R., Owen, M.J., Williams, J., Kehoe, P.G., *et al.,* 2010. Distribution and expression of picalm in Alzheimer disease. J. Neuropathol. Exp. Neurol. 69, 1071–1077.

Ballatore, C., Lee, V.M.Y., Trojanowski, J.Q., 2007. Tau-mediated neurodegeneration in Alzheimer's disease and related disorders. Nat. Rev. Neurosci. 8, 663-672.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., Shendure, J., 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. 12, 745–755.

Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W. a, Jiang, H., Feng, G., 2014. Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. Lib. Acad. 13, 67–82.

Barber, I.S., Braae, A., Clement, N., Patel, T., Guetta-Baranes, T., Brookes, K., Medway, C., *et al.,* 2017. Mutation analysis of sporadic early-onset Alzheimer's disease using the NeuroX array. Neurobiol. Aging 49, 215.e1-215.e8.

Beach, T.G., Monsell, S.E., Phillips, L.E., Kukull, W., 2012. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010. J. Neuropathol. Exp. Neurol. 71, 266–273.

Beecham, G.W., Hamilton, K., Naj, A.C., Martin, E.R., Huentelman, M., Myers, A.J., Corneveaux, J.J., *et al.,* 2014. Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. PLoS Genet. 10, e1004606.

Beecham, G.W., Martin, E.R., Li, Y.-J., Slifer, M.A., Gilbert, J.R., Haines, J.L., Pericak-Vance, M.A., 2009. Genome-wide Association Study Implicates a Chromosome 12 Risk Locus for Late-Onset Alzheimer Disease. Am. J. Hum. Genet. 84, 35–43.

Beel, A.J., Sakakura, M., Barrett, P.J., Sanders, C.R., 2010. Direct binding of cholesterol to the amyloid precursor protein: An important interaction in lipid–Alzheimer's disease relationships? Biochim. Biophys. Acta - Mol. Cell Biol. Lipids 1801, 975–982.

Beisang, D., Reilly, C., Bohjanen, P.R., 2014. Alternative polyadenylation regulates CELF1/CUGBP1 target transcripts following T cell activation. Gene. 550, 93-100.

Bekris, L.M., Yu, C.-E., Bird, T.D., Tsuang, D.W., 2010. Review Article: Genetics of Alzheimer Disease. J. Geriatr. Psychiatry Neurol. 23, 213–227.

Benitez, B.A., Jin, S.C., Guerreiro, R., Graham, R., Lord, J., Harold, D., Sims, R., *et al.,* 2014. Missense variant in TREML2 protects against Alzheimer's disease. Neurobiol. Aging 35, e19-26.

Benitez, B.A., Karch, C.M., Cai, Y., Jin, S.C., Cooper, B., Carrell, D., Bertelsen, S., *et al.,* 2013. The PSEN1, p.E318G Variant Increases the Risk of Alzheimer's Disease in APOE-ε4 Carriers. PLoS Genet. 9, e1003685.

Bertram, L., Lange, C., Mullin, K., Parkinson, M., Hsiao, M., Hogan, M.F., Schjeide, B.M.M., *et al.,* 2008. Genome-wide Association Analysis Reveals Putative Alzheimer's Disease Susceptibility Loci in Addition to APOE. Am. J. Hum. Genet. 83, 623–632.

Bertram, L., Lill, C.M., Tanzi, R.E., 2010. The genetics of Alzheimer's disease: Back to the future. Neuron 68, 270–281.

Bertram, L., McQueen, M.B., Mullin, K., Blacker, D., Tanzi, R.E., 2007. Systematic meta-analyses of Alzheimer's disease genetic association studies: The AlzGene database. Nat. Genet. 39, 17–23.

Bettens, K., Brouwers, N., Engelborghs, S., Lambert, J.C., Rogaeva, E., Vandenberghe, R., Le Bastard, N., *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.

Bettens, K., Sleegers, K., Van Broeckhoven, C., Broeckhoven, C. Van, 2013. Genetic insights in Alzheimer's disease. Lancet Neurol. 12, 92–104.

Bianchin, M.M., Capella, H.M., Chaves, D.L., Steindel, M., Grisard, E.C., Ganev, G.G., Da Silva, J.P., *et al.,* 2004. Nasu-Hakola Disease (Polycystic Lipomembranous Osteodysplasia with Sclerosing Leukoencephalopathy-PLOSL): A Dementia Associated with Bone Cystic Lesions. From Clinical to Genetic and Molecular Aspects. Cell. Mol. Neurobiol. 24, 1-24.

Bird, T., 2013. Alzheimer Disease Overview. GeneReviews.

Blacker, D., Albert, M.S., Bassett, S.S., Go, R.C.., Harrell, L.E., 1994. Reliability and validity of NINCDS-ADRDA criteria for Alzheimer's disease: The National Institute of Mental Health Genetics Initiative. Arch. Neurol. 51, 1198-1204.

Blennow, K., 2004. CSF biomarkers for mild cognitive impairment. J. Intern. Med. 256, 224–234.

Blennow, K., Hampel, H., 2003. CSF markers for incipient Alzheimer's disease. Lancet. Neurol. 2, 605–613.

Bloom, G.S., 2014. Amyloid-β and tau: The trigger and bullet in Alzheimer's disease pathogenesis. JAMA Neurol. 71, 505.

Braae, A., Medway, C., Carrasquillo, M., Younkin, S., Kehoe, P.G., Morgan, K., Craig, D., *et al.,* 2015. Blood type gene locus has no influence on ACE association with Alzheimer's disease. Neurobiol. Aging 36, 1767.e1-1767.e2.

Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H., Del Tredici, K., 2006. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. Acta Neuropathol. 112, 389–404.

Braak, H., Braak, E., 1991. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 82, 239–259.

Bras, J., Guerreiro, R., Darwent, L., Parkkinen, L., Ansorge, O., Escott-Price, V., Hernandez, D.G., *et al.,* 2014. Genetic analysis implicates APOE, SNCA and suggests lysosomal dysfunction in the aetiology of dementia with Lewy bodies. Hum. Mol. Genet. 23, 6139–6146.

Brookes, K.J., Morgan, K., 2017. Genetics of Alzheimer's disease. In: ELS. John Wiley & Sons, Ltd, Chichester, UK, pp. 1–16.

Burns, A., Iliffe, S., 2009. Alzheimer's Disease. BMJ (British Med. Journal) 338, 467–471.

Burns, M.P., Igbavboa, U., Wang, L., Wood, W.G., Duff, K., 2006. Cholesterol distribution, not total levels, correlate with altered amyloid precursor protein processing in statin-treated mice. Neuromolecular Med. 8, 319–328.

Burns, M.P., Noble, W.J., Olm, V., Gaynor, K., Casey, E., LaFrancois, J., Wang, L., *et al.,* 2003. Co-localization of cholesterol, apolipoprotein E and fibrillar Abeta in amyloid plaques. Brain Res. Mol. Brain Res. 110, 119–125.

Bush, W.S., Moore, J.H., 2012. Chapter 11: Genome-Wide Association Studies. PLoS Comput. Biol.

Carrasquillo, M.M., Belbin, O., Hunter, T.A., Ma, L., Bisceglio, G.D., Zou, F., Crook, J.E., *et al.,* 2010. Replication of CLU, CR1, and PICALM associations with Alzheimer's disease. Arch. Neurol. 67, 961–964.

Carrasquillo, M.M., Zou, F., Pankratz, V.S., Wilcox, S.L., Ma, L., Walker, L.P., Younkin, S.G., *et al.,* 2009. Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. Nat. Genet. 41, 192–198.

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

Castellani, R.J., Lee, H., Siedlak, S.L., Nunomura, A., Hayashi, T., Nakamura, M., Zhu, X., *et al.,* 2009. Reexamining Alzheimer's disease: Evidence for a protective role for amyloid-beta protein precursor and amyloid-beta. J. Alzheimer's Dis. 18, 447–452.

Chan, S.L., Kim, W.S., Kwok, J.B., Hill, A.F., Cappai, R., Rye, K.-A., Garner, B., 2008. ATP-binding cassette transporter A7 regulates processing of amyloid precursor protein in vitro. J. Neurochem. 106, 793–804.

Chapuis, J., Hansmannel, F., Gistelinck, M., Mounier, A., Van Cauwenberghe, C., Kolen, K. V., Geller, F., *et al.,* 2013. Increased expression of BIN1 mediates Alzheimer genetic risk by modulating tau pathology. Mol. Psychiatry 18, 1225–1234.

Chaudhry, M., Wang, X., Bamne, M.N., Hasnain, S., Demirci, F.Y., Lopez, O.L., Kamboh, M.I., 2015. Genetic variation in imprinted genes is associated with risk of late-onset Alzheimer's disease. J. Alzheimer's Dis. 44, 989–994.

Chaudhury, S., Patel, T., Barber, I.S., Guetta-Baranes, T., Brookes, K.J., Chappell, S., Turton, J., *et al.,* 2017. Polygenic risk score in post-mortem diagnosed sporadic early onset Alzheimer's disease. Neurobiol. Aging. 62, e1-e8.

Cheng, J., Guo, X., Zhang, T., Zhong, L., Bu, G., Chen, X., 2016. TREMs in Alzheimer's disease: Genetic and clinical investigations. 463, 88-95.

Cheng, L., Doecke, J.D., Sharples, R.A., Villemagne, V.L., Fowler, C.J., Rembach, A., Martins, R.N., *et al.,* 2015. Prognostic serum miRNA biomarkers associated with Alzheimer's disease shows concordance with neuropsychological and neuroimaging assessment. Mol. Psychiatry 20, 1188–1196.

Chilamakuri, C.S.R., Lorenz, S., Madoui, M.-A., Vodák, D., Sun, J., Hovig, E., Myklebost, O., *et al.,* 2014. Performance comparison of four exome capture systems for deep sequencing. BMC Genomics 15, 449.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., *et al.,* 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing 106, 19096–101.

Chun, S., Fay, J.C., 2009. Identification of deleterious mutations within three human genomes. Genome Res. 19, 1553–1561.

Churcher, I., 2006. Tau therapeutic strategies for the treatment of Alzheimer's disease. Curr. Top. Med. Chem. 6, 579-595.

Cingolani, P., Platts, A., Wang, L.L.L., Coon, M., Nguyen, T., Wang, L.L.L., Land, S.J., *et al.,* 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 6, 80–92.

Cirrito, J.R., Kang, J.-E., Lee, J., Stewart, F.R., Verges, D.K., Silverio, L.M., Bu, G., *et al.,* 2008. Endocytosis is required for synaptic activity-dependent release of amyloid-beta in vivo. Neuron 58, 42–51.

Cirulli, V., Yebra, M., 2007. Netrins: beyond the brain. Nat. Rev. Mol. Cell Biol. 8, 296–306.

Clark, R.I., Tan, S.W.S., Péan, C.B., Roostalu, U., Vivancos, V., Bronda, K., Pilátová, M., *et al.,* 2013. MEF2 is an in vivo immune-metabolic switch. Cell. 155. 155, 435-447.

Clarke, G.M., Anderson, C.A., Pettersson, F.H., Cardon, L.R., Morris, A.P., Zondervan, K.T., 2011. Basic statistical analysis in genetic case-control studies. Nat. Protoc. 6, 121-133.

Clement, N., Braae, A., Turton, J., Lord, J., Guetta Baranes, T., Medway, C., Brookes, K.J., *et al.,* 2016. Investigating splicing variants uncovered by next-generation sequencing the Alzheimer's disease candidate genes, CLU, PICALM, CR1, ABCA7, BIN1, the MS4A locus, CD2AP, EPHA1 and CD33. J. Alzheimer's Dis. Park. 6, 1–7.

Cobb, J.L., Wolf, P.A., Au, R., White, R., D'Agostino, R.B., 1995. The effect of education on the incidence of dementia and Alzheimer's disease in the Framingham Study. Neurology. 45, 1707-1712.

Colonna, M., 2003. TREMs in the immune system and beyond. Nat. Rev. Immunol. 3, 445–453.

Čolović, M.B.., Krstić, D.Z.., Lazarević-Pašti, T.D.., Bondžić, A.M.., Vasić, V.M.., 2013. Acetylcholinesterase inhibitors: Pharmacology and toxicology. Curr. Neuropharmacol.11, 315-335.

Coon, K.D., Myers, A.J., Craig, D.W., Webster, J.A., Pearson, J. V., Lince, D.H., Zismann, V.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, 613–618.

Corder, E.H., Saunders, A.M., Risch, N.J., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Rimmler, J.B., *et al.,* 1994. Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. Nat. Genet. 7, 180–184.

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

Corrada, M.M., Brookmeyer, R., Paganini-Hill, A., Berlau, D., Kawas, C.H., 2010. Dementia incidence continues to increase with age in the oldest old the 90+ study. Ann. Neurol. 67, 114-121.

Costello, H., Hayes, G.M., Highton-Williamson, E., Nurock, S., Hanbury, D., Francis, P.T., 2016. A pilot study of potential brain donor satisfaction and attitudes towards telephone assessment. Int. J. Geriatr. Psychiatry. 32, 1247-1256.

Crocker, P.R., Paulson, J.C., Varki, A., 2007. Siglecs and their roles in the immune system. Nat. Rev. Immunol. 7, 255-266.

Cruts, M., Theuns, J., Van Broeckhoven, C., 2012. Locus-specific mutation databases for neurodegenerative brain diseases. Hum. Mutat. 33, 1340–1344.

Currinn, H., Guscott, B., Balklava, Z., Rothnie, A., Wassmer, T., 2016. APP controls the formation of PI(3,5)P2 vesicles through its binding of the PIKfyve complex. Cell. Mol. Life Sci. 73, 393–408.

Curtis, D., 2017. Construction of an Exome-Wide Risk Score for Schizophrenia Based on a Weighted Burden Test. Ann. Hum. Genet. 82, 11-22.

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

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., *et al.,* 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158.

Darst, B.F., Koscik, R.L., Racine, A.M., Oh, J.M., Krause, R.A., Carlsson, C.M., Zetterberg, H., *et al.,* 2017. Pathway-specific polygenic risk scores as predictors of amyloid-β deposition and cognitive function in a sample at increased risk for Alzheimer's disease. J. Alzheimers. Dis. 55, 473–484.

De Roeck, A., Van den Bossche, T., van der Zee, J., Verheijen, J., De Coster, W., Van Dongen, J., Dillen, L., *et al.,* 2017. Deleterious ABCA7 mutations and transcript rescue mechanisms in early onset Alzheimer's disease. Acta Neuropathol. 134, 475–487.

De Strooper, B., Saftig, P., Craessaerts, K., Vanderstichele, H., Guhde, G., Annaert, W., Von Figura, K., *et al.,* 1998. Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. Nature 391, 387–390.

De Summa, S., Malerba, G., Pinto, R., Mori, A., Mijatovic, V., Tommasi, S., 2017. GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC Bioinformatics 18, 119.

Delay, C., Mandemakers, W., Hébert, S.S., 2012. MicroRNAs in Alzheimer's disease. Neurobiol. Dis. 46, 285–290.

DeMattos, R.B., Cirrito, J.R., Parsadanian, M., May, P.C., O'Dell, M.A., Taylor, J.W., Harmony, J.A.K., *et al.,* 2004. ApoE and clusterin cooperatively suppress Abeta levels and deposition: evidence that ApoE regulates extracellular Abeta metabolism in vivo. Neuron 41, 193–202.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J.R., Hartl, C., Philippakis, A.A., *et al.,* 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498.

Desikan, R.S., Fan, C.C., Wang, Y., Schork, A.J., Cabral, H.J., Cupples, L.A., Thompson, W.K., *et al.,* 2017. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. PLoS Med. 14, e1002258.

Desikan, R.S., Schork, A.J., Wang, Y., Witoelar, A., Sharma, M., McEvoy, L.K., Holland, D., *et al.,* 2015. Genetic overlap between Alzheimer's disease and Parkinson's disease at the MAPT locus. Mol. Psychiatry 20, 1588–1595.

Di Paolo, G., Sankaranarayanan, S., Wenk, M.R., Daniell, L., Perucco, E., Caldarone, B.J., Flavell, R., *et al.,* 2002. Decreased synaptic vesicle recycling efficiency and cognitive deficits in amphiphysin 1 knockout mice. Neuron 33, 789–804.

Dodson, S.E., Andersen, O.M., Karmali, V., Fritz, J.J., Cheng, D., Peng, J., Levey, A.I., *et al.,* 2008. Loss of LR11/SORLA Enhances Early Pathology in a Mouse Model

of Amyloidosis: Evidence for a Proximal Role in Alzheimer's Disease. J. Neurosci. 28, 12877–12886.

Doecke, J.D., Laws, S.M., Faux, N.G., Wilson, W., Burnham, S.C., Lam, C.-P., Mondal, A., *et al.,* 2012. Blood-Based Protein Biomarkers for Diagnosis of Alzheimer Disease. Arch. Neurol. 69, 1318.

Doody, R.S., Thomas, R.G., Farlow, M., Iwatsubo, T., Vellas, B., Joffe, S., Kieburtz, K., *et al.,* 2014. Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. N. Engl. J. Med. 370, 311–321.

Dosunmu, R., Wu, J., Basha, M.R., Zawia, N.H., 2007. Environmental and dietary risk factors in Alzheimer's disease. Expert Rev. Neurother. 7, 887-900.

Dourlen, P., Fernandez-Gomez, F.J., Dupont, C., Grenier-Boley, B., Bellenguez, C., Obriot, H., Caillierez, R., *et al.,* 2017. Functional screening of Alzheimer risk loci identifies PTK2B as an in vivo modulator and early marker of Tau pathology. Mol. Psychiatry 22, 874–883.

Dubois, B., Feldman, H.H., Jacova, C., DeKosky, S.T., Barberger-Gateau, P., Cummings, J., Delacourte, A., *et al.,* 2007. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol. 6, 734-746.

Dudbridge, F., 2013. Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet. 9. 9, e10003348.

Dustin, M.L., Olszowy, M.W., Holdorf, A.D., Li, J., Bromley, S., Desai, N., Widder, P., *et al.,* 1998. A novel adaptor protein orchestrates receptor patterning and cytoskeletal polarity in T-cell contacts. Cell. 94, 667-677.

Ebbert, M.T.W., Ridge, P.G., Wilson, A.R., Sharp, A.R., Bailey, M., Norton, M.C., Tschanz, J.T., *et al.,* 2014. Population-based analysis of Alzheimer's disease risk alleles implicates genetic interactions. Biol. Psychiatry 75, 732–737.

Ebbert, M.T.W., Wadsworth, M.E., Staley, L.A., Hoyt, K.L., Pickett, B., Miller, J., Duce, J., *et al.,* 2016. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics 17, 239.

Edwards, S.L., Beesley, J., French, J.D., Dunning, M., 2013. Beyond GWASs: Illuminating the dark road from association to function. Am. J. Hum. Genet. 93, 779-797.

Efthymiou, A.G., Goate, A.M., 2017. Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. Mol. Neurodegener. 12, 43.

ENCODE Project Consortium, Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., *et al.,* 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Endrullat, C., Glökler, J., Franke, P., Frohme, M., 2016. Standardization and quality management in next-generation sequencing. Appl. Transl. Genomics 10, 2–9.

Ertekin-Taner, N., 2007. Genetics of Alzheimer's Disease: A Centennial Review. Neurol. Clin.25, 611-667.

Escott-Price, V., Myers, A.J., Huentelman, M., Hardy, J., 2017. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. Ann. Neurol. 82, 311–314.

Escott-Price, V., Shoai, M., Pither, R., Williams, J., Hardy, J., 2017. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. Neurobiol. Aging 49, 214.

Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., Badarinarayan, N., *et al.,* 2015. Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain 138, 3673–3684.

Euesden, J., Lewis, C., O'Reilly, P., 2014. PRSice: polygenic risk score software. Bioinformatics. 31, 1466-1468.

Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048.

Ewers, M., Mattsson, N., Minthon, L., Molinuevo, J.L., Antonell, A., Popp, J., Jessen, F., *et al.,* 2015. CSF biomarkers for the differential diagnosis of Alzheimer's disease: A large-scale international multicenter study. Alzheimer's Dement. 11, 1306–1315.

Farlow, M., Arnold, S.E., van Dyck, C.H., Aisen, P.S., Snider, B.J., Porsteinsson, A.P., Friedrich, S., *et al.,* 2012. Safety and biomarker effects of solanezumab in patients with Alzheimer's disease. Alzheimer's Dement. 8, 261–271.

Farrer, L.A., 1997. Effects of age, sex and ethnicity on the association between Apolipoprotein E genotype and Alzheimer's disease. JAMA 278, 1349.

Ferrando-Miguel, R., Rosner, M., Freilinger, A., Lubec, G., Hengstschläger, M., 2005. Tuberin – A New Molecular Target in Alzheimer's Disease? Neurochem. Res. 30, 1413–1419.

Ferrari, R., Kapogiannis, D., Huey, E.D., Momeni, P., 2011. FTD and ALS: a tale of two diseases. Curr. Alzheimer Res. 8, 273–294.

Ferrari, R., Wang, Y., Vandrovcova, J., Guelfi, S., Witeolar, A., Karch, C.M., Schork, A.J., *et al.,* 2017. Genetic architecture of sporadic frontotemporal dementia and overlap with Alzheimer's and Parkinson's diseases. J. Neurol. Neurosurg. Psychiatry 88, 152–164.

Ferri, C., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., Hall, K., *et al.,* 2005. Global prevalence of dementia: a Delphi consensus study. Lancet 366, 2112–2117.

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12, 189–198.

Fonseca, M.I., Chu, S., Pierce, A.L., Brubaker, W.D., Hauhart, R.E., Mastroeni, D., Clarke, E. V., *et al.,* 2016. Analysis of the putative role of CR1 in Alzheimer's disease: Genetic association, expression and function. PLoS One. 11, e0149792.

Forlenza, O. V., Radanovic, M., Talib, L.L., Aprahamian, I., Diniz, B.S., Zetterberg, H., Gattaz, W.F., 2015. Cerebrospinal fluid biomarkers in Alzheimer's disease: Diagnostic accuracy and prediction of dementia. Alzheimer's Dement. Diagnosis, Assess. Dis. Monit. 1, 455–463.

Formichi, P., Battisti, C., Radi, E., Federico, A., 2006. Cerebrospinal fluid tau, Aß, and phosphorylated tau protein for the diagnosis of Alzheimer's disease. J. Cell. Physiol. 208, 39–46.

Förstl, H., Kurz, A., 1999. Clinical features of Alzheimer's disease. Eur. Arch. Psychiatry Clin. Neurosci. 249, 288–290.

Fox, E.J., Reid-Bayliss, K.S., Emond, M.J., Loeb, L.A., 2014. Accuracy of next-generation sequencing platforms. Next Gener. Seq. Appl. 1, 1000106.

Francis, P.T., Palmer, A.M., Snape, M., Wilcock, G.K., 1999. The cholinergic hypothesis of Alzheimer's disease: a review of progress. J. Neurol. Neurosurg. Psychiatry. 67, 558.

Fraser, J., Williamson, I., Bickmore, W.A., Dostie, J., 2015. An overview of genome organisation and how we got there: From FISH to Hi-C. Microbiol. Mol. Biol. Rev. 79, 347–372.

Fuller, C.W., Middendorf, L.R., Benner, S.A., Church, G.M., Harris, T., Huang, X., Jovanovich, S.B., *et al.,* 2009. The challenges of sequencing by synthesis. Nat. Biotechnol. 27, 1013–1023.

Gamba, P., Testa, G., Sottero, B., Gargiulo, S., Poli, G., Leonarduzzi, G., 2012. The link between altered cholesterol metabolism and Alzheimer's disease. Ann. N. Y. Acad. Sci. 1259, 54–64.

Gatz, M., Reynolds, C.A., Fratiglioni, L., Johansson, B., Mortimer, J.A., Berg, S., Fiske, A., *et al.,* 2006. Role of Genes and Environments for Explaining Alzheimer Disease. Arch. Gen. Psychiatry 63, 168.

Gauderman, W.J., 2002. Sample size requirements for matched case-control studies of gene-environment interaction. Stat. Med. 21, 35–50.

Ghosh, A., Khandelwal, N., Kumar, A., Bera, A.K., 2017. Leucine-rich repeat-containing 8B protein is associated with the endoplasmic reticulum Ca2+ leak in HEK293 cells. J. Cell Sci. 130, 3818–3828.

Gibson, G., 2012. Rare and common variants: Twenty arguments. Nat. Rev. Genet.

Giri, M., Zhang, M., Lü, Y., 2016. Genes associated with Alzheimer's disease: An overview and current status. Clin. Interv. Aging. 11, 665-681.

Goedert, M., Jakes, R., Vanmechelen, E., 1995. Monoclonal antibody AT8 recognises tau protein phosphorylated at both serine 202 and threonine 205. Neurosci. Lett. 189, 167-169.

Goedert, M., Spillantini, M.G., 2006. A Century of Alzheimer's Disease. Science (80-.). 314, 777–781.

Goh, G., Choi, M., 2012. Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. Genomics Inform. 10, 214–219.

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., *et al.,* 2006. Comparative analysis identifies exonic splicing regulatory sequences: The complex definition of enhancers and silencers. Mol. Cell 22, 769–781.

Grada, A., Weinbrecht, K., 2013. Next-generation sequencing: Methodology and application. J. Invest. Dermatol. 133, e11.

Gratten, J., Visscher, P.M., 2016. Genetic pleiotropy in complex traits and diseases: Implications for genomic medicine 8, 78.

Griciuc, A., Serrano-Pozo, A., Parrado, A.R., Lesinski, A.N., Asselin, C.N., Mullin, K., Hooli, B., *et al.,* 2013. Alzheimer's disease risk gene CD33 inhibits microglial uptake of amyloid beta. Neuron 78, 631–643.

Grupe, A., Abraham, R., Li, Y., Rowland, C., Hollingworth, P., Morgan, A., Jehu, L., *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, 865–873.

Gu, L., Guo, Z., 2013. Alzheimer's Aβ42 and Aβ40 peptides form interlaced amyloid fibrils. J. Neurochem. 126, 305–311.

Guerreiro, R., Bras, J., Hardy, J., Singleton, A., 2014. Next generation sequencing techniques in neurological diseases: Redefining clinical and molecular associations. Hum. Mol. Genet. 23, 47–53.

Guerreiro, R., Hardy, J., 2013. TREM2 and Neurodegenerative Disease. N. Engl. J. Med. 369, 1564–1570.

Guerreiro, R., Hardy, J., 2014. Genetics of Alzheimer's disease. Neurotherapeutics 11, 732–737.

Guerreiro, R., Lohmann, E., Bras, J., Gibbs, J.R., Rohrer, J.D., Gurunlian, N., Dursun, B., *et al.,* 2013. Using exome sequencing to reveal mutations in TREM2 presenting as a frontotemporal dementia–like syndrome without bone involvement. JAMA Neurol. 70, 78.

Guerreiro, R., Wojtas, A., Bras, J., Carrasquillo, M., Rogaeva, E., Majounie, E., Cruchaga, C., *et al.,* 2013. TREM2 variants in Alzheimer's disease. N. Engl. J. Med. 368, 117–127.

Guo, Y., Long, J., He, J., Li, C.I., Cai, Q., Shu, X.O., Zheng, W., *et al.,* 2012. Exome sequencing generates high quality data in non-target regions. BMC Genomics. 13, 194.

Guo, Y., Ye, F., Sheng, Q., Clark, T., Samuels, D.C., 2014. Three-stage quality control strategies for DNA re-sequencing data. Brief. Bioinform. 15, 879–889.

Hampel, H., Buerger, K., Zinkowski, R., Teipel, S.J., Goernitz, A., Andreasen, N., Sjoegren, M., *et al.,* 2004. Measurement of Phosphorylated Tau Epitopes in the Differential Diagnosisof Alzheimer Disease. Arch. Gen. Psychiatry 61, 95.

Hansson, O., Zetterberg, H., Vanmechelen, E., Vanderstichele, H., Andreasson, U., Londos, E., Wallin, A., *et al.,* 2010. Evaluation of plasma Aβ40 and Aβ42 as predictors of conversion to Alzheimer's disease in patients with mild cognitive impairment. Neurobiol. Aging 31, 357–367.

Hardy, J., Allsop, D., 1991. Amyloid deposition as the central event in the aetiology of Alzheimer's disease. Trends Pharmacol. Sci. 12, 383-388.

Hardy, J., Guerreiro, R., Lovestone, S., 2011. Clusterin as an Alzheimer's biomarker. Arch. Neurol. 68, 1459.

Harms, A.S., Cao, S., Rowse, A.L., Thome, A.D., Li, X., Mangieri, L.R., Cron, R.Q., *et al.,* 2013. MHCII is required for α-synuclein-induced activation of microglia, CD4 T cell proliferation and dopaminergic neurodegeneration. J. Neurosci. 33, 9592–9600.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M.L., Pahwa, J.S., *et al.,* 2009. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat. Genet. 41, 1088–1093.

He, F., Umehara, T., Saito, K., Harada, T., Watanabe, S., Yabuki, T., Kigawa, T., *et al.,* 2010. Structural insight into the zinc finger CW domain as a histone modification reader. Structure 18, 1127–1139.

Heinrich, V., Stange, J., Dickhaus, T., Imkeller, P., Krüger, U., Bauer, S., Mundlos, S., *et al.,* 2012. The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. Nucleic Acids Res. 40, 2426–2431.

Heneka, M.T., Carson, M.J., Khoury, J. El, Landreth, G.E., Brosseron, F., Feinstein, D.L., Jacobs, A.H., *et al.,* 2015. Neuroinflammation in Alzheimer's disease, The Lancet Neurology. 14, 388-405.

Herholz, K., Westwood, S., Haense, C., Dunn, G., 2011. Evaluation of a Calibrated 18F-FDG PET Score as a Biomarker for Progression in Alzheimer Disease and Mild Cognitive Impairment. J. Nucl. Med. 52, 1218–1226.

Hickman, S.E., El Khoury, J., 2014. TREM2 and the neuroimmunology of Alzheimer's disease. Biochem. Pharmacol. 88, 495–498.

Hippen, A.A., Ebbert, M.T.W., Norton, M.C., Tschanz, J.T., Munger, R.G., Corcoran, C.D., Kauwe, J.S.K., 2016. Presenilin E318G variant and Alzheimer's disease risk: the Cache County study. BMC Genomics 17 Suppl 3, 438.

Hofman, A., Ott, A., Breteler, M.M., Bots, M.L., Slooter, A.J., van Harskamp, F., van Duijn, C.N., *et al.,* 1997. Atherosclerosis, apolipoprotein E, and prevalence of dementia and Alzheimer's disease in the Rotterdam Study. Lancet 349, 151–154.

Hohsfield, L.A., Humpel, C., 2015. Migration of blood cells to β-amyloid plaques in Alzheimer's disease. Exp. Gerontol. 65, 8–15.

Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.-C., Carrasquillo, M.M., Abraham, R., *et al.,* 2011. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat. Genet. 43, 429–435.

Holtzman, D.M., Bales, K.R., Tenkova, T., Fagan, A.M., Parsadanian, M., Sartorius, L.J., Mackey, B., *et al.,* 2000. Apolipoprotein E isoform-dependent amyloid deposition and neuritic degeneration in a mouse model of Alzheimer's disease. Proc. Natl. Acad. Sci. 97, 2892-2897.

Holtzman, D.M., Morris, J.C., Goate, A.M., 2011. Alzheimer's disease: The challenge of the second century. Sci. Transl. Med. 3, 77.

Hu, Y.Y.-S., Xin, J., Hu, Y.Y.-S., Zhang, L., Wang, J., 2017. Analyzing the genes related to Alzheimer's disease via a network and pathway-based approach. Alzheimers. Res. Ther. 9, 29.

Huang, K., Marcora, E., Pimenova, A.A., Di Narzo, A.F., Kapoor, M., Jin, S.C., Harari, O., *et al.,* 2017. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. Nat. Neurosci. 20, 1052–1061.

Huang, Y., Lu, W., Ali, D.W., Pelkey, K.A., Pitcher, G.M., Lu, Y.M., Aoto, H., *et al.,* 2001. CAKbeta/Pyk2 kinase is a signaling link for induction of long-term potentiation in CA1 hippocampus. Neuron 29, 485–496.

Huang, Y., Mucke, L., 2012. Alzheimer mechanisms and therapeutic strategies. Cell. 148, 1204-1222.

Humpel, C., 2011. Identifying and validating biomarkers for Alzheimer's disease. Trends Biotechnol. 29, 26–32.

Hung, S.-Y., Fu, W.-M., 2017. Drug candidates in clinical trials for Alzheimer's disease. J. Biomed. Sci. 24, 47.

Hutton, M., Lendon, C.L., Rizzu, P., Baker, M., Froelich, S., Houlden, H.H., Pickering-Brown, S., *et al.,* 1998. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. Nature. 393, 702-705.

Ihalainen, J., Sarajärvi, T., Rasmusson, D., Kemppainen, S., Keski-Rahkonen, P., Lehtonen, M., Banerjee, P.K., *et al.,* 2011. Effects of memantine and donepezil on cortical and hippocampal acetylcholine levels and object recognition memory in rats. Neuropharmacology. 5-6, 891-899.

International HapMap Consortium, Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., *et al.,* 2003. The International HapMap project. Nature 426, 789–796.

International Schizophrenia Consortium, Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., *et al.,* 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748–752.

Itan, Y., Casanova, J.-L., 2015. Can the impact of human genetic variations be predicted? Proc. Natl. Acad. Sci. U. S. A. 112, 11426–11427.

Ivanov, A., Romanovsky, A., 2006. Putative dual role of ephrin-Eph receptor interactions in inflammation. IUBMB Life (International Union Biochem. Mol. Biol. Life) 58, 389–394.

Iwamoto, N., Abe-Dohmae, S., Sato, R., Yokoyama, S., 2006. ABCA7 expression is regulated by cellular cholesterol through the SREBP2 pathway and associated with phagocytosis. J. Lipid Res. 47, 1915–1927.

Jack, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., *et al.,* 2013. Update on hypothetical model of dynamic biomarkers. Lancet Neurol. 12, 207–216.

Jacobsen, N., Bentzen, J., Meldgaard, M., Jakobsen, M.H., Fenger, M., Kauppinen, S., Skouv, J., 2002. LNA-enhanced detection of single nucleotide polymorphisms in the apolipoprotein E. Nucleic Acids Res. 30, e100.

Jan, A., Gokce, O., Luthi-Carter, R., Lashuel, H., 2008. The ratio of monomeric to aggregated forms of Abeta40 and Abeta42 is an important determinant of amyloid-beta aggregation, fibrillogenesis, and toxicity. J. Biol. Chem. 283, 28179-28189.

Jefferson, A.L., Cosentino, S. a, Ball, S.K., Bogdanoff, B., Leopold, N., Kaplan, E., Libon, D.J., 2002. Errors produced on the mini-mental state examination and neuropsychological test performance in Alzheimer's disease, ischemic vascular dementia, and Parkinson's disease. J. Neuropsychiatry Clin. Neurosci. 14, 311-320.

Jehle, A.W., Gardai, S.J., Li, S., Linsel-Nitschke, P., Morimoto, K., Janssen, W.J., Vandivier, R.W., *et al.,* 2006. ATP-binding cassette transporter A7 enhances phagocytosis of apoptotic cells and associated ERK signaling in macrophages. J. Cell Biol. 174, 547–556.

Jickling, G.C., Ander, B.P., Stamova, B., Zhan, X., Liu, D., Rothstein, L., Verro, P., *et al.,* 2013. RNA in blood is altered prior to hemorrhagic transformation in ischemic stroke. Ann. Neurol. 74, 232-240.

Jin, S.C., Benitez, B.A., Karch, C.M., Cooper, B., Skorupa, T., Carrell, D., Norton, J.B., *et al.,* 2014. Coding variants in TREM2 increase risk for Alzheimer's disease. Hum. Mol. Genet. 23, 5838–5846.

Jobst, K.A., Barnetson, L.P.D., Shepstone, B.J., 1998. Accurate prediction of histologically confirmed Alzheimer's disease and the differential diagnosis of dementia: The use of NINCDS-ADRDA and DSM-III-R criteria, SPECT, X-ray CT, and Apo E4 in medial temporal lobe dementias. Int. Psychogeriatrics. 10, 271-302.

Johnson, K.A., Fox, N.C., Sperling, R.A., Klunk, W.E., 2012. Brain imaging in Alzheimer disease. Cold Spring Harb. Perspect. Med. 2, e006213.

Jones, L., Holmans, P.A., Hamshere, M.L., Harold, D., Moskvina, V., Ivanov, D., Pocklington, A., *et al.,* 2010. Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer's disease. PLoS One 5, e13950.

Jonsson, T., Atwal, J.K., Steinberg, S., Snaedal, J., Jonsson, P. V., Bjornsson, S., Stefansson, H., *et al.,* 2012. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature 488, 96–99.

Jonsson, T., Stefansson, H., Steinberg, S., Jonsdottir, I., Jonsson, P. V, Snaedal, J., Bjornsson, S., *et al.,* 2013. Variant of TREM2 associated with the risk of Alzheimer's disease. N. Engl. J. Med. 368, 107–116.

Jorm, A.F., Jolley, D., 1998. The incidence of dementia: A meta-analysis. Neurology.

Jun, G., Naj, A.C., Beecham, G.W., Wang, L.-S., Buros, J., Gallins, P.J., Buxbaum, J.D., *et al.,* 2010. Meta-analysis confirms CR1, CLU, and PICALM as Alzheimer's disease risk loci and reveals interactions with APOE genotypes. Arch. Neurol. 67, 1473–1484.

Kaerst, L., Kuhlmann, A., Wedekind, D., Stoeck, K., Lange, P., Zerr, I., 2013. Cerebrospinal fluid biomarkers in Alzheimer's disease, vascular dementia and ischemic stroke patients: a critical analysis. J. Neurol. 260, 2722–2727.

Kajiho, H., Saito, K., Tsujita, K., Kontani, K., Araki, Y., Kurosu, H., Katada, T., 2003. RIN3: a novel Rab5 GEF interacting with amphiphysin II involved in the early endocytic pathway. J. Cell Sci. 116, 4159–4168.

Kakio, A., Nishimoto, S.I., Yanagisawa, K., Kozutsumi, Y., Matsuzaki, K., 2001. Cholesterol-dependent formation of GM1 ganglioside-bound amyloid β-protein, an endogenous seed for Alzheimer amyloid. J. Biol. Chem. 276, 24985–24990.

Karch, C.C.M., Goate, A.M.A., 2015. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. Biol. Psychiatry 77, 43–51.

Karch, C.C.M.C., Jeng, A.T.A., Nowotny, P., Cady, J., Cruchaga, C., Goate, A.M., 2012. Expression of novel Alzheimer's disease risk genes in control and Alzheimer's disease brains. PLoS One 7, e50976.

Karch, C.M., Ezerskiy, L.A., Bertelsen, S., Goate, A.M., Albert, M.S., Albin, R.L., Apostolova, L.G., *et al.,* 2016. Alzheimer's disease risk polymorphisms regulate gene expression in the ZCWPW1 and the CELF1 loci. PLoS One 11, 1–22.

Karran, E., Mercken, M., Strooper, B. De, 2011. The amyloid cascade hypothesis for Alzheimer's disease: An appraisal for the development of therapeutics. Nat. Rev. Drug Discov. 10, 698-712.

Kehoe, P.G., Wilcock, G.K., 2007. Is inhibition of the renin–angiotensin system a new treatment option for Alzheimer's disease? Lancet Neurol. 6, 373–378.

Khatri, P., Sirota, M., Butte, A.J., 2012. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput. Biol. 8, e1002375.

Kim, W.S., Li, H., Ruberu, K., Chan, S., Elliott, D.A., Low, J.K., Cheng, D., *et al.,* 2013. Deletion of Abca7 increases cerebral amyloid accumulation in the J20 mouse model of Alzheimer's disease. J. Neurosci. 33, 4387–4394.

Kimura, N., Yanagisawa, K., 2017. Traffic jam hypothesis: Relationship between endocytic dysfunction and Alzheimer's disease. Neurochem. Int. 186, 20249-30258.

Kircher, M., Kelso, J., 2010. High-throughput DNA sequencing - concepts and limitations. BioEssays 32, 524–536.

Kivipelto, M., Solomon, A., 2006. Cholesterol as a risk factor for Alzheimer's disease - epidemiological evidence. Acta Neurol. 185, 50-57.

Klein, C., Lohmann, K., Ziegler, A., 2012. The Promise and Limitations of Genome-wide Association Studies. JAMA 308, 1867.

Klein, W.L., 2013. Synaptotoxic amyloid-β oligomers: A molecular basis for the cause, diagnosis, and treatment of Alzheimer's disease? J. Alzheimer's Dis. 33, 49-65.

Knopman, D.S., DeKosky, S.T., Cummings, J.L., Chui, H., Corey-Bloom, J., Relkin, N., Small, G.W., *et al.,* 2007. Diagnosis of dementia: An evidence-based review. Contin. Lifelong Learn. Neurol.

Knopman, D.S., Parisi, J.E., Salviati, A., Floriach-Robert, M., Boeve, B.F., Ivnik, R.J., Smith, G.E., *et al.,* 2003. Neuropathology of Cognitively Normal Elderly. J. Neuropathol. Exp. Neurol. 62, 1087–1095.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., Mardis, E.R., 2013. The next-generation sequencing revolution and its impact on genomics. Cell 155, 27–38.

Kokmen, E., Beard, M.C., O'Brien, P., Kurland, L., 1996. Epidemiology of dementia in Rochester, Minnesota. Mayo Clin. Proc. 71, 275-282.

Koopman, K., Le Bastard, N., Martin, J.-J., Nagels, G., De Deyn, P.P., Engelborghs, S., 2009. Improved discrimination of autopsy-confirmed Alzheimer's disease (AD) from non-AD dementias using CSF P-tau181P. Neurochem. Int. 55, 214–218.

Kreisl, W.C., Lyoo, C.H., Liow, J.-S., Wei, M., Snow, J., Page, E., Jenko, K.J., *et al.,* 2016. 11C-PBR28 binding to translocator protein increases with progression of Alzheimer's disease. Neurobiol. Aging 44, 53–61.

Kullander, K., Klein, R., 2002. Mechanisms and functions of Eph and ephrin signalling. Nat. Rev. Mol. Cell Biol. 3, 475-486.

Kumar, P., Henikoff, S., Ng, P.C., 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4, 1073–1081.

Lacour, A., Espinosa, A., Louwersheimer, E., Heilmann, S., Hernández, I., Wolfsgruber, S., Fernández, V., *et al.,* 2016. Genome-wide significant risk factors for Alzheimer's disease: Role in progression to dementia due to Alzheimer's disease among subjects with mild cognitive impairment. Mol. Psychiatry 22, 1–8.

Lai, K.O., Ip, N.Y., 2009. Synapse development and plasticity: Roles of ephrin/Eph receptor signaling. Curr. Opin. Neurobiol. 19, 275-283.

Lambert, J.-C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., Combarros, O., *et al.,* 2009. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat. Genet. 41, 1094–1099.

Lambert, J., Ibrahim-Verbaas, C., Harold, D., Naj, A., 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature. 45, 1452-1458.

Lanoiselée, H.-M., Nicolas, G., Wallon, D., Rovelet-Lecrux, A., Lacour, M., Rousseau, S., Richard, A.-C., *et al.,* 2017. APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases. PLOS Med. 14, e1002270.

Larsson, M., Duffy, D.L.L., Zhu, G., Liu, J.Z.Z., Macgregor, S., McRae, A.F.F., Wright, M.J.J., *et al.,* 2011. GWAS findings for human iris patterns: Associations with variants in genes that influence normal neuronal pattern development. Am. J. Hum. Genet. 89, 334–343.

Launer, L.J., Ross, G.W., Petrovitch, H., Masaki, K., Foley, D., White, L.R., Havlik, R.J., 2000. Midlife blood pressure and dementia: The Honolulu–Asia aging study. Neurobiol. Aging 21, 49–55.

Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., Jung, S., *et al.,* 2014. Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell 159, 1312–1326.

Le Couteur, D.G., Hunter, S., Brayne, C., 2016. Solanezumab and the amyloid hypothesis for Alzheimer's disease. BMJ 355, 6771.

Lee, S., Abecasis, G.R., Boehnke, M., Lin, X., 2014. Rare-variant association analysis: Study designs and statistical tests, American Journal of Human Genetics. Elsevier. 95, 5-23.

Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., NHLBI GO Exome Sequencing Project—ESP Lung Project Team, N.G.E.S.P.L.P., *et al.,* 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. 91, 224–37.

Lee, S., Miropolsky, L., Wu, M., 2013. SKAT: SNP-set (Sequence) kernel association test. 0.82, URL: http//CRAN. R-project. org/package= SKAT.

Lehmann, M., Ghosh, P.M., Madison, C., Laforce, R., Corbetta-Rastelli, C., Weiner, M.W., Greicius, M.D., *et al.,* 2013. Diverging patterns of amyloid deposition and hypometabolism in clinical variants of probable Alzheimer's disease. Brain 136, 844–858.

Leidinger, P., Backes, C., Deutscher, S., Schmitt, K., Mueller, S.C., Frese, K., Haas, J., *et al.,* 2013. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol. 14, 78.

Leung, W.H., Tarasenko, T., Bolland, S., 2009. Differential roles for the inositol phosphatase SHIP in the regulation of macrophages and lymphocytes. Immunol. Res. 43, 243-251.

Levy-Lahad, E., Poorkaj, P., Wang, K., Fu, Y.H., Oshima, J., Mulligan, J., Schellenberg, G.D., 1996. Genomic structure and expression of STM2, the chromosome 1 familial Alzheimer's disease gene. Genomics. 34, 198-204.

Lewis, F., Schaffer, S.K., Sussex, J., O'Neill, P., Cockcroft, L., 2014. The trajectory of dementia in the UK – making a difference. Rep. Alzheimer's Res. UK by OHE Consult.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595.

Li, H., Homer, N., 2010. A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. 11, 473–483.

Li, H., Ruan, J., Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18, 1851–1858.

Li, X.-L., Hu, N., Tan, M.-S., Yu, J.-T., Tan, L., 2014. Behavioral and psychological symptoms in Alzheimer's disease. Biomed Res. Int. 2014, 927804.

Li, X., Wang, S., Sitaram, R.T., Andersson, C., Ljungberg, B., Li, A., 2013. Single Nucleotide Polymorphisms in the Wilms' Tumour Gene 1 in Clear Cell Renal Cell Carcinoma. PLoS One 8, e58396.

Li, Y., Rowland, C., Catanese, J., Morris, J., Lovestone, S., O'Donovan, M.C., Goate, A., *et al.,* 2008. SORL1 variants and risk of late-onset Alzheimer's disease. Neurobiol. Dis. 29, 293–296.

Li, Y., Willer, C., Sanna, S., Abecasis, G., 2009. Genotype imputation. Annu. Rev. Genomics Hum. Genet. 10, 387–406.

Liao, F., Yoon, H., Kim, J., 2017. Apolipoprotein E metabolism and functions in brain and its role in Alzheimer's disease. Curr. Opin. Lipidol. 28, 60–67.

Lichtenthaler, S.F., 2012. Alpha-Secretase Cleavage of the Amyloid Precursor Protein: Proteolysis Regulated by Signaling Pathways and Protein Trafficking. Curr. Alzheimer Res. 9, 165-177.

Lindquist, S., Schwartz, M., Batbayli, M., Waldemar, G., Nielsen, J., 2009. Genetic testing in familial AD and FTD: Mutation and phenotype spectrum in a Danish cohort. Clin. Genet. 76, 205–209.

Linnertz, C., Lutz, M.W., Ervin, J.F., Allen, J., Miller, N.R., Welsh-Bohmer, K.A., Roses, A.D., *et al.,* 2014. The genetic contributions of SNCA and LRRK2 genes to Lewy Body pathology in Alzheimer's disease. Hum. Mol. Genet. 23, 4814–4821.

Lipton, S.A., 2006. Paradigm shift in neuroprotection by NMDA receptor blockade: Memantine and beyond. Nat. Rev. Drug Discov. 5, 160-170.

Liu, C.-C.C.-C., Liu, C.-C.C.-C., Kanekiyo, T., Xu, H., Bu, G., 2013. Apolipoprotein E and Alzheimer disease: Risk, mechanisms and therapy. Nat. Rev. Neurol. 9, 106–118.

Liu, D., Niu, Z.-X., 2009. The structure, genetic polymorphisms, expression and biological functions of complement receptor type 1 (CR1/CD35). Immunopharmacol. Immunotoxicol. 31, 524–535.

Liu, Y., Yu, J.-T., Wang, H.-F., Hao, X.-K., Yang, Y.-F., Jiang, T., Zhu, X.-C., *et al.,* 2014. Association between NME8 locus polymorphism and cognitive decline, cerebrospinal fluid and neuroimaging biomarkers in Alzheimer's disease. PLoS One 9, e114777.

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S.G., Huntley, J., Ames, D., Ballard, C., *et al.,* 2017. Dementia prevention, intervention, and care. Lancet. 390, 2673-2734.

Lleo, A., Blesa, R., Gendre, J., Castellvi, M., Pastor, P., Queralt, R., Oliva, R., 2001. A novel presenilin 2 gene mutation (D439A) in a patient with early-onset Alzheimer's disease. Neurology. 57, 1926-1928.

Llorens, F., Schmitz, M., Ferrer, I., Zerr, I., 2016. CSF biomarkers in neurodegenerative and vascular dementias. Prog. Neurobiol. 138-140, 36-53.

Lord, J., Turton, J., Medway, C., Shi, H., Brown, K., Lowe, J., Mann, D., *et al.,* 2012. Next generation sequencing of CLU, PICALM and CR1: pitfalls and potential solutions. Int. J. Mol. Epidemiol. Genet. 3, 262–275.

Lott, I.T., Head, E., 2005. Alzheimer's disease and Down syndrome: Factors in pathogenesis. In: Neurobiology of Aging. 26, 383-389.

Lue, L.-F.F., Schmitz, C., Walker, D.G., Walker, A.D.G., 2015. What happens to microglial TREM2 in Alzheimer's disease: Immunoregulatory turned into immunopathogenic?, Neuroscience. 302, 138-150.

Mahdieh, N., Rabbani, B., 2013. An overview of mutation detection methods in genetic disorders. Iran. J. Pediatr. 23, 375-388.

Malik, M., Simpson, J.F., Parikh, I., Wilfred, B.R., Fardo, D.W., Nelson, P.T., Estus, S., 2013. CD33 Alzheimer's risk-altering polymorphism, CD33 expression and exon 2 splicing. J. Neurosci. 33, 13320–13325.

Manolio, T.A., 2010. Genome-wide association studies and assessment of the risk of disease. N. Engl. J. Med. 363, 166–176.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., *et al.,* 2009. Finding the missing heritability of complex diseases. Nature 461, 747–753.

Maquat, L.E., 2001. The power of point mutations. Nat. Genet. 27, 5–6.

Marden, J.R., Mayeda, E.R., Walter, S., Vivot, A., Tchetgen Tchetgen, E.J., Kawachi, I., Glymour, M.M., 2016. Using an Alzheimer's disease polygenic risk score to predict memory decline in black and white Americans over 14 years of follow-up. Alzheimer's Dis. Assoc. Disord. 30, 195–202.

Mardis, E.R., 2013. Next-Generation Sequencing Platforms. Annu. Rev. Anal. Chem. 6, 287–303.

Marioni, R.E., Campbell, A., Hagenaars, S.P., Nagy, R., Amador, C., Hayward, C., Porteous, D.J., *et al.,* 2017. Genetic Stratification to Identify Risk Groups for Alzheimer's Disease. J. Alzheimer's Dis. 57, 275–283.

Marsden, I.T., Laurie, L.S., Bamburg, J.R., 2011. Amyloid-β-induced amyloid-β secretion: A possible feed-forward mechanism in Alzheimer's disease. J. Alzheimer's Dis. 24, 681-691.

Martínez, A., Otal, R., Sieber, B.A., Ibáñez, C., Soriano, E., 2005. Disruption of ephrin-A/EphA binding alters synaptogenesis and neural connectivity in the hippocampus. Neuroscience. 135, 451-461.

Matarin, M., Salih, D.A., Yasvoina, M., Cummings, D.M., Guelfi, S., Liu, W., Nahaboo Solim, M.A., *et al.,* 2015. A genome-wide gene expression analysis and database in transgenic mice during development of amyloid or tau pathology. Cell Rep. 10, 633–644.

Mattsson, N., 2011. CSF biomarkers in neurodegenerative diseases. Clin. Chem. Lab. Med. 49, 345–352.

Mattsson, N., Zetterberg, H., Hansson, O., Andreasen, N., Parnetti, L., Jonsson, M., Herukka, S.-K., *et al.,* 2009. CSF Biomarkers and Incipient Alzheimer Disease in Patients With Mild Cognitive Impairment. JAMA 302, 385.

Mayeux, R., Tang, M.X., Jacobs, D.M., Manly, J., Bell, K., Merchant, C., Small, S.A., *et al.,* 1999. Plasma amyloid beta-peptide 1-42 and incipient Alzheimer's disease. Ann. Neurol. 46, 412–416.

McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.-B., Donnelly, P., 2014. Choice of transcripts and software has a large effect on variant annotation. Genome Med. 6, 26.

McGowan, E., Eriksen, J., Hutton, M., 2006. A decade of modeling Alzheimer's disease in transgenic mice. Trends Genet. 22, 281-289.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., *et al.,* 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., 1984. Clinical Diagnosis fo 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. 34, 939-944.

McKhann, G.G.M., Knopman, D.S.D., Chertkow, H., Hyman, B.B.T., Jack, C.R., Kawas, C.C.H., Klunk, W.E.W., *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. Alzheimer's Dement. 7, 1–7.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., *et al.,* 2016. The Ensembl Variant Effect Predictor. bioRxiv. 42374.

Medway, C., Morgan, K., 2014. Review: The genetics of Alzheimer's disease; putting flesh on the bones. Neuropathol. Appl. Neurobiol. 40, 97–105.

Mehta, D., Jackson, R., Paul, G., Shi, J., Sabbagh, M., 2017. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. Expert Opin. Investig. Drugs 26, 735–739.

Mendez, M.F., 2012. Early-onset Alzheimer's Disease: Nonamnestic Subtypes and Type 2 AD. Arch. Med. Res. 43, 677-685.

Metzner, A., Precht, C., Fehse, B., Fiedler, W., Stocking, C., Günther, A., Mayr, G.W., *et al.,* 2009. Reduced proliferation of CD34(+) cells from patients with acute myeloid leukemia after gene transfer of INPP5D. Gene Ther. 16, 570-573.

Meunier, B., Quaranta, M., Daviet, L., Hatzoglou, A., Leprince, C., 2009. The membrane-tubulating potential of amphiphysin 2/BIN1 is dependent on the microtubule-binding cytoplasmic linker protein 170 (CLIP-170). Eur. J. Cell Biol. 88, 91–102.

Michikawa, M., 2003. The role of cholesterol in pathogenesis of Alzheimer's disease: Dual metabolic interaction between amyloid β-protein and cholesterol. Mol. Neurobiol. 27, 1–12.

Minati, L., Edginton, T., Bruzzone, M.G., Giaccone, G., 2009. Reviews: Current concepts in Alzheimer's disease: A multidisciplinary review. Am. J. Alzheimers. Dis. Other Demen. 95–121.

Miosge, L.A., Field, M.A., Sontani, Y., Cho, V., Johnson, S., Palkova, A., Balakishnan, B., *et al.,* 2015. Comparison of predicted and actual consequences of missense mutations. Proc. Natl. Acad. Sci. U. S. A. 112, 5189-5198.

Mistur, R., Mosconi, L., Santi, S. De, Guzman, M., Li, Y., Tsui, W., de Leon, M.J., 2009. Current challenges for the early detection of Alzheimer's disease: Brain imaging and CSF studies. J. Clin. Neurol. 5, 153.

Moreau, K., Fleming, A., Imarisio, S., Lopez Ramirez, A., Mercer, J.L., Jimenez-Sanchez, M., Bento, C.F., *et al.,* 2014. PICALM modulates autophagy activity and tau accumulation. Nat. Commun. 5, 4998.

Moreno, H., Yu, E., Pigino, G., Hernandez, A.I., Kim, N., Moreira, J.E., Sugimori, M., *et al.,* 2009. Synaptic transmission block by presynaptic injection of oligomeric amyloid beta. Proc. Natl. Acad. Sci. U. S. A. 106, 5901-5906.

Morgan, K., 2011. The three new pathways leading to Alzheimer's disease. Neuropathol. Appl. Neurobiol. 37, 353–357.

Morris, G.P., Clark, I.A., Vissel, B., 2014. Inconsistencies and controversies surrounding the amyloid hypothesis of Alzheimer's disease. Acta Neuropathol. Commun. 2, 135.

Motter, R., Vigo-Pelfrey, C., Kholodenko, D., Barbour, R., Johnson-Wood, K., Galasko, D., Chang, L., *et al.,* 1995. Reduction of beta-amyloid peptide42 in the cerebrospinal fluid of patients with Alzheimer's disease. Ann. Neurol. 38, 643–648.

Mousseau, D.D., Banville, D., L'Abbé, D., Bouchard, P., Shen, S.H., 2000. PILRalpha, a novel immunoreceptor tyrosine-based inhibitory motif-bearing protein, recruits SHP-

1 upon tyrosine phosphorylation and is paired with the truncated counterpart PILRbeta. J. Biol. Chem. 275, 4467–4474.

Mungas, D., 1991. In-office mental status testing: A practical guide. Geriatrics.

Naj, A.C., Jun, G., Beecham, G.W., Wang, L.-S., Vardarajan, B.N., Buros, J., Gallins, P.J., *et al.,* 2011. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat. Genet. 43, 436–441.

Naj, A.C., Schellenberg, G.D., 2017. Genomic variants, genes, and pathways of Alzheimer's disease: An overview. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. 174, 5–26.

National Institute for Health and Clinical Excellence, 2011. Donepezil, galantamine, rivastigmine, and memantine for the treatment of Alzheimer's disease. Natl. Inst. Heal. Clin. Excell.

Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., *et al.,* 2011. Testing for an unusual distribution of rare variants. PLoS Genet. 7, e1001322.

Nelson, A.R., Sagare, A.P., Zlokovic, B. V, 2017. Role of clusterin in the brain vascular clearance of amyloid-β. Proc. Natl. Acad. Sci. U. S. A. 114, 8681–8682.

Neumann, M., Tolnay, M., Mackenzie, I.R., 2009. The molecular basis of frontotemporal dementia. Expert Rev Mol Med. 11, e23.

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., *et al.,* 2010. Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. 42, 30–35.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., *et al.,* 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272–276.

Nho, K., Horgusluoglu, E., Kim, S., Risacher, S.L., Kim, D., Foroud, T., Aisen, P.S., *et al.,* 2016. Integration of bioinformatics and imaging informatics for identifying rare PSEN1 variants in Alzheimer's disease. BMC Med. Genomics 9 Suppl 1, 30.

Nicolas, G., Charbonnier, C., Wallon, D., Quenez, O., Bellenguez, C., Grenier-Boley, B., Rousseau, S., *et al.,* 2016. SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease. Mol. Psychiatry 21, 831–836.

Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12, 443–451.

Nobili, F., Morbelli, S., 2010. [18F]FDG-PET as a biomarker for early Alzheimer's disease. Open Nucl. Med. J. 2, 46–52.

Nochlin, D., Bird, T.D., Nemens, E.J., Ball, M.J., Sumi, S.M., 1998. Amyloid angiopathy in a Volga German family with Alzheimer's disease and a presenilin-2 mutation (N141I). Ann. Neurol. 43, 131-135.

Novak, P., Schmidt, R., Kontsekova, E., Zilka, N., Kovacech, B., Skrabana, R., Vince-Kazmerova, Z., *et al.,* 2016. Safety and immunogenicity of the tau vaccine AADvac1 in patients with Alzheimer's disease: a randomised, double-blind, placebo-controlled, phase 1 trial. Lancet Neurol. 16, 123-134.

Nuchnoi, P., Nantakomol, D., Chumchua, V., Plabplueng, C., Isarankura-Na-Ayudhya, C., Prachayasittikul, V., 2011. The Identification of Functional Non-Synonymous SNP in Human ATPBinding Cassette (ABC), Subfamily Member 7 Gene: Application of Bioinformatics Tools in Biomedicine. J. Bioanal. Biomed. 3.

Nussbaum, J.M., Seward, M.E., Bloom, G.S., 2013. Alzheimer's disease: A tale of two prions. Prion.

Nuutinen, T., Suuronen, T., Kauppinen, A., Salminen, A., 2009. Clusterin: A forgotten player in Alzheimer's disease. Brain Res. Rev. 61, 89-104.

Nuytemans, K., Theuns, J., Cruts, M., Van Broeckhoven, C., 2010. Genetic etiology of Parkinson disease associated with mutations in the SNCA, PARK2, PINK1, PARK7, and LRRK2 genes: a mutation update. Hum. Mutat. 31, 763–780.

Office for National Statistics, 2017. Deaths Registered in England and Wales. Off. Natl. Stat.

Oliver, G.R., Hart, S.N., Klee, E.W., 2015. Bioinformatics for clinical next generation sequencing. Clin. Chem. 61, 124-135.

Olsson, B., Lautner, R., Andreasson, U., Öhrfelt, A., Portelius, E., Bjerke, M., Hölttä, M., *et al.,* 2016. CSF and blood biomarkers for the diagnosis of Alzheimer's disease: A systematic review and meta-analysis. Lancet Neurol. 15, 673–684.

Orth, M., Bellosta, S., 2012. Cholesterol: Its regulation and role in central nervous system disorders. Cholesterol 2012, 1–19.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., *et al.,* 2014. A survey of tools for variant analysis of next-generation genome sequencing data. Brief. Bioinform. 15, 256–278.

Paloneva, J., Manninen, T., Christman, G., Hovanes, K., Mandelin, J., Adolfsson, R., Bianchin, M., *et al.,* 2002. Mutations in two genes encoding different subunits of a receptor signaling complex result in an identical disease phenotype. Am. J. Hum. Genet. 71, 656–662.

Pandey, P., Avraham, S., Kumar, S., Nakazawa, A., Place, A., Ghanem, L., Rana, A., *et al.,* 1999. Activation of p38 mitogen-activated protein kinase by PYK2/related adhesion focal tyrosine kinase-dependent mechanism. J. Biol. Chem. 274, 10140–10144.

Pangman, V.C., Sloan, J., Guse, L., Brook, C., Reiss, B., German, P., Robinson, L., 2000. An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: implications for clinical practice. Appl. Nurs. Res. 13, 209–213.

Pant, S., Sharma, M., Patel, K., Caplan, S., Carr, C.M., Grant, B.D., 2009. AMPH-1/Amphiphysin/Bin1 functions with RME-1/Ehd1 in endocytic recycling. Nat. Cell Biol. 11, 1399–1410.

Panza, F., Solfrizzi, V., Seripa, D., Imbimbo, B.P., Lozupone, M., Santamato, A., Zecca, C., *et al.,* 2016. Tau-Centric Targets and Drugs in Clinical Development for the Treatment of Alzheimer's Disease. Biomed Res. Int. 2016, 3245935.

Patel, T., Brookes, K.J., Turton, J., Chaudhury, S., Guetta-Baranes, T., Guerreiro, R., Bras, J., *et al.,* 2018. Whole-exome sequencing of the BDR cohort: Evidence to support the role of the PILRA gene in Alzheimer's disease. Neuropathol. Appl. Neurobiol. 12452.

Patnala, R., Clements, J., Batra, J., 2013. Candidate gene association studies: a comprehensive guide to useful in silico tools. BMC Genet. 14, 39.

Peng, Q., Malhotra, S., Torchia, J.A., Kerr, W.G., Coggeshall, K.M., Humphrey, M.B., 2010. TREM2- and DAP12-dependent activation of PI3K requires DAP10 and is inhibited by SHIP1. Sci. Signal. 3, 38.

Peng, X., Nelson, E.S., Maiers, J.L., Demali, K. a, 2015. Family-based genome scan for age at onset of late onset Alzheimer's disease in whole exome sequencing data. Genes, Brain Behav. 14, 607–617.

Petersen, R.C., Roberts, R.O., Knopman, D.S., Boeve, B.F., Geda, Y.E., Ivnik, R.J., Smith, G.E., *et al.,* 2009. Mild cognitive impairment: Ten years later. Arch. Neurol. 66, 1447–1455.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: clinical characterization and outcome. Arch. Neurol. 56, 303–308.

Pfeifer, S.P., 2016. From next-generation resequencing reads to a high-quality variant data set. Heredity (Edinb). 118, 1–14.

Piaceri, I., Nacmias, B., Sorbi, S., 2013. Genetics of familial and sporadic Alzheimer's disease. Front. Biosci. (Elite Ed). 5, 167–177.

Pirinen, M., Donnelly, P., Spencer, C.C.A., 2012. Including known covariates can reduce power to detect genetic effects in case-control studies. Nat. Genet. 44, 848–851.

Pirooznia, M., Kramer, M., Parla, J., Goes, F.S., Potash, J.B., McCombie, W., Zandi, P.P., 2014. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum. Genomics 8, 14.

Plomin, R., Deary, I.J., 2015. Genetics and intelligence differences: five special findings. Mol. Psychiatry 20, 98–108.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., Siepel, A., 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20, 110–121.

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Auwera, G.A. Van der, Kling, D.E., *et al.,* 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 201178.

Postina, R., 2012. Activation of α-secretase cleavage. J. Neurochem. 120, 46-54.

Pottier, C., Hannequin, D., Coutant, S., Rovelet-Lecrux, A., Wallon, D., Rousseau, S., Legallic, S., *et al.,* 2012. High frequency of potentially pathogenic SORL1 mutations in autosomal dominant early-onset Alzheimer disease. Mol. Psychiatry 17, 875–879.

Pottier, C., Wallon, D., Rousseau, S., Rovelet-Lecrux, A., Richard, A.-C., Rollin-Sillaire, A., Frebourg, T., *et al.,* 2013. TREM2 R47H Variant as a Risk Factor for Early-Onset Alzheimer's Disease. J. Alzheimer's Dis. 35, 45–49.

Pozueta, J., Lefort, R., Shelanski, M.L., 2013. Synaptic changes in Alzheimer's disease and its models. Neuroscience 251, 51–65.

Prince, M., Wimo, A., Guerchet, M.E., 2015. World Alzheimer's Report 2015: The global impact of dementia, Alzheimer's Disease International, London.

Proitsi, P., Lee, S.H., Lunnon, K., Keohane, A., Powell, J., Troakes, C., Al-Sarraj, S., *et al.,* 2014. Alzheimer's disease susceptibility variants in the MS4A6A gene are associated with altered levels of MS4A6A expression in blood. Neurobiol. Aging 35, 279–290.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., *et al.,* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses 81, 559–575.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., *et al.,* 2012. A tale of three next generation sequencing platforms: comparison of

Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13, 341.

Raghavan, N., Tosto, G., 2017. Genetics of Alzheimer's Disease: the Importance of Polygenic and Epistatic Components. Curr. Neurol. Neurosci. Rep. 17, 78.

Raina, P., Santaguida, P., Ismaila, A., Patterson, C., Cowan, D., Levine, M., Booker, L., *et al.,* 2008. Annals of Internal Medicine Clinical Guidelines Effectiveness of Cholinesterase Inhibitors and Memantine for Treating Dementia: Evidence Review for a Clinical Practice Guideline. Ann. Intern. Med. 149, 358-359.

Ray, S., Britschgi, M., Herbert, C., Takeda-Uchimura, Y., Boxer, A., Blennow, K., Friedman, L.F., *et al.,* 2007. Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. Nat. Med. 13, 1359–1362.

Refai, F.S., Ng, S.H., Tan, E.K., 2015. Evaluating LRRK2 genetic variants with unclear pathogenicity. Biomed Res. Int. 2015, 1–6.

Reich, D.E., Lander, E.S., 2001. On the allelic spectrum of human disease. Trends Genet. 17, 502–510.

Reiman, E.M., Webster, J.A., Myers, A.J., Hardy, J., Dunckley, T., Zismann, V.L., Joshipura, K.D., *et al.,* 2007. GAB2 Alleles Modify Alzheimer's Risk in APOE ε4 Carriers. Neuron 54, 713–720.

Reitz, C., 2014. Genetic loci associated with Alzheimer's disease. Future Neurol. 9, 119-122.

Riancho, J., Vázquez-Higuera, J.L., Pozueta, A., Lage, C., Kazimierczak, M., Bravo, M., Calero, M., *et al.,* 2017. MicroRNA profile in patients with Alzheimer's disease: Analysis of miR-9-5p and miR-598 in raw and exosome enriched cerebrospinal fluid samples. J. Alzheimer's Dis. 57, 483–491.

Richens, J.L., Vere, K.-A., Light, R.A., Soria, D., Garibaldi, J., Smith, A.D., Warden, D., *et al.,* 2014. Practical detection of a definitive biomarker panel for Alzheimer's disease; comparisons between matched plasma and cerebrospinal fluid. Int. J. Mol. Epidemiol. Genet. 5, 53–70.

Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., *et al.,* 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. 43, 1066–1073.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26.

Rogaeva, E., Meng, Y., Lee, J.H., Gu, Y., Kawarai, T., Zou, F., Katayama, T., *et al.,* 2007. The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer's disease. Nat. Genet. 39, 168–177.

Rohn, T.T., 2013. The Triggering Receptor Expressed on Myeloid Cells 2: "TREM-ming" the Inflammatory Component Associated with Alzheimer's Disease. Oxid. Med. Cell. Longev. 2013, 1–8.

Rosén, C., Hansson, O., Blennow, K., Zetterberg, H., 2013. Fluid biomarkers in Alzheimer's disease - current concepts. Mol. Neurodegener. 8, 20.

Rosenthal, S.L., Kamboh, M.I., 2014. Late-onset Alzheimer's disease genes and the potentially implicated pathways. Curr. Genet. Med. Rep. 2, 85–101.

Ross, O., Soto-Ortolaza, A., Heckman, M., Aasly, J., Abahuni, N., Annesi, G., Bacon, J.A., *et al.,* 2011. Association of LRRK2 exonic variants with susceptibility to Parkinson's disease: A case–control study. Lancet Neurol. 10, 898–908.

Roy, S., Zhang, B., Lee, V.M.Y., Trojanowski, J.Q., 2005. Axonal transport defects: A common theme in neurodegenerative diseases. Acta Neuropathol. 109, 5-13.

Saad, M., Brkanac, Z., Wijsman, E.M., 2015. Family-based genome scan for age at onset of late-onset Alzheimer's disease in whole exome sequencing data. Genes, Brain Behav. 14, 607–617.

Saadipour, K., 2017. TREM1: A potential therapeutic target For Alzheimer's disease. Neurotox. Res. 32, 14–16.

Saiki, M., Baker, A., Williams-Gray, C.H., Foltynie, T., Goodman, R.S., Taylor, C.J., Compston, D.A.S., *et al.,* 2010. Association of the human leucocyte antigen region with susceptibility to Parkinson's disease. J. Neurol. Neurosurg. Psychiatry 81, 890–891.

Salamon, M., Millino, C., Raffaello, A., Mongillo, M., Sandri, C., Bean, C., Negrisolo, E., *et al.,* 2003. Human MYO18B, a novel unconventional myosin heavy chain expressed in striated muscles moves into the myonuclei upon differentiation. J. Mol. Biol. 326, 137–149.

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74, 5463–5467.

Sao, T., Yoshino, Y., Yamazaki, K., Ozaki, Y., Mori, Y., Ochi, S., Yoshida, T., *et al.,* 2017. MEF2C mRNA expression and cognitive function in Japanese patients with Alzheimer's disease. Psychiatry Clin. Neurosci. 72, 160-167.

Sassi, C., Nalls, M.A., Ridge, P.G., Gibbs, J.R., Ding, J., Lupton, M.K., Troakes, C., *et al.,* 2016. ABCA7 p.G215S as potential protective factor for Alzheimer's disease. Neurobiol. Aging 46, 235.e1-235.e9.

Sato-Harada, R., Okabe, S., Umeyama, T., Kanai, Y., Hirokawa, N., 1996. Microtubule-associated proteins regulate microtubule function as the track for intracellular membrane organelle transports. Cell Struct. Funct. 21, 283-295.

Sauna, Z.E., Kimchi-Sarfaty, C., 2011. Understanding the contribution of synonymous mutations to human disease. Nat. Rev. Genet. 12, 683–691.

Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C.A., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., *et al.,* 2011. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476, 214–219.

Sbrissa, D., Ikonomov, O.C., Fu, Z., Ijuin, T., Gruenberg, J., Takenawa, T., Shisheva, A., 2007. Core protein machinery for mammalian phosphatidylinositol 3,5-bisphosphate synthesis and turnover that regulates the progression of endosomal transport. J. Biol. Chem. 282, 23878–23891.

Scheltens, P., Fox, N., Barkhof, F., De Carli, C., 2002. Structural magnetic resonance imaging in the practical assessment of dementia: beyond exclusion. Lancet. Neurol. 1, 13–21.

Scherzer, C.R., Offe, K., Gearing, M., Rees, H.D., Fang, G., Heilman, C.J., Schaller, C., *et al.,* 2004. Loss of apolipoprotein E receptor LR11 in Alzheimer's disease. Arch. Neurol. 61, 1200–1205.

Schnetz-Boutaud, N.C., Hoffman, J., Coe, J.E., Murdock, D.G., Pericak-Vance, M.A., Haines, J.L., 2012. Identification and confirmation of an exonic splicing enhancer variation in exon 5 of the Alzheimer disease associated PICALM gene. Ann. Hum. Genet. 76, 448–453.

Schork, N.J., Murray, S.S., Frazer, K.A., Topol, E.J., 2009. Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev. 19, 212–219.

Schrijvers, E.M.C., Koudstaal, P.J., Hofman, A., Breteler, M.M.B., 2011. Plasma Clusterin and the Risk of Alzheimer Disease. JAMA 305, 1322.

Schwarz, J.M., Rödelsperger, C., Schuelke, M., Seelow, D., 2010. MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods. 7, 575-576.

Seshadri, S., Fitzpatrick, A.L., Ikram, M.A., DeStefano, A.L., Gudnason, V., Boada, M., Bis, J.C., *et al.,* 2010. Genome-wide analysis of genetic loci associated with Alzheimer's disease. J. Am. Med. Assoc. 303, 1832–1840.

Seshadri, S., Fitzpatrick, A.L., Ikram, M.A., DeStefano, A.L., Gudnason, V., Boada, M., Bis, J.C., *et al.,* 2010. Genome-wide analysis of genetic loci associated with Alzheimer disease. J. Am. Med. Assoc. 303, 1832–1840.

Seshadri, S., Wolf, P.A., 2007. Lifetime risk of stroke and dementia: current concepts, and estimates from the Framingham Study. Lancet Neurol. 6, 1106–1114.

Sevigny, J., Chiao, P., Bussière, T., Weinreb, P.H., Williams, L., Maier, M., Dunstan, R., *et al.,* 2016. The antibody aducanumab reduces Aβ plaques in Alzheimer's disease. Nature 537, 50–56.

Sharfe, N., Nikolic, M., Cimpeon, L., Van De Kratts, A., Freywald, A., Roifman, C.M., 2008. EphA and ephrin-A proteins regulate integrin-mediated T lymphocyte interactions. Mol. Immunol. 45, 1208-1220.

Sharma, N., Singh, A.N., 2016. Exploring biomarkers for Alzheimer's disease. J. Clin. Diagnostic Res. 10, 1-6.

Shen, J.N., Wang, D.S., Wang, R., 2012. The protection of acetylcholinesterase inhibitor on β-amyloid-induced the injury of neurite outgrowth via regulating axon guidance related genes expression in neuronal cells. Int. J. Clin. Exp. Pathol. 5, 900–913.

Shen, Y., Yang, L., Li, R., 2013. What does complement do in Alzheimer's disease? Old molecules with new insights. Transl. Neurodegener. 2, 21.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K., 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–11.

Shigemizu, D., Momozawa, Y., Abe, T., Morizono, T., Boroevich, K.A., Takata, S., Ashikawa, K., *et al.,* 2015. Performance comparison of four commercial human whole-exome capture platforms. Nat. Publ. Gr. 5, 12742.

Shulman, J.M., Imboywa, S., Giagtzoglou, N., Powers, M.P., Hu, Y., Devenport, D., Chipendo, P., *et al.,* 2014. Functional screening in Drosophila identifies Alzheimer's disease susceptibility genes and implicates tau-mediated mechanisms. Hum. Mol. Genet. 23, 870–877.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., Ponting, C.P., 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15, 121–132.

Sims, R., van der Lee, S.J., Naj, A.C., Bellenguez, C., Badarinarayan, N., Jakobsdottir, J., Kunkle, B.W., *et al.,* 2017. Rare coding variants in PLCG2, ABI3 and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. Nat. Genet. 49, 1373–1384.

Singh, R.K., Cooper, T.A., 2012. Pre-mRNA splicing in disease and therapeutics. Trends Mol. Med. 18, 472–482.

Sleegers, K., Cuyvers, E., Sleegers, K., 2016. Genetic variations underlying Alzheimer's disease: Evidence from genome-wide association studies and beyond. Lancet Neurol. 15, 857–868.

Sleegers, K., Lambert, J.C., Bertram, L., Cruts, M., Amouyel, P., Van Broeckhoven, C., 2010. The pursuit of susceptibility genes for Alzheimer's disease: Progress and prospects. Trends Genet. 26, 84-93.

Smith, A.R., Mill, J., Smith, R.G., Lunnon, K., 2016. Elucidating novel dysfunctional pathways in Alzheimer's disease by integrating loci identified in genetic and epigenetic studies. Neuroepigenetics 6, 32–50.

Spillantini, M.G., Goedert, M., 2013. Tau pathology and neurodegeneration. Lancet Neurol. 12, 609–622.

Spires-Jones, T.L., Hyman, B., 2014. The Intersection of Amyloid Beta and Tau at Synapses in Alzheimer's Disease. Neuron 82, 756–771.

St George-Hyslop, P.H., 2000. Molecular genetics of Alzheimer's disease. Biol. Psychiatry. 47, 183-199.

Steinberg, S., Stefansson, H., Jonsson, T., Johannsdottir, H., Ingason, A., Helgason, H., Sulem, P., *et al.,* 2015. Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. Nat. Genet. 47, 445–447.

Stitziel, N.O., Kiezun, A., Sunyaev, S., 2011. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol. 12, 227.

Streit, W.J., Xue, Q.S., 2012. Alzheimer's disease, neuroprotection, and CNS immunosenescence. Front. Pharmacol. 3, 138.

Stringer, S., Wray, N.R., Kahn, R.S., Derks, E.M., 2011. Underestimated Effect Sizes in GWAS: Fundamental Limitations of Single SNP Analysis for Dichotomous Phenotypes. PLoS One 6, e27964.

Suárez-Calvet, M., Kleinberger, G., Araque Caballero, M.Á., Brendel, M., Rominger, A., Alcolea, D., Fortea, J., *et al.,* 2016. sTREM2 cerebrospinal fluid levels are a potential biomarker for microglia activity in early-stage Alzheimer's disease and associate with neuronal injury markers. EMBO Mol. Med. 8, e201506123.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., *et al.,* 2017. The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. 45, 362–368.

Taddei, K., Fisher, C., Laws, S.M., Martins, G., Paton, A., Clarnette, R.M., Chung, C., *et al.,* 2002. Association between presenilin-1 Glu318Gly mutation and familial Alzheimer's disease in the Australian population. Mol. Psychiatry. 7, 776-781.

Tan, C.H., Fan, C.C., Mormino, E.C., Sugrue, L.P., Broce, I.J., Hess, C.P., Dillon, W.P., *et al.,* 2018. Polygenic hazard score: an enrichment marker for Alzheimer's associated amyloid and tau deposition. Acta Neuropathol. 135, 85–93.

Tan, M.-S., Yu, J.-T., Tan, L., 2013. Bridging integrator 1 (BIN1): Form, function, and Alzheimer's disease. Trends Mol. Med. 19, 594–603.

Tanaka, N., Abe-Dohmae, S., Iwamoto, N., Yokoyama, S., 2011. Roles of ATP-Binding Cassette Transporter A7 in cholesterol homeostasis and host defense system. J. Atheroscler. Thromb. 18, 274–281.

Tanzi, R.E., Bertram, L., 2001. New frontiers in Alzheimer's disease genetics. Neuron 32, 181–184.

Tanzi, R.E., Bertram, L., 2005. Twenty years of the Alzheimer's disease amyloid hypothesis: A genetic perspective. Cell 120, 545–555.

Thal, D.R., Rüb, U., Orantes, M., Braak, H., 2002. Phases of A beta-deposition in the human brain and its relevance for the development of AD. Neurology 58, 1791–1800.

Thambisetty, M., Lovestone, S., 2010. Blood-based biomarkers of Alzheimer's disease: challenging but feasible. Biomark. Med. 4, 65–79.

Thambisetty, M., Simmons, A., Velayudhan, L., Hye, A., Campbell, J., Zhang, Y., Wahlund, L.-O., *et al.,* 2010. Association of plasma clusterin concentration with severity, pathology, and progression in Alzheimer disease. Arch. Gen. Psychiatry 67, 739–748.

Thorvaldsdottir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192.

Toft, M., Sando, S.B., Melquist, S., Ross, O.A., White, L.R., Aasly, J.O., Farrer, M.J., 2005. LRRK2 mutations are not common in Alzheimer's disease. Mech. Ageing Dev. 126, 1201–1205.

Tomita, T., Maruyama, K., Saido, T.C., Kume, H., Shinozaki, K., Tokuhiro, S., Capell, A., *et al.,* 1997. The presenilin 2 mutation (N141I) linked to familial Alzheimer disease (Volga German families) increases the secretion of amyloid protein ending at the 42nd (or 43rd) residue. Proc. Natl. Acad. Sci. 4, 2025-2030.

Tosto, G., Reitz, C., 2013. Genome-wide Association Studies in Alzheimer's Disease: A Review. Curr. Neurol. Neurosci. Rep. 13, 381.

Treusch, S., Cyr, D.M., Lindquist, S., 2009. Amyloid deposits: Protection against toxic protein species? Cell Cycle 8, 1668–1674.

Turner, S., Armstrong, L.L., Bradford, Y., Carlson, C.S., Dana, C., Crenshaw, A.T., Andrade, M. De, *et al.,* 2011. Quality control procedures for genome wide association studies. Curr. Proc. Hum. Genet. 68, 1–24.

Vafadar-Isfahani, B., Ball, G., Coveney, C., Lemetre, C., Boocock, D., Minthon, L., Hansson, O., *et al.,* 2012. Identification of SPARC-like 1 Protein as Part of a Biomarker Panel for Alzheimer's Disease in Cerebrospinal Fluid. J. Alzheimer's Dis. 28, 625–636.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., *et al.,* 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinforma. 43, 1-33.

van Oijen, M., Hofman, A., Soares, H.D., Koudstaal, P.J., Breteler, M.M., 2006. Plasma Aβ1–40 and Aβ1–42 and the risk of dementia: A prospective case-cohort study. Lancet Neurol. 5, 655–660.

Vardarajan, B.N., Zhang, Y., Lee, J.H., Cheng, R., Bohm, C., Ghani, M., Reitz, C., *et al.,* 2015. Coding mutations in SORL1 and Alzheimer's disease. Ann. Neurol. 77, 215–227.

Villegas-Llerena, C., Phillips, A., Garcia-Reitboeck, P., Hardy, J., Pocock, J.M., 2016. Microglial genes regulating neuroinflammation in the progression of Alzheimer's disease. Curr. Opin. Neurobiol. 36, 74–81.

Villemagne, V.L., Fodero-Tavoletti, M.T., Pike, K.E., Cappai, R., Masters, C.L., Rowe, C.C., 2008. The ART of loss: Aβ imaging in the evaluation of Alzheimer's disease and other dementias. Mol. Neurobiol. 38, 1–15.

Viña, J., Lloret, A., 2010. Why Women Have More Alzheimer's Disease Than Men: Gender and Mitochondrial Toxicity of Amyloid-β Peptide. J. Alzheimer's Dis. 20, 527–533.

Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J., 2012. Five years of GWAS discovery, American Journal of Human Genetics. 90, 7-24.

Vivash, L., OBrien, T.J., 2016. Imaging microglial activation with TSPO PET: Lighting up neurologic diseases? J. Nucl. Med. 57, 165–168.

Vlassenko, A.G., Benzinger, T.L., Morris, J.C., 2012. PET amyloid-beta imaging in preclinical Alzheimer's disease. Biochim. Biophys. Acta - Mol. Basis Dis. 1822, 370-379.

Voyle, N., Hodges, A., 2017. Does genetic risk help to predict amyloid burden in a non-demented population? PLoS One.

Wang, K., Li, M., Hakonarson, H., 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164.

Warden, C.D., Adamson, A.W., Neuhausen, S.L., Wu, X., 2014. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. PeerJ. 2, e600.

Warner, P., 2013. Testing association with Fisher's Exact test. J. Fam. Plan. Reprod. Heal. Care 39, 281–284.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., Watson, M., 2015. Exome Sequencing: Current and Future Perspectives. 3, 5.

Weirauch, M., 2011. Gene co-expression networks for the analysis of DNA microarray data. Appl. Stat. Netw. Biol. Methods.

Wenk, G.L., 2003. Neuropathologic Changes in Alzheimer's Disease. Neuropathol. Chang. J Clin PsychiatryJ Clin Psychiatry 6464, 7–10.

Whitwell, J.L., 2010. Progression of atrophy in Alzheimer's disease and related disorders. Neurotox. Res. 18, 339-346.

Whitwell, J.L., Przybelski, S.A., Weigand, S.D., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack, C.R., 2007. 3D maps from multiple MRI illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to Alzheimer's disease. Brain 130, 1777–1786.

Wilkinson, D., 2012. A review of the effects of memantine on clinical progression in Alzheimer's disease. Int. J. Geriatr. Psychiatry. 27, 769-776.

Wilson, R.S., Barral, S., Lee, J.H., Leurgans, S.E., Foroud, T.M., Sweet, R. a, Graff-Radford, N., *et al.,* 2011. Heritability of different forms of memory in the Late Onset Alzheimer's Disease Family Study. J. Alzheimers. Dis. 23, 249-255.

Wray, N.R., Goddard, M.E., Visscher, P.M., 2008. Prediction of individual genetic risk of complex disease. Curr. Opin. Genet. Dev. 18, 257–263.

Wu, M.C.M., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X., 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89, 82–93.

Xiao, Q., Gil, S.-C., Yan, P., Wang, Y., Han, S., Gonzales, E., Perez, R., *et al.,* 2012. Role of phosphatidylinositol clathrin assembly lymphoid-myeloid leukemia (PICALM) in intracellular amyloid precursor protein (APP) processing and amyloid plaque pathogenesis. J. Biol. Chem. 287, 21279–21289.

Xu, W., Tan, L., Yu, J.-T., 2015. The Role of PICALM in Alzheimer's Disease. Mol. Neurobiol. 52, 399–413.

Xu, W.L., Atti, A.R., Gatz, M., Pedersen, N.L., Johansson, B., Fratiglioni, L., 2011. Midlife overweight and obesity increase late-life dementia risk: A population-based twin study. Neurology 76, 1568–1574.

Yaari, R., Corey-Bloom, J., 2007. Alzheimer's disease. Semin. Neurol. 27, 32-41.

Yu, J.-T., Tan, L., Hardy, J., 2014. Apolipoprotein E in Alzheimer's disease: An update. Annu. Rev. Neurosci. 37, 79–100.

Yu, L., Chibnik, L.B., Srivastava, G.P., Pochet, N., Yang, J., Xu, J., Kozubek, J., *et al.,* 2015. Association of Brain DNA methylation in SORL1, ABCA7, HLA-DRB5, SLC24A4, and BIN1 with pathological diagnosis of Alzheimer disease. JAMA Neurol. 72, 15–24.

Zanetti, O., Solerte, S.B., Cantoni, F., 2009. Life expectancy in Alzheimer's disease. Arch. Gerontol. Geriatr. 49, 237–243.

Zhang, J., Liu, Q., 2015. Cholesterol metabolism and homeostasis in the brain. Protein Cell 6, 254–264.

Zhao, Q.-F., Yu, J.-T., Tan, M.-S., Tan, L., 2015. ABCA7 in Alzheimer's disease. Mol. Neurobiol. 51, 1008–1016.

Zhao, Y., Bhattacharjee, S., Dua, P., Alexandrov, P.N., Lukiw, W.J., 2015. MicroRNA-based biomarkers and the diagnosis of Alzheimer's disease. Front. Neurol. 6, 162.

Zheng, H., Liu, C.-C., Atagi, Y., Chen, X.-F., Jia, L., Yang, L., He, W., *et al.,* 2016. Opposing roles of the triggering receptor expressed on myeloid cells 2 and triggering receptor expressed on myeloid cells-like transcript 2 in microglia activation. Neurobiol. Aging 42, 132–141.

Zhou, D.F., Wu, C.S., Qi, H., Fan, J.H., Sun, X.D., Como, P., Qiao, Y.L., *et al.,* 2006. Prevalence of dementia in rural China: Impact of age, gender and education. Acta Neurol. Scand. 114, 273-280.

Zhu, M., Zhao, S., 2007. Candidate gene identification approach: progress and challenges. Int. J. Biol. Sci. 3, 420–427.

Zimmer, E.R., Leuzy, A., Benedet, A.L., Breitner, J., Gauthier, S., Rosa-Neto, P., 2014. Tracking neuroinflammation in Alzheimer's disease: The role of Positron Emission Tomography imaging. J. Neuroinflammation 11, 120.

Zimprich, A., Biskup, S., Leitner, P., Lichtner, P., Farrer, M., Lincoln, S., Kachergus, J., *et al.,* 2004. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. Neuron 44, 601–607.

Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., *et al.,* 2014. Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. U. S. A. 111, 455-464.

Appendix

Appendix 7.1. Sample dropout at each quality check stage.

The number of samples of samples excluded at each stage of quality control and their criteria for removal. The number of samples passing QC at each step is also stated. These numbers have been classified by their phenotype.

***** 2 duplicate samples were not extracted again as tissue was already available.

Processing Step	Total Samples	N Samples Excluded	Passed QC	Criteria for Removal	Case	Control	Other
Initial Samples	382	2*	380	NA	161	73	146
DNA extracted	380	43	337	Concentration < 5µg or DIN < 6	161	73	146
Sent for whole-exome sequencing	337	45	292	Concentration < 3µg and smear on gel	158	68	111
Data Processing	292	3	288	Sex discrepancy	149	55	88
Final dataset	288	0	288	NA	148	55	85

Appendix 7.2. Burden2SetID.py

The script converts PLINK-SEQ (PSEQ) gene information output to the Set ID format required for the R SKAT package.

```
Converts PLINK-SEQ burden analysis output with SNPs and the RefSeq genes
that are contained in to the SetID file format for R SKAT package.

$ python Burden2SetID.py <input_pseq_file> <output_file_name>

#!/usr/bin/python
import sys

def hasNumbers(inputString):
	return any(char.isdigit() for char in inputString)

if len(sys.argv)!=3:
	print "Please provide: <input_pseq_file> <output_file_name>"
else:
	with open(sys.argv[1], 'r') as input, open(sys.argv[2], 'w') as
output:
		l=[]
		for line in input:
			if hasNumbers(line.split('\t')[2]):
				for i in l:
					output.write(i[:-
1].split('\t')[1]+'\t'+line.split('\t')[2].split(',')[1]+'\n')
				l=[]
			else:
				l.append(line)
```

Appendix 7.3. BurdenAnalysisSKATO.R

The script calculates the burden of variants within a gene together on a disease.

```
library(SKAT)
file.bed<-"bdr.bed"
file.bim<-"bdr.bim"
file.fam<-"bdr.fam"
file.setID<-"./SKATSetIDfile.txt"
file.cov<-"./AD.covar"

Generate_SSD_SetID(file.bed, file.bim, file.fam, file.setID,
 file.ssd, file.info)
SSD.INFO<-Open_SSD(file.ssd, file.info)

frame1=data.frame(matrix(NA, nrow=SSD.INFO$nSets, ncol=5))
names(frame1)<-c("P.value", "Test Type", "Q", "N Marker", "N Marker
 Test")

FAM_Cov<-Read_Plink_FAM_Cov(file.fam, file.cov, cov_header=TRUE,
 Is.binary=TRUE)

FAM_Cov<-Read_Plink_FAM_Cov(file.fam, file.cov, cov_header=TRUE,
 flag1=0, Is.binary=TRUE)
FAM_Cov[1:5,]
x1 = FAM_Cov$Sex
x2 = FAM_Cov$Cal_AAD
x3 = FAM_Cov$APOE_No_E4_alleles
x4 = FAM_Cov$APOE_No_E2_alleles
y<-FAM_Cov$Phenotype

#some covariates with weights
obj<-SKAT_Null_Model(y ~ x1 + x2 + x3 + x4, out_type="D")
for(i in 1:39) {
id<-i
try(z<-Get_Genotypes_SSD(SSD.INFO, id, is_ID=FALSE), silent=TRUE)
try(weights<-Get_Logistic_Weights(z, par1=0.07, par2=150),
 silent=TRUE)
try(out<-SKAT(z,obj,method="optimal.adj", weights.beta=c(0.5,0.5)),
 silent=TRUE)
try(frame1[i,1]<-out$p.value, silent=TRUE)
try(frame1[i,4]<-out$param$n.marker, silent=TRUE)
try(frame1[i,5]<-out$param$n.marker.test, silent=TRUE)
}

# Common and rare SNPs C=combined sum test
allsnps<-SKAT_CommonRare.SSD.All(SSD.INFO, obj, method="C",
 weights.beta.rare=c(1,25), weights.beta.common=c(0.5,0.5))

# Rare/Common SNPs only
allsnps_rare<-SKAT_CommonRare.SSD.All(SSD.INFO, obj, method="C",
 test.type="Rare.Only")
allsnps_common<-SKAT_CommonRare.SSD.All(SSD.INFO, obj, method="C",
 test.type="Common.Only")

Close_SSD()
```

Appendix 7.4. Chromosomal coordinates used for candidate gene analyses

The chromosomal location of all AD and familial neurodegenerative disease genes investigated for candidate gene analysis.

Chromosome	Start	End	Gene	Disease
21	27252861	27543446	*APP*	Familial AD
14	73603143	73690399	*PSEN1*	
1	227057885	227083804	*PSEN2*	
9	27546543	27573864	*C9orf72*	Frontotemporal dementia
3	87276413	87304698	*CHMP2B*	
16	31191431	31206192	*FUS*	
17	42422491	42430474	*GRN*	
12	40618813	40763087	*LRRK2*	Parkinson's disease
17	43971702	44105700	*MAPT*	
6	161768590	163148834	*PARK2*	
1	8021714	8045342	*PARK7*	
1	20959948	20978004	*PINK1*	
4	89724099	89838324	*SNCA*	
1	11072462	11085549	*TARDBP*	Amyotrophic lateral sclerosis
9	35056065	35072739	*VCP*	
20	4666797	4682235	*PRNP*	Prion disease
19	1040102	1065571	*ABCA7*	Late-onset Alzheimer's disease
2	127805599	127864903	*BIN1*	
6	47445482	47594999	*CD2AP*	
8	27454434	27472328	*CLU*	
1	207669473	207815110	*CR1*	
7	143088205	143105985	*EPHA1*	
11	59939080	59952139	*MS4A6A*	
11	85668214	85780923	*PICALM*	
20	54987168	55034396	*CASS4*	
11	47487489	47574792	*CELF1*	
14	53323989	53417856	*FERMT2*	
6	32546546	32557613	*HLA-DRB1/*	
6	32485151	32498006	*HLA-DRB5*	
2	233925036	233995881	*INPP5D*	
5	88014058	88199922	*MEF2C*	
7	37887930	37940003	*NME8*	
8	27168999	27316908	*PTK2B*	
14	92788925	92967825	*SLC24A4*	
14	92980125	93155339	*RIN3*	
11	121322912	121504471	*SORL1*	
7	99998449	100027037	*ZCWPW1*	
19	51728335	51743274	*CD33*	
19	45409039	45412650	*APOE4*	
6	41126244	41130924	*TREM2*	