Investigating polygenic risk scores in Alzheimer's disease

Sultan Raja Chaudhury, BSc, MRes

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

March 2023

Abstract

Background

Alzheimer's disease (AD), the most common cause of dementia, is one of the most studied diseases in the UK due to its impact on quality of life, symptoms of neurodegeneration, and burden on health and social care. AD is most common in the elderly, as age is a significant risk factor, but can also be found in young people. Diagnosis and treatment has developed over the years through improved therapies and screening methods, but a cure or definitive disease prevention has not been found.

Polygenic risk scoring is a relatively new approach, enabled by advancements in genotyping and sequencing technologies. They are used to quantify individual risk based on variants with observed effect within genes associated with the disease, calculated from genome-wide association studies of case v control data. Modelling is completed at various significance thresholds to identify the threshold at which the greatest predictive ability is achieved. Polygenic risk scoring has become increasingly popular as a tool for screening cohorts used in research, selecting candidates for trials, and further understanding complex genetic diseases and relationships between endophenotypes and disease status.

Methods

This project investigates polygenic risk scores in Alzheimer's disease, analysing late-onset AD (LOAD) cases, controls and undiagnosed samples recruited by the Brains for Dementia Research resource; mild cognitive impairment (MCI) cases recruited by the Inflammation, Cognition and Stress study; and sporadic early-onset AD (sEOAD) cases and controls recruited from research centres across the UK.

Genetic data was collected on either the NeuroX or NeuroChip array, quality controlled using recommended software and methodology, and imputed using the Michigan Imputation Server. Polygenic risk score (PRS) analysis was undertaken using PRSice-2 software, to calculate likelihood of developing AD and identify effective models for determining disease status in LOAD and sEOAD; the most predictive LOAD model was then used to predict likelihood in undiagnosed samples and MCI cases; a subset of genes expressed at the synapse were also analysed to understand their predictive ability in AD. This method utilised the most up-to-date analysis software and improved data sources to build on previously published work.

Results

The results for LOAD using updated PRS software identified a model with similar levels of predictive ability (AUPRC = 81.5%) as previously reported. Imputation identified additional variants within the best model threshold which implicate more genes in AD risk.

The sEOAD model using updated PRS software also confirmed a model with similar levels of prediction (AUROC = 73.0%) as previously reported. Analysis of imputed data identified a predictive model (AUROC = 72.9%) at a more significant p-value threshold and also implicated many more genes in AD risk.

PRS analysis of synaptic genes using updated PRS software showed greater levels of predictive ability for LOAD requiring fewer SNPs (AUPRC = 85.5%) than previously reported. A predictive model was also seen when analysing sEOAD (AUROC = 72.5%), and when combining LOAD and sEOAD cases (AUROC = 74.2%; AUPRC = 77.5%).

Utilisation of the best model for LOAD for predicting AD likelihood in MCI cases and undiagnosed samples successfully distributed individuals into tiers of risk; when most recent conversion status was cross-referenced with MCI samples, distribution was

seen across all risk tiers with most converters found to have moderate followed by high risk.

Conclusion

Identification of predictive models for LOAD and sEOAD which remain consistent with changes to methods, successful modelling of synaptic genes for LOAD and sEOAD, and moderate success in stratifying risk in undiagnosed samples highlight the utility of PRS in AD research. Continuous improvement in these analyses, through access to larger, more comprehensive datasets and advancements in software and methods, can enable greater accuracy and utility. This can ultimately establish polygenic risk scoring as a mechanism for understanding genetic risk for AD and other dementia sub-types, but further research in other complex diseases.

Publications

Lawingco, T., **Chaudhury, S.**, Brookes, K. J., Guetta-Baranes, T., Guerreiro, R., Bras, J., Hardy, J., Francis, P., Thomas, A., Belbin, O., and Morgan, K. (2020) Genetic variants in glutamate- Aβ- and tau-related pathways determine polygenic risk score for Alzheimer's disease. Neurobiology of Aging. 4580 (20), Pages 30391-30392.

Chaudhury, S.*, Brookes, K. J.*, Patel, T., Fallows, A., Guetta-Baranes, T., Turton, J. C., Guerreiro, R., Bras, J., Hardy, J., Francis, P. T., Croucher, R., Holmes, C., Morgan, K., and Thomas, A. J. (2019) Alzheimer's disease polygenic risk scoring as a predictor of conversion from mild-cognitive impairment. Translational Psychiatry. 9 (1), Page 167.

Brookes, K. J., McConnell, G., Williams, K., **Chaudhury, S.**, Madhan, G., Patel, T., Turley, C., Guetta-Baranes, T., Bras, J., Guerreiro, R., Hardy, J., Francis, P. T., and Morgan, K. (2018) Genotyping of the Alzheimer's disease genome-wide association study index single nucleotide polymorphisms in the Brains for Dementia Research cohort. Journal of Alzheimer's disease. 64 (2), Pages 355-362.

Patel, T., Brookes, K. J., Turton, J. C., **Chaudhury, S.**, Guetta-Baranes, T., Guerreiro, R., Bras, J., Hernandez, D., Singleton, A., Hardy, J., Francis, P. T., 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. 44 (5), Pages 506-521.

Chaudhury, S.*, Patel, T., Barber, I. S., Guetta-Baranes, T., Brookes, K. J., Chappell, S., Turton, J. C., Guerreiro, R., Bras, J., Hernandez, D. Singleton, A., Hardy, J., Mann, D., ARUK Consortium, and Morgan, K. (2018) Polygenic risk score in postmortem diagnosed sporadic early-onset Alzheimer's disease. Neurobiology of Aging. 62 (244), Page 244.

Contents

Abstract i
Background i
Methods i
Results ii
Conclusion iii
Publications iv
Contents v
Acknowledgements x
List of Tables xi
List of Figures xii
1 Introduction 1
1.1 Background 1
1.1.1 Symptoms 2
1.1.2 Pathogenesis 3
1.1.2.1 Amyloid β deposition 4
1.1.2.3 Tauopathy 5
1.1.2.3 Alternative hypotheses 6
1.1.3 Treatment 7
1.2 Aetiology and Comorbidities 9
1.2.1 Infection 9
1.2.2 Vascular diseases 9
1.2.3 Autoimmune diseases 10
1.2.4 Psychiatric disorders 10
1.2.5 Toxic disorders 10
1.2.6 Metabolic disorders 11
1.2.7 Poor nutrition 11
1.2.8 Lifestyle 12
1.3 Diagnosis 13
1.3.1 Biomarkers 14
1.3.1.1 Neuroimaging 14
1.3.1.2 Cerebrospinal fluid (CSF) 15
1.3.1.3 Plasma 16
1.3.2 Age 17

1.3.2.1 Late-onset AD 17
1.3.2.2 Early-onset familial AD 17
1.3.2.3 Sporadic early onset AD 18
1.3.2.4 Mild cognitive impairment 18
1.4 Genetics 19
1.4.1 Genome-wide association studies 19
1.4.2 Gene Clusters 20
1.4.2.1 *APOE*: high frequency, moderate risk 21
1.4.2.2 Causative: low frequency, high risk 21
1.4.2.3 Common: high frequency, low risk 22
1.4.2.4 Rare: low frequency, moderate risk 22
1.4.3 Complex Aetiology 23
1.5 Genotyping and Sequencing 23
1.5.1 Whole-genome sequencing 24
1.5.2 Whole-exome sequencing 24
1.5.3 Genotyping 25
1.5.3.1 NeuroX array 25
1.5.3.2 NeuroChip array 25
1.5.4 Imputation 26
1.6 Polygenic risk scoring 27
1.8.1 Literature 28
1.8.2 Association 29
1.8.3 Prediction 29
1.8.4 Specialisation 30
1.8.5 Software 31
1.7 Project Aims 31
2 Methods 33
2.1 Laboratory 33
2.1.1 Samples 33
2.1.2 DNA extraction 34
2.1.3 DNA quality control 35
2.1.4 TaqMan assays 36
2.1.5 Comparison of *APOE* genotyping methods 37
2.2 Bioinformatics 38

2.2.1 Genetic data 38
2.2.1.1 Variant call format 38
2.2.1.2 Genome Analysis Toolkit 38
2.2.1.3 PLINK 39
2.2.2 Sequencing pipeline 40
2.2.2.1 Initial QC 40
2.2.2.2 Alignment 41
2.2.2.3 Sorting 41
2.2.2.4 Read Groups 42
2.2.2.5 Validation 43
2.2.2.6 Mark duplicates 43
2.2.2.7 Realignment around InDels 44
2.2.2.8 Fix mate information 45
2.2.2.9 Base Quality Score Recalibration 46
2.2.2.10 Variant calling 47
2.2.2.11 Variant Quality Score Annotation and Recalibration 48
2.2.2.12 Functional annotation 50
2.2.2.13 Quality checks 51
2.2.3 Genotyping pipeline 52
2.2.3.1 Initial clustering 52
2.2.3.2 Manual re-clustering 53
2.2.3.3 Non-autosomal SNPs 56
2.2.3.4 Rare variants 58
2.2.3.5 Alignment 60
2.2.3.6 Sample call rate 60
2.2.3.6 Gender mismatch 61
2.2.3.7 Relatedness 61
2.2.3.8 Ancestry 62
2.2.3.9 Hardy-Weinberg equilibrium 63
2.2.3.10 Heterozygosity 64
2.2.4 Imputation pipeline 65
2.2.4.1 Pre-imputation quality control 65
2.2.4.2 Imputation 66
2.2.4.3 Quality control 67
2.2.5 Polygenic risk scoring 67

2.2.5.1 PRSice 68
3 Polygenic risk scoring of late-onset Alzheimer's disease 71
3.1 Perspective 71
3.2 Samples 72
3.3 Polygenic risk scoring 75
3.3.1 Best model selection 77
3.3.2 Association and risk prediction 82
3.4 SNPs 85
3.4.1 TaqMan GWAS SNPs 85
3.4.2 Imputation 88
3.5 Discussion 88
4 Polygenic risk scoring of sporadic early-onset Alzheimer's disease 93
4.1 Perspective 93
4.2 Samples 94
4.3 Polygenic risk scoring 96
4.3.1 Best model selection 98
4.3.2 Association and risk prediction 103
4.4 SNPs 106
4.4.1 Imputation 106
4.5 Discussion 107
5 Polygenic risk scoring of expression networks 111
5.1 Perspective 111
5.2 Samples 112
5.3 Polygenic risk scoring 114
5.3.1 Quality controls and gene set discovery 114
5.3.2 Best model selection 116
5.4 SNPs 122
5.5 Discussion 123
6 Polygenic risk scoring as a predictor of Alzheimer's disease 128
6.1 Perspective 128
6.2 Samples 129
6.3 Polygenic risk scoring 131
6.3.1 Summary 131
6.3.2 Risk prediction 135

6.4 Discussion 139
7 Discussion 142
7.1 Summary 142
7.2 Genotyping 144
7.3 PRSice 145
7.3.1 Decile scoring 147
7.4 Outcomes 148
7.4.1 Gene discovery 148
7.5 Limitations 150
7.6 Future work 150
7.6.1 BDR resource 150
7.6.2 NeuroBooster array 151
7.6.3 Targets for analysis 151
7.6.4 Machine learning and AI 151
7.7 Conclusion 153
References 154

Acknowledgements

I would like to thank Kevin Morgan for giving me this significant opportunity and the many incredible opportunities over the years, from supervising my research to encouraging my development, and supporting me until the very end. I would also like to thank Keeley Brookes and Tamar Guetta-Baranes for all their advice and support, for everything they taught me and every time they challenged me. Thank you to Tulsi for teaching me everything she knew, Jenny and Frankie for being with me at the beginning and Gaby for being there at the end. My time in the lab would not have been the same without the presence of Jim, as well as all the students whose projects contributed to my work and experience – especially Ted, Tom and Szymon.

I give thanks to the Neuroscience Support Group at the QMC and the University of Nottingham School of Life Sciences for funding this studentship but also enabling my aspirations. Additional thanks go to the University of Nottingham for their support during my period of sabbatical leave. This would not have been possible if not for all those individuals who participated in research - thank you to you and your families. This work was also only achieved through collaboration and advocacy by research groups, charities and organisations, most importantly the BDR.

To all the members of my family who have believed in me and prayed for my success – Jazaka'Allah Khair. To all my friends who kept me sane, singing, and happy, especially through some difficult times - thank you. To those especially dear, my companions for parts of this journey: Vivialyn, Mary-Claire, Barnaby, Jake, Ruth and Lucy – thank you.

Daisy, I would not have been able to do this without you.

List of Tables

1. APOE genotype calls by TaqMan assay....37
2. Demographics of BDR samples....73
3. Additional GWAS SNPs genotyped using TaqMan assays....74
4. PRSice modelling of LOAD with the NeuroChip....79
5. Modelling PRS of LOAD with covariates....81
6. Distribution of LOAD samples by decile scoring....83
7. Effect of GWAS SNPs on PRS of LOAD modelling....87
8. Demographics of sEOAD samples....95
9. PRSice modelling of sEOAD with the NeuroX....100
10. Modelling PRS of sEOAD with covariates....102
11. Distribution of sEOAD samples by decile scoring....104
12. Demographics of samples for synaptic PRS....113
13. Distribution of synaptic PRS samples by decile scoring....120
14. Demographics of predictive PRS samples....130
15. Distribution of predictive PRS samples by decile scoring....133
16. Distribution of samples by quartile scoring....136
17. Distribution of MCI conversion by quartile scoring....138
18. Genes identified in PRS analyses....149

List of Figures

1. Examples of manual re-clustering in GenomeStudio 55

2. Examples of manual re-clustering autosomal SNPs in GenomeStudio 59

3. Linkage disequilibrium heat map of PRSice clumping algorithm 70

4. Analysis pipeline for PRS of LOAD with the NeuroChip 76

5. Output figures of PRSice modelling of LOAD with the NeuroChip 78

6. Distribution of LOAD PRS by decile scoring 84

7. Analysis pipeline for PRS of sEOAD with the NeuroX 97

8. Output figures of PRS modelling of sEOAD with the NeuroX 99

9. Distribution of sEOAD PRS by decile scoring 105

10. Analysis pipeline for synaptic PRS of AD 115

11. High-resolution plot of synaptic PRS modelling of LOAD and sEOAD with the NeuroChip and NeuroX 118

12. Distribution of synaptic PRS by decile scoring 121

13. Analysis pipeline for predictive PRS of undiagnosed and MCI samples 132

14. Distribution of predictive PRS by decile scoring 134

1 Introduction

1.1 Background

Dementia categorises a group of neurodegenerative diseases estimated to affect nearly 1 million people in the UK and cost the NHS over £34 billion in health and social care by 2020 (Lewis et al., 2014). It is estimated nearly 70% of individuals over the age of 65 in the UK thought to have dementia are currently diagnosed, leaving almost one third without sufficient support and treatment. This is expected to continue to rise to nearly 1.6 million individuals and £94 billion by 2040 (Wittenberg et al., 2019). The loss of life due to the Covid-19 pandemic, especially in elderly populations, has skewed these projections; in recent years, a consistently lower prevalence of dementia has been observed amongst those over 65 in England (Public Health England, 2021).

Dementia is identified by symptoms of impairment through cognitive decline, changes in personality and behaviour, the loss of memory as well as language and visuospatial skills beyond the limits of normal ageing (Burns & Illife, 2009).

Whilst there is complex genetic component to dementia, their aetiology is diverse and can be caused or affected by historic infection; they can be concomitant with vascular, autoimmune and psychiatric disorders; toxic disorders caused by drug abuse or chemical exposure, and metabolic disorders can also increase risk (Budson, 2016; Gatz et al., 2006). Risk is also increased by environmental factors such as poor nutrition, educational attainment, and lack of physical or social activity, especially in old age (Budson, 2016; Corrada et al., 2010). These factors allow dementia to manifest as degenerative diseases, the most common of which is Alzheimer's disease (AD) (Prince et al., 2015).

AD is usually signified by the presence of amyloid beta (Aβ) plaques and hyperphosphorylated tau neurofibrillary tangles (NFT) in the brain; these biomarkers can be identified from cerebrospinal fluid (CSF) and confirmed by post-mortem analysis (Goedert & Spillantini, 2006; W. W. Li et al., 2020; G. McKhann et al., 1984).

1.1.1 Symptoms

AD prevalence increases with age, being a correlative risk factor; prevalence before the age of 65 is around 5% and reaches 7% at 65, increasing to 16% at the age of 85 (Corrada et al., 2010; Jorm & Jolley, 1998; Prince et al., 2015; Seshadri & Wolf, 2007).

Before the onset of AD symptoms, there is often a recognised prodromal stage of disease called mild cognitive impairment (MCI), where individuals may suffer mild amnesia but not necessarily present with other symptoms (Petersen et al., 1999). Once the symptoms of AD become apparent, individuals are likely to progressively worsen; the progression of the disease directly affects the severity of symptoms (Zanetti et al., 2009).

It is understood around 10-15% of individuals with MCI convert to AD each year with around one third progressing to AD over their lifetime (Adams et al., 2015; Hansson et al., 2010; Pyun et al., 2021; Rodríguez-Rodríguez et al., 2013). However, AD symptoms progress from mild to moderate to severe dementia; individuals may live decades with mild AD, however, with more severe cases of dementia, fatality is likely within years; the average time until death from disease onset is estimated as 8 years (X. L. Li et al., 2014; Zanetti et al., 2009).

Early stages of AD are classified by a struggle with short- and long-term memory recall, identified difficulty following conversation, disorganisation, and disorientation. When the disease progresses from mild to moderate severity, clinical symptoms worsen and individuals show signs of confusion: unable to recognise friends and

family, loss of spatial awareness with mood swings and other behavioural changes. The final stages of dementia are described as severe loss of memory, cognition, communication and the ability to walk, talk or swallow is almost completely lost; full-time care is required for the rest of their life (Förstl & Kurz, 1999).

1.1.2 Pathogenesis

In post-mortem examination of brains of patients with AD, the primary observations are of cortical shrinkage throughout many regions of the brain known as brain atrophy. Lesions are present in post-mortem tissue, caused by neuronal cell death which leads to the enlargement of the fluid-filled ventricles. Neuroimaging of patients with AD identifies correlation between the regions affected by AD pathology and the presentation of symptoms (Whitwell, 2010).

The first region affected by the disease is the hippocampus, which is involved in the formation and application of memories; the degeneration of the hippocampus leads to symptoms of amnesia. The temporal lobe, associated with communication and comprehension, is often the next region affected by disease and understood to cause the symptoms observed in mild to moderate AD (Wenk, 2003). Symptoms of moderate severity, such as changes in behaviour, mood swings, and further cognitive impairment, occur during the deterioration of the cingulate gyrus at a later stage of disease progression (Whitwell et al., 2007). The parietal lobe is associated with visuospatial skills and spatial awareness, and affected during the stages of moderate AD. The disease progresses to the occipital lobe alongside these regions and elevates symptoms of confusion and can lead to hallucinations and difficulties with recognition (Förstl & Kurz, 1999). The cerebellum is the region least affected by AD, which may be due to a protective nature of the region or the last region the disease reaches; before

the cerebellum exhibits atrophy, patients' exhibit symptoms of late-stage, severe AD as the brainstem begins to deteriorate (Thal et al., 2002).

The process by which brain atrophy occurs is complex and multifaceted. In post-mortem examination of patients with dementia, a diagnosis of AD as opposed to other dementia subtypes, is confirmed by the presence of pathological hallmarks of extracellular plaques of Aβ and NFTs of intracellular hyperphosphorylated tau protein (P-tau) (Goedert & Spillantini, 2006). The presence of these hallmarks, as well as symptoms of neurodegeneration are required for diagnosis of AD, however, they can also be present in individuals without symptoms of dementia and even absent in individuals with dementia (Villemagne et al., 2008; Whitwell, 2010). Why these hallmarks occur and their involvement in neuronal cell death has been studied since they were first identified in 1907 by Alois Alzheimer (G. McKhann et al., 1984).

1.1.2.1 Amyloid β deposition

The aggregates of Aβ found in diseased brain tissue are found to consist of higher-than-normal levels of an insoluble form of the peptide, produced during subsidiary processing of extracellular amyloid precursor protein (APP), $A\beta_{42}$ (Pozueta et al., 2013). The predominant, non-amyloidogenic pathway for APP processing involves cleavage by α-secretase resulting in a soluble $A\beta_{40}$ peptide (Hampel et al., 2021; Lichtenthaler, 2012; Postina, 2012). But when APP is initially cleaved by β-secretase instead of α-secretase, and then by γ-secretase peptides of varying lengths, $A\beta_{36\text{-}43}$, can result (Bird, 2018; Hampel et al., 2021; Marsden et al., 2011). The presence of Aβ in its many isoforms can be measured in CSF, blood plasma and brain interstitial fluid (Jan et al., 2008).

As hydrophobic, insoluble peptides of Aβ build up within the brain they begin to form oligomers, which are thought to disrupt synaptic transmission (Gu & Guo, 2013; Klein,

2013; Pozueta et al., 2013). The immune response, coordinated by microglia, acts to breakdown the Aβ oligomers and clear them from the central nervous system (CNS) in healthy individuals (Pozueta et al., 2013). Aβ plaques are hypothesised to form when the oligomers are not successfully cleared by the microglial response, which may lead to an apoptotic response and therefore loss of neurones (Hardy & Allsop, 1991; Karran et al., 2011).

1.1.2.3 Tauopathy

Microtubule-associated protein tau (*MAPT*) is the gene which codes for the soluble protein, tau, found within the axon and known to modulate stability of microtubules and therefore maintain cell structure and transport of organelles and nutrients within the neuron (Ballatore et al., 2007; Spillantini & Goedert, 2013). The protein can be translated into many isoforms with a varying number of microtubule binding domains, where a greater number of domains is associated with greater stabilisation (Ballatore et al., 2007). Stabilisation of microtubules by tau is also controlled by protein kinases, as hyperphosphorylation of tau reduces its microtubule binding affinity (Ballatore et al., 2007; Churcher, 2006).

When tau is hyperphosphorylated, the protein aggregates in the somatodendritic compartment instead of the axon and forms paired helical filaments which lead to NFT formation (Sato-Harada et al., 1996; Spillantini & Goedert, 2013). The NFTs disrupt the activity of tau and the effectiveness of nutrient transport along the axons, which eventually result in neuronal cell death (Goedert et al., 1995; Nussbaum et al., 2013; Roy et al., 2005; Spillantini & Goedert, 2013). Tauopathies are the collective term for the presence and pathological effect of tau NFTs in disease; tauopathies can occur in other dementia sub-types, like frontotemporal dementia (FTD), as well as

other neurological disorders like Down's syndrome (Hutton et al., 1998; Neumann et al., 2009; Spillantini & Goedert, 2013).

The spread of tau tangles across different regions of the brain is also concordant with disease progression (Spires-Jones & Hyman, 2014). The regions within which tau NFTs are found are classified by Braak staging, which acts as an indicator of the disease severity and progression (Braak et al., 2006; Braak & Braak, 1991). Tau first propagates in the transentorhinal layer of the brain, where the severity of neuronal alteration ranges between stage I or II; tau pathology in these Braak stages is either symptomatically unobserved or prodromal (Braak & Braak, 1991; Spires-Jones & Hyman, 2014). The tauopathy spreads to the limbic layer of the brain, and its severity is classed as stage III or IV; the damage caused by tau NFTs in this region occurs alongside mild and moderate dementia (Spires-Jones & Hyman, 2014). The final stages, V and VI, occur in the neocortex and are associated with the most disease severity (Knopman et al., 2003).

Tau pathology and Aβ plaque formation have been studied to identify correlation or relationship (Nussbaum et al., 2013). Evidence suggests Aβ may induce tau misfolding as well as observed positive correlation of the two disease hallmarks (Nussbaum et al., 2013). Like Aβ, tau NFTs are observed in elderly individuals without AD; individuals may have up to Braak stage III without disease (Braak et al., 2006).

1.1.2.3 Alternative hypotheses

Aβ and tau NFTs are considered hallmarks of AD, observed as the effect of disease on the brain (Goedert & Spillantini, 2006). Nevertheless, the presence of these hallmarks in non-diseased brains suggests other factors lead to disease symptoms; this could mean these hallmarks could mark the consequence of disease rather than the pathological cause (Villemagne et al., 2008). The alternative mechanisms by which AD

symptoms could be presenting were hypothesised and explored; these hypotheses did not identify causative factors for AD but had led to the definition of other dementia sub-types and their association with disease risk (Lipton, 2006).

The early hypotheses on symptom aetiology considered neurotransmitter deficiency, specifically Acetylcholine (ACh), due to low levels observed in diseased patients as well as benefits of acetylcholinesterase inhibitor (AChEI) treatments on symptom management (Francis et al., 1999). Excitotoxicity, specifically the overstimulation of the NMDA glutamate receptor, was considered as a potential cause of AD due to its ability to cause neurone cell death, as seen in neurological disorders like Multiple Sclerosis (Lipton, 2006).

1.1.3 Treatment

No cure has been established for AD, as the aetiology remains unclear (Zanetti et al., 2009). Investment in drug research and development has produced numerous unsuccessful drugs to combat AD symptoms (Cummings et al., 2014). Treatments have been often targeted the Aβ pathway with monoclonal antibodies and gamma secretase inhibitors, targeting P-tau with aggregation inhibitors, and drugs regulating neurotransmitter levels (Doody et al., 2014; Novak et al., 2017). Few drugs have passed clinical trials, principally those which provide temporary symptomatic relief at different stages of disease progression.

Donepezil, an AChEI, acts by preventing the breakdown of ACh, which is present in lower concentrations in patients with AD as a consequence of cholinergic neurone loss (National Institute for Health and Clinical Excellence, 2011; Wenk, 2003). The cognitive symptoms may reduce with greater levels of ACh being present at synapses, but the drug is most effective against mild to moderate AD and improves experience

for only around 40-70% of patients (National Institute for Health and Clinical Excellence, 2011).

Memantine is an example of an NMDA receptor antagonist, used to alleviate the cognitive symptoms of AD as a result of excessive levels of glutamate released through neuronal damage in AD patients (National Institute for Health and Clinical Excellence, 2011; Wilkinson, 2012). The drug acts to prevent the uptake of glutamate at active NMDA receptors which would usually lead to excitotoxicity and cell death; Memantine is recommended with moderate to severe AD (National Institute for Health and Clinical Excellence, 2011).

Recent advancements in immunotherapies have led to successful trials of lecanemab and donanemab, which target cerebral Aβ plaques at high levels of potency; where previous drugs have had limited clinical benefit, patients have been observed as amyloid-negative within 12 months of treatment (Alzheimer's Society, 2023; Ramanan & Day, 2023). Following approval, these drugs are expected to have significant impact on the rate of cognitive decline and may potentially be used for disease prevention (Alzheimer's Society, 2023).

1.2 Aetiology and Comorbidities

1.2.1 Infection

Many disorders caused by infection have been linked to dementia risk due to common symptoms and association studies, some infections can cause dementia in the absence of other risk factors (Budson, 2016).

Lyme neuroborreliosis, caused by Lyme disease bacterial infection, causes memory problems and symptoms affecting attention and learning; this can be identified as Lyme dementia, and is treatable (Blanc et al., 2014; Budson, 2016).

Human Immunodeficiency Virus (HIV) has an associated neurocognitive disorder due to the impacts on cognition, processing and executive function; cognitive assessments are carried out on patients with HIV and treated with combination antiretroviral therapies (Alzheimer's Society, 2021; Budson, 2016).

Studies into the presence of herpes simplex virus type 1 in postmortem AD patients, its correlation with other AD risk factors, and other herpesviruses implicated in dementia risk have led to wider studies into the association of microbial infection with dementia pathology (Abbott, 2020; Itzhaki et al., 1997; Seaksid & Wilcockid, 2020). This includes the most recent association of chronic periodontitis, often caused by Porphyromonas gingivalis, and Alzheimer's disease pathogenesis (Abbott, 2020; Chen et al., 2017).

1.2.2 Vascular disease

Mild vascular cognitive impairment (VCI) is determined when effects of cerebrovascular disease (CVD) advance pathogenesis and severity of dementia symptoms; when impairment is more significant or severe, this is considered vascular dementia (VaD) (Skrobot et al., 2018). When VaD is combined with dementia pathology this is identified as either mixed dementia unless the dementia subtype can

be identified, i.e., VCI-AD when in combination with AD (Budson, 2016; Skrobot et al., 2018).

Atherosclerosis has been found to be significantly associated with both AD and vascular dementia; this is determined to be involved with interactions with Apolipoprotein E (Andrews et al., 2021; Chandler et al., 2019; Hofman et al., 1997; Kivipelto & Solomon, 2006; Launer et al., 2000; Xu et al., 2011).

1.2.3 Autoimmune diseases

Type 2 diabetes is a common autoimmune disease with known association with dementia risk and symptoms, including memory loss (Budson, 2016). Diabetes leading to cerebrovascular disease or hypoglycaemia, which cause damage to the brain, has been linked to increased risk of dementia as well as a lower age at onset (Andrews et al., 2021; Budson, 2016).

1.2.4 Psychiatric disorders

The relationship between anxiety, depression and psychiatric disorders and dementia is understood to be concomitant, as anxiety and depression are often a symptom of dementia whilst a history of depression in early life can increase risk (Budson, 2016; Kokmen et al., 1996; Panza et al., 2010).

1.2.5 Toxic disorders

Alcoholism caused by long-term exposure to alcohol can cause dysfunction to the frontal lobes, limbic system and cerebellum of the brain; changes in personality and emotion are seen alongside aggressive and inappropriate behavioural traits (Budson, 2016; Matloff et al., 2020). Other comorbidities combined with alcoholism including schizophrenia, liver disease and head injury result in worsening of symptoms of dementia (Budson, 2016).

Whilst some metals, such as Zinc, Iron and Copper, are essential to biological functions, excess amounts can disrupt the equilibrium required for normal protein expression, similarly to exposure to neurotoxic metals like Lead and Aluminium (Dosunmu et al., 2007). Exposure to concentrated amounts of certain metals during the lifetime can have acute effects and increase risk and severity of dementia, observed by the high concentrations of Zinc and Iron in Aβ plaques due to their role in altering APP expression and promoting Aβ aggregation (Dosunmu et al., 2007).

1.2.6 Metabolic disorders

Metabolic disorders can affect attention and memory, presenting symptoms of dementia; there is also a distinction between this and symptoms of patients with the comorbidity of metabolic disorders and underlying dementia pathology (Andrews et al., 2020; Budson, 2016). This can be understood with Hypercalcemia, when symptoms of cognitive impairment are observed to be worse than usual and then return post-treatment, in later life as dementia symptoms (Budson, 2016).

The relationship between mitochondria, mitonuclear interactions and metabolic disease and how they link to dementia risk has been studied; whilst some correlation has been observed with specific mitochondrial haplogroups, results remain inconclusive and warrant further study (Andrews et al., 2020).

1.2.7 Poor nutrition

Dementia risk can be increased by a lifestyle of historic poor nutrition as well as deficiency of certain vitamins and minerals (Budson, 2016; Dosunmu et al., 2007; Xu et al., 2011). Nutrients are necessary for healthy function of the central nervous system through their impact on APP metabolism, Aβ isomer concentrations and neurodegeneration (Dosunmu et al., 2007). Cholesterol levels, which can be

controlled by diet, recorded at elevated levels in midlife are considered a significant risk factor for dementia in later life (Kivipelto & Solomon, 2006).

Many symptoms of vitamin B12 deficiency are common to those with dementia: memory loss, psychosis, irritability and personality changes (Budson, 2016). B12 deficiency can occur in patients who are vegetarian, with conditions which reduce absorption, and the elderly; treatment such as supplements may improve B12 levels and reduce symptoms (Budson, 2016). Long-term deficiency of vitamin B12, vitamin B6 or folic acid can lead Hyperhomocysteinemia, a risk factor for vascular disease and dementia (Dosunmu et al., 2007).

1.2.8 Lifestyle

Educational attainment and length have been studied to identify association with dementia, however results are mixed (Cobb et al., 1995; Zhou et al., 2006). Lower education can also be linked to low socioeconomic status which has been more often associated with poor health and disease (Cobb et al., 1995). Similarly, lower education when linked with nutritional deficiency, association with dementia risk can be observed (Zhou et al., 2006). Conversely, intellectual stimulation, occupation, and leisurely activity in later life as extensions to education have been associated with reduced risk of dementia and quality of life (Andel et al., 2005).

1.3 Diagnosis

Post-mortem pathological diagnosis is necessary to confirm disease status and dementia subtype; however, clinical diagnosis is mostly concordant with this, achieving 77% accuracy where misclassification may be due to the variability in dementia subtypes (Beach et al., 2012). Diagnosis has been accomplished through various methods; when distinguishing dementia from other neurological disorders and later diagnosing the sub-type, mental and physical assessments are recommended alongside testing for biomarkers and conducting imaging (Blacker et al., 1994; Dubois et al., 2007; Jack et al., 2016; G. McKhann et al., 1984; G. M. McKhann et al., 2011; Whitwell et al., 2007).

In the UK, dementia is identified by a General Practitioner and diagnosed by a specialist. Mental assessments are used to identify cognitive decline, assessing short and long-term memory, ability to concentrate and attention span, language and communication skills, and temporal and spatial awareness; tests are often repeated over time to identify if abilities are progressively worsening (Dubois et al., 2007). Recommended screening tools include the Mini-Mental State Exam (MMSE) and the Montreal Cognitive Assessment (MoCA), which are both short, 30-point tests where score-bands differentiate between those who are cognitively healthy, in cognitive decline or severely impaired due to dementia (Folstein et al., 1975; Mungas, 1991; Pangman et al., 2000; Yaari & Corey-Bloom, 2007). As results can be influenced by age, educational background, and other impairments, further testing is required to rule out a dementia diagnosis (Jefferson et al., 2002; Pangman et al., 2000).

Blood testing is currently arranged to check liver, kidney and thyroid function, vitamin and folate levels, and whether the patient may have diabetes; other investigations are considered to rule out infections. Computerised Tomography (CT) scans can be used

to search for signs of stroke or brain tumours, in order to rule them out, but are not adequate for assessing brain structure (Johnson et al., 2012).

An Electroencephalogram (EEG) is used to rule out epilepsy as a cause of dementia symptoms (Johnson et al., 2012). Magnetic Resonance Imaging (MRI) scans are more effective at imaging the brain structure and are often used to define dementia sub-type; damage to blood vessels is an indicator of VAD, and shrinkage in the frontal as well as the temporal lobe is an indicator of FTD (Jobst et al., 1998; Johnson et al., 2012). Other imaging assessments include Positron Emission Tomography (PET) scans, which when incorporating radiotracers, can improve diagnosis by observing pathological phenotypes like Aβ plaque deposition, and its levels within the brain (Johnson et al., 2012; G. M. McKhann et al., 2011; Minati et al., 2009; Vlassenko et al., 2012).

1.3.1 Biomarkers

Clinical diagnosis usually occurs after the onset of symptoms; however, the pathological changes which occur within the AD brain may begin years to decades beforehand, meaning treatments to modify or prevent degeneration are being administered too late in the disease process (Mattsson, 2011). Preclinical dementia diagnosis can be achieved through testing for the presence or concentration of biomarkers associated with AD (Ewers et al., 2015; Humpel, 2011; Mattsson, 2011).

1.3.1.1 Neuroimaging

MRI and CT scanning, used in diagnosis, can also be utilised preclinically when measuring brain volume, cortical atrophy and NFTs; brain structural and functional information can be collected, and changes can be observed over time through repeat scanning (Scheltens et al., 2002). PET scanning with the radioligands Pittsburgh compound (PiB-PET) and 18F-2-deoxy-2-fluoro-D-glucose (FDG-PET) can be used to

identify Aβ plaque and glucose levels respectively (Mistur et al., 2009; Nobili & Morbelli, 2010). The use of these neuroimaging biomarkers has resulted in success tracking and predicting AD conversion amongst those with MCI, with results quoting around 75% accuracy using PET alone (Herholz et al., 2011; Nobili & Morbelli, 2010). Using imaging as a tool to observe preclinical pathology could be considered a successful, non-invasive utility; however, comprehensive regular neuroimaging would be expensive and could not be justified as a preclinical procedure available to either the entire population or even individuals who could be more at risk.

1.3.1.2 Cerebrospinal fluid (CSF)

CSF transports nutrients, hormones, and waste products around the CNS. A sample of CSF is extracted from patients by an invasive lumbar puncture but can provide details of proteins and their levels at the point of extraction; model CSF biomarkers are identified as AD-related proteins which vary in concentration when comparing between samples from individuals with dementia, without dementia (controls) and with MCI. Established AD-related proteins which can be used as CSF biomarkers include $A\beta_{42}$, which is significantly lower in AD than controls, indicative of $A\beta_{42}$ trapped in insoluble plaques and therefore not being cleared or circulated; total tau (t-tau) levels, which tend to increase with age but are significantly higher in dementia patients than controls and more likely to be higher in those with MCI who convert to AD than non-converters; P-tau levels are generally similar to t-tau but more predictive in regards to sensitivity and specificity when discerning between AD and other dementias (Blennow & Hampel, 2003; Forlenza et al., 2015; Humpel, 2011; Llorens et al., 2016; Motter et al., 1995).

Inflammatory regulation and microglial activation could be measured by levels of soluble Triggering receptor expressed on myeloid cells 2 (TREM2), a protein expressed

on microglia which correlates with t-tau and P-tau, and therefore a biomarker for inflammation during early symptomatic phases of AD (Suárez-Calvet et al., 2016). As several biomarkers can be tested from one sample, CSF sampling may be a good source for comprehensive pre-clinical diagnostic information (Mattsson et al., 2009). The current method of sampling however is intrusive and may not be suitable for all patients or regular sampling.

1.3.1.3 Plasma

The most cost effective and reasonably non-invasive source of biomarkers may be in blood plasma collection. Plasma carries tens of thousands of proteins, with most in low, immeasurable concentrations; the proteins present in plasma are produced by peripheral tissue and do not cross the blood-brain barrier, therefore are not useful for observing activity and expression of AD-related proteins in the brain (Hansson et al., 2010; Mehta et al., 2000; van Oijen et al., 2006).

Although plasma levels of proteins like Aβ and tau produce conflicting results for risk prediction, platforms built to measure a combination of plasma biomarkers with other markers may show high predictive accuracy between AD and controls (Thambisetty & Lovestone, 2010). Candidate biomarkers have been identified and tested based on pathways associated with AD pathology, such as total serum cholesterol and oxysterols for cholesterol metabolism and F_2-isoprostanes for oxidative stress; predictive abilities of 85-90% have been recorded when measuring plasma concentrations of proteins, such as interleukins, associated with inflammation (Ray et al., 2007; Thambisetty & Lovestone, 2010). Most recently, Neurofilament light chain (NfL) has been identified as a promising biomarker based on its ability to distinguish between AD and MCI cases and controls (Giacomucci et al., 2022).

1.3.2 Age

1.3.2.1 Late-onset AD

AD most commonly occurs in the elderly, with symptoms first presenting at the age of 65 or later; late-onset Alzheimer's disease (LOAD) represents most individuals who suffer from dementia (Piaceri et al., 2013). Through twin studies, it is apparent there is a genetic component to LOAD as well as being affected by environmental factors (Gatz et al., 2006; Wilson et al., 2011). AD can occur in patients before 65, presenting symptoms in early life or before 65, the symptoms are often more severe, and the progression can be more accelerated (Mendez, 2012; Minati et al., 2009). The identification of individuals with early, severe symptoms of AD resulted in many studies and possible explanations for the aetiology of AD (Antonell et al., 2013; Piaceri et al., 2013).

1.3.2.2 Early-onset familial AD

Amongst patients of early onset AD, there are two further sub-types (Ertekin-Taner, 2007). A small proportion of individuals with early-onset AD were found to have genetic variants in one of few genes, which were recognised for their involvement in APP processing (Bertram et al., 2010; R. Guerreiro & Hardy, 2014). The variants are present within the genes of Amyloid precursor protein (APP), Presenilin 1 (PSEN1) and Presenilin 2 (PSEN2); the genes are autosomal-dominant due to the variants showing high penetrance in family studies (M. Choi et al., 2009). Many variants have been identified within these genes, mostly within coding regions; patients with early-onset symptoms of AD are screened for variants in these genes, the presence of which leads to a diagnosis of early-onset familial Alzheimer's disease (EOFAD) (Hampel & Lista, 2012). Although rare, familial variants can also be found in LOAD patients, suggesting some variants may also have lower penetrance (Antonell et al., 2013).

1.3.2.3 Sporadic early onset AD

The remaining patients with early-onset symptoms without the autosomal-dominant familial aspect of EOFAD are classified as sporadic early-onset Alzheimer's disease (sEOAD) (Piaceri et al., 2013). The hypothetical cause of sEOAD is thought to be a result of a greater concentration of or more penetrant set of gene variants than observed in LOAD, which leads to an earlier onset of symptoms but otherwise identical disease traits (Piaceri et al., 2013).

1.3.2.4 Mild cognitive impairment

Where only minor symptoms of cognitive decline present in individuals before or around the age of 65, a diagnosis of mild cognitive impairment (MCI) is given (Petersen et al., 1999). Amongst this cohort, there is the likelihood individuals will either develop dementia in later life, with MCI being a prodrome of dementia; continue to have symptoms like amnesia throughout life; or revert to healthy levels of cognition (Adams et al., 2015; Hansson et al., 2010; Pyun et al., 2021; Rodríguez-Rodríguez et al., 2013).

1.4 Genetics

Genetic research involves the study of genes and gene variants and their associations with disease phenotypes. Through twin, family and cohort studies, the genetic architecture of diseases can be defined, and heritability can be estimated. Genetic studies have been undertaken in AD research since the discovery and further study of the pathological pathways which may be involved in AD (Tanzi & Bertram, 2005). The various methods of gene variant discovery have resulted in some understanding of the complex genetic nature of AD, its sub-types and other dementia sub-types (Gatz et al., 2006).

1.4.1 Genome-wide association studies

Gene discovery was revolutionised by the utility of genome-wide association studies (GWAS) (Bertram et al., 2010; Bush & Moore, 2012; Manolio, 2010). The study takes the genotype data of a cohort of individuals with (cases) and without (controls) a disease phenotype and compares the frequency of each genetic variant between each group and calculates the significance of the association of the genotype with the phenotype (Anderson et al., 2010; Clarke et al., 2011). The more often a variant appears amongst cases than controls the more likely it is to be associated with increasing likelihood of developing the disease (Anderson et al., 2010; Manolio, 2010). The more individuals involved in the study, the more accurate the effect the variant has can be calculated, and the more power the study has finding variants of significant, genome-wide impact (Clarke et al., 2011). This concept was recognised and used across a spectrum of diseases, aiding discovery and confirmation of many significant genetic loci associated with a disease phenotype (Manolio, 2010).

A widely used GWAS completed on LOAD using genotype data from 35,274 clinically diagnosed and later post-mortem pathologically confirmed cases of AD and 59,163

controls of Non-Hispanic White, Caucasian ethnicity (J. C. Lambert et al., 2013). The results identified 25 independent genetic loci with association to AD at genome-wide level of significance, 20 of which were previously found in previous GWAS analyses (Beecham et al., 2009; Bertram et al., 2007; Coon et al., 2007; Grupe et al., 2007; Harold et al., 2009; Hollingworth et al., 2011; J. C. Lambert et al., 2009; Y. Li et al., 2008; Reiman et al., 2007; Seshadri et al., 2010). The most recent GWAS identifies 75 risk loci associated with AD and related dementias, requiring further study (Bellenguez et al., 2022).

1.4.2 Gene Clusters

Results of GWAS are often shared in the form of a Manhattan plot; the figure identifies the genetic loci along the x-axis, and a logarithmic scale of significance (p-value) on the y-axis (J. C. Lambert et al., 2013). There is usually a line which outlines the significance level which needs to be achieved for the variant to be considered genome-wide significant, $p<5\times10^{-8}$ (J. C. Lambert et al., 2013). GWAS results can also be shared in a plot of minor allele frequency on the x-axis and disease risk on the y-axis; this is useful for understanding the genetic architecture of diseases.

Genetic research to date has aided in categorising genetic loci associated with AD as causative, common and rare; the Apolipoprotein E (*APOE*) gene locus falls outside of this categorisation (Bush & Moore, 2012; Manolio et al., 2009). The associated risk observed in *APOE* variants is similar to that of rare variants, yet the risk allele is common. The hypothesis for this considers individuals who harbour the *APOE* ε4 allele and common variants to have fecundity and therefore manage to survive and reproduce, passing on the variant to the next generation and maintaining the allele frequency. As the disease presents in later life; there has been no selective pressure against harbouring it.

1.4.2.1 *APOE*: high frequency, moderate risk

One of the first genes to be associated with AD, before the use of GWAS was *APOE*; the gene was first identified in linkage studies and its presence and role in the Aβ pathway is still being determined (Corder et al., 1993; Liao et al., 2017). This exists in three isoforms (i.e., E2, E3, and E4), where individuals carry two copies; the identity of which alleles an individual harbours can be determined from the genotypes of two SNPs (i.e., rs7412 and rs429358) (Farrer et al., 1997). The specific *APOE* allele associated with increased risk produces the E4 isoform of the protein, where the presence of one copy increases AD risk 4-fold and individuals who harbour two have a 12 to 16-fold risk of developing AD compared to individuals with the most common genotype, ε3ε3 (Bertram et al., 2007; Corder et al., 1993, 1994). The ε2 allele is associated with a reduced risk of developing AD as well as a delayed age at onset in those who do develop AD (Corder et al., 1994; Farrer et al., 1997).

The pathological role of *APOE* in AD has been studied based on its site of expression, and role and interaction with amyloid β (Corder et al., 1994; Holtzman et al., 2000; Liao et al., 2017; Sleegers et al., 2010). The study of *APOE* and its association continues, as further evidence highlights the entire locus site of interest, with suggestions implicating the gene to be the most significant association to AD whilst others consider the *APOE* gene to be linked to and inherited with another nearby loci, which may be the true causative source of disease risk observed with *APOE* ε4; both *TOMM40* and *APOC1* are situated within the same locus as *APOE* and have variants with association to AD risk (Kulminski et al., 2021; Ware et al., 2020).

1.4.2.2 Causative: low frequency, high risk

Variants with the highest risks are clustered together and usually considered to have such high risk it is likely an individual carrying this variant would develop the disease

(Hampel & Lista, 2012). Examples of variants in AD which may have this effect include the three genes which lead to EOFAD (i.e., *APP*, *PSEN1* and *PSEN2*) as well as a rare variant found in the *TREM2* gene (Abduljaleel et al., 2014; Bekris et al., 2010; De Strooper et al., 1998; Desikan et al., 2015; Ertekin-Taner, 2007; R. Guerreiro et al., 2013; R. J. Guerreiro et al., 2013; Hutton et al., 1998; S. C. Jin et al., 2014; Jonsson et al., 2013; Levy-Lahad et al., 1996; Lleó et al., 2001; Pottier et al., 2013; Taddei et al., 2002; Tomita et al., 1997).

1.4.2.3 Common: high frequency, low risk

Most variants identified as significant by GWAS happen to be common, present in both cases and controls, but more frequently in cases (Bush & Moore, 2012; Manolio, 2010; Manolio et al., 2009; Visscher et al., 2012). This localisation of these variants is usually due to the variants having little independent disease risk but when present with other risk variants in genes which interact, the low pathological effect propagates into greater risk (Morgan, 2011; Tosto & Reitz, 2013).

1.4.2.4 Rare: low frequency, moderate risk

Rare variants form another cluster of variants associated with AD, sharing similar inheritance and disease risk (Schork et al., 2009). They tend to have a higher risk and may drive disease pathology more than *APOE* is observed to, but the variant remains rare as accumulation of too many rare variants or the specific rare variant with too high penetrance may lead to fatality or premature death and thus prevent the variant being passed down (Pabinger et al., 2014). Rare variants are often so infrequent they are not included in the genotyping arrays developed for studies and their discovery occurs through analysis of whole genome sequence data, where rare variants with higher risk are sequenced in smaller association studies and validated in later studies (Y. Li et al., 2008; Tosto & Reitz, 2013; Vardarajan et al., 2015; Visscher et al., 2012).

1.4.3 Complex Aetiology

The results of GWAS identify a spectrum of variants of increasing risk and frequency, this describes the complex architecture now known to lead to development of AD (Pabinger et al., 2014; Pottier et al., 2012; Reich & Lander, 2001). Further hypotheses on the mechanism by which disease pathology occurs consider the onset of AD to be a combination of genetic risk, environment and upbringing, lifestyle and care in later life (Gatz et al., 2006).

1.5 Genotyping and Sequencing

The first human genome sequenced was completed using shotgun sequencing methods, however, in the process of bringing down costs alternative sequencing methods were used and refined (Grada & Weinbrecht, 2013; Ng et al., 2010). Nowadays, high throughput next-generation sequencing is the established method for large genomics studies and automated sanger sequencing is the 'gold standard' for accurately detecting polymorphisms; methods primarily vary depending on the desired product and study design (Goh & Choi, 2012; Sanger et al., 1977).

The ability to genotype large cohorts has provided great advancements with the development of GWAS (Manolio, 2010). Association studies identify susceptibility loci which are frequently observed alongside disease traits; the studies identify numerous loci across the genome in complex diseases, some of which are considered to reach genome-wide significance, one in a million ($p \leq 5 \times 10^{-8}$) (J. C. Lambert et al., 2013). These loci are found within gene coding regions as well as intronic and intergenic regions; the variants within the loci outside of gene coding regions are usually more common and have smaller effect on disease risk (J. C. Lambert et al., 2013).

GWAS begins with the identification of loci from where a signal is derived but, due to linkage disequilibrium (LD), it is difficult to identify the risk variant amongst those it is

inherited with (Bush & Moore, 2012; Rosenthal & Kamboh, 2014). The loci are identified as an LD block which harbours the variant of interest (Manolio, 2010; Rosenthal & Kamboh, 2014). Follow-up studies are required to identify these specific variants, this process can be easier for exonic LD blocks whilst some signals from intergenic regions are yet to be explained (Clarke et al., 2011). When genotyping individuals for previously established genetic variants it is most practical to use a genotyping array; when looking at variation within exonic gene coding regions, the most cost-effective sequencing method would be whole-exome sequencing (WES); when looking for any and all genetic variation for cohorts the whole-genome sequencing (WGS) method is most thorough (Bamshad et al., 2011; Ng et al., 2010; R. Sims et al., 2017; Visscher et al., 2012).

1.5.1 Whole-genome sequencing

The most comprehensive and expensive sequencing method for an individual is whole genome sequencing (WGS) (Grada & Weinbrecht, 2013; Sanger et al., 1977; Visscher et al., 2012). This method includes all coding and non-coding regions and, during its processing stages, can produce multiple long reads which are especially useful for data analysis investigating copy number variants (CNVs), indels, regulatory elements and structural variants (Depristo et al., 2011; Koboldt et al., 2013; D. Sims et al., 2014).

1.5.2 Whole-exome sequencing

Many studies are more specifically interested in looking at variation within genes, as they code for proteins and make up around 2% of the genome and therefore require a fraction of the data storage space (M. Choi et al., 2009; Goh & Choi, 2012; Ng et al., 2010). WES requires less sequencing time and money and is a more popular

sequencing method than WGS for its efficiency and specificity of its approach (Bamshad et al., 2011; Guo et al., 2012; Warr et al., 2015).

1.5.3 Genotyping

Genotyping arrays are the most popular method for sequencing larger cohorts for specific variants of interest. Independent sequencing assays are completed for every variant present on the array, plates are designed to carry up to 92 individuals. Genotyping arrays can be designed to detect variation across the genome for association with one or a group of diseases, such as the NeuroX and NeuroChip arrays and neurological disorders (Blauwendraat et al., 2017; Nalls et al., 2015).

1.5.3.1 NeuroX array

The NeuroX genotyping array is the first iteration of a series of arrays designed to capture genetic data associated with neurological disorders (Barber et al., 2017; Nalls et al., 2015). The NeuroX was built on the HumanExome BeadChip v1.1 which consists of 242,901 SNPs across the whole exome, with an additional 24,706 SNPs which were either candidates for or known to be associated with diseases such as Alzheimer's disease and frontotemporal dementia, Parkinson’s disease, amyotrophic lateral sclerosis and multiple sclerosis, multiple systems atrophy, Charcot Marie Tooth and myasthenia gravis (Barber et al., 2017; Illumina, 2011; Nalls et al., 2015).

1.5.3.2 NeuroChip array

The NeuroChip array is the successor to the NeuroX (Blauwendraat et al., 2017; Nalls et al., 2015). The NeuroChip was built on the backbone of the Infinium HumanCore-24 v1.0 array consisting of 306,670 SNPs, with an additional 179,467 SNPs of custom content (Blauwendraat et al., 2017). Use of the HumanCore-24 backbone allows for greater coverage of non-exonic variants and improved genome-wide resolution; alongside variants more recently implicated in the disorders covered by the NeuroX

array, the NeuroChip includes variants implicated in dementia with Lewy bodies, progressive supranuclear palsy and corticobasal degeneration (Blauwendraat et al., 2017).

1.5.4 Imputation

Imputing array data can increase the number of variants individuals are genotyped for in a more cost-effective way than other NGS methods. Imputation involves phasing inputted sample data into haplotypes to identify blocks of linkage disequilibrium. These regions are compared to a reference panel of individuals' phased data to statistically calculate the likelihood of variants present in the reference to also be present in the sample. The accuracy of calls were given as INFO scores, which is affected by the number of initially genotyped variants in the LD block and the size of the population in the reference panel. Imputation requires a lot of computation and memory space especially when using large reference panels.

In recent years servers have become an alternative and user-friendly utility to impute array data; the Michigan Imputation Server (MIS) is often used as a free imputation service which uses the Minimac4 engine and stores data on Cloudgene (Das et al., 2016).

Imputation gives rise to more variants genotyped per individual, leading to coverage of relevant variants identified as associated with disease after the architecture of an array. The SNPs are identified when compared to reference panel which depending on its size can provide around 39 million SNPs per individual, compared to arrays which may potentially carry up to one million variants. When comparing imputed data to the original array data, there may be greater cross-over with datasets used in analysis, improving overall accuracy.

1.6 Polygenic risk scoring

Polygenic risk scores (PRS) are understood to determine the risk associated with a specific phenotype across multiple genes, usually across the whole genome (Euesden et al., 2015; S. A. Lambert et al., 2019). Differing nomenclature includes polygenic hazard score (PHS) and omnigenic risk score (ORS), sometimes due to minor variations in methodology. The fundamental method uses variants which are given an estimated value of effect from a base dataset, the number of which harboured by an individual is summated to identify their overall risk of a phenotype, and compared at different thresholds of inclusion of variants based on the significance of their association with the phenotype (Escott-Price, Myers, et al., 2017; Euesden et al., 2015; S. A. Lambert et al., 2019). This method is completed on groups of individuals in a target dataset with and without the phenotype, to determine which PRS model is most useful in differentiating the groups and therefore most effective at predicting likelihood of developing the phenotype (S. A. Lambert et al., 2019).

Base variant effect scores used as in PRS analyses are derived from summary statistics from GWAS meta-analyses or association testing of case-control data and are required to be independent from the target dataset due to be validated (Euesden et al., 2015).

PRS has many utilities, with more in development; association of genetic risk with other co-factors can be useful in more accurately determining disease likelihood and estimating age at disease onset (Darst et al., 2017; Foo et al., 2021; W. W. Li et al., 2020; Porter et al., 2018). Modelling PRS between cases and controls can be used to derive predictability of disease risk for undiagnosed individuals or individuals diagnosed mild cognitive impairment (Chaudhury et al., 2019; Logue et al., 2019). Selection of individuals at greatest risk of disease for clinical trials has the utility of

observing the maximum potential effects of treatment on disease onset (Euesden et al., 2020; S. A. Lambert et al., 2019).

Further study of number of gene variants associated with a disease required formatting the variants into sub-groups (Darst et al., 2017; Lawingco et al., 2021). The methods of variant classification include allele frequency, by chromosome, by genes which are active in the same pathway or cascade, and genes which are all expressed or upregulated in the same system (Andrews et al., 2020; Darst et al., 2017; Femminella et al., 2021; Hu et al., 2017; Lawingco et al., 2021).

1.8.1 Literature

PRS analysis has been used to investigate AD since methods for doing so were determined, now varying methods and sample cohorts have been used in studies. Exploring literature using the search criteria of "polygenic risk score" AND "Alzheimer's disease" identified 55 articles between 2016 and 2021, varying from primary studies to reviews (Bellou et al., 2020; S. A. Lambert et al., 2019).

The meta-analysis by the International Genomics of Alzheimer's disease Project (IGAP) consortium has been a significant contributor to most studies of PRS in AD (Andrews et al., 2020; Axelrud et al., 2019; Chandler et al., 2019; Chaudhury et al., 2018, 2019; Cruchaga et al., 2018; Darst et al., 2017; Del-Aguila et al., 2018; Ebenau et al., 2021; Elman et al., 2020; Escott-Price et al., 2015, 2019; Escott-Price, Myers, et al., 2017; Escott-Price, Shoai, et al., 2017; Escott-Price & Schmidt, 2021; Euesden et al., 2020; Femminella et al., 2021; Foo et al., 2021; Fulton-Howard et al., 2021; Ge et al., 2018; Gibson et al., 2017; Hong et al., 2020; Huq et al., 2021; Kauppi et al., 2020; Korologou-Linden et al., 2019; Kremen et al., 2019; Lawingco et al., 2021; Leonenko et al., 2021; Leonenko, Shoai, et al., 2019; Leonenko, Sims, et al., 2019; Logue et al., 2019; Matloff et al., 2020; Tasaki et al., 2018, 2019; Wehby et al., 2018; Yesavage et al., 2020).

Used as the base dataset, the summary statistics on data of 74,046 individuals, provided sufficient power to determine many variants to be identified to have genome-wide significance (J. C. Lambert et al., 2013). Recently, this was increased to include data of 94,437 individuals and used in subsequent studies (Kunkle et al., 2019; Najar et al., 2021; Skoog et al., 2021; Stocker et al., 2021). The most comprehensive meta-analysis to-date utilises 1,126,563 individuals, its potential benefits still to be determined (Wightman et al., 2021).

1.8.2 Association

The results of PRS studies have identified associations between PRS and biomarkers or other diagnostic criteria (Darst et al., 2017; Foo et al., 2021; W. W. Li et al., 2020; Porter et al., 2018). Patients with CSF biomarkers for AD have been found to have high PRS, identifying high PRS individuals as targets for early screening of biomarkers before dementia symptoms (Darst et al., 2017; W. W. Li et al., 2020). The association between higher PRS and lower hippocampal subfield volume has been observed when measured over time; as individuals with high PRS develop a reduced hippocampal volume endophenotype in old age, similar PRS in younger individuals may indicate likelihood of developing this in later life (Foo et al., 2021).

1.8.3 Prediction

Predictions can also be made from the PRS of a cohort with known phenotypes, resulting in the successful identification of at-risk individuals before the onset of disease symptoms (Chaudhury et al., 2018, 2019; Dudbridge, 2013; Escott-Price, Myers, et al., 2017; S. A. Lambert et al., 2019; Leonenko et al., 2021; Logue et al., 2019; Porter et al., 2018; Stocker et al., 2021). The PRS of patients diagnosed with mild cognitive impairment compared to a case-control cohort has found trends which

indicate the likelihood of conversion to AD in patients with highest PRS (Chaudhury et al., 2019; Logue et al., 2019).

Deriving PRS using selection criteria (e.g., common variants, MAF > 5%) has been a popular approach to efficiently determine the risk of AD in the general population (X. Jin et al., 2021). Many of the variants used in analyses have been previously identified in GWAS meta-analyses, enabling replication of methods and substantial research undertaken to understand the role of the variants in AD pathology (Desikan et al., 2017; Leonenko, Sims, et al., 2019). A Polygenic Hazard Score has often been classified as using a limited number of SNPs identified by GWAS as the basis for measuring genetic risk (Desikan et al., 2017).

1.8.4 Specialisation

Expression network and pathway driven PRS approaches are a method by which the genetic risk of a smaller subset of variants can be more specific in predicting the risk of disease from the variants in the genes as well as the association of the pathway or network to disease pathology (Darst et al., 2017; Femminella et al., 2021; Hu et al., 2017; Lawingco et al., 2021; Morgan, 2011).

Understanding the relationship between proteins expressed at the synapse was developed into PRS analysis to determine the gene variants with most significant association to AD alongside the utility of predicting likelihood of AD in patients exhibiting synaptic dysfunction (Lawingco et al., 2021; Lleó et al., 2019). These approaches have identified 6 potential CSF biomarkers for AD and a PRS model consisting of 8 SNPs across 6 genes sufficient to predict likelihood of AD greater than the *APOE* SNPs alone.

PRS have been derived from 19,630 variants in 1,158 mitonuclear genes, those affecting the expression and function of mitochondria; this study has shown strong

associations with risk of AD as well as age at onset, developing the understanding of mitochondrial activity on AD pathology (Andrews et al., 2020). The utility of specialised PRS analyses in understanding AD pathology advocates for similar approaches to be taken for other AD associated pathways and could support improvements in discerning between dementia sub-types.

1.8.5 Software

Calculating PRS can be achieved by utilising various bioinformatics tools; dedicated software has also been developed for this purpose, such as PRSice (Euesden et al., 2015). The first iteration of this software, PRSice v1.25, utilises other bioinformatics tools for calculating scores, computing regression models and producing figures to present this (Euesden et al., 2015). The second iteration, PRSice-2, incorporates complex coding to maximise its efficiency, enabling analysis at larger scale in terms of size of data and number of models (S. W. Choi & O'Reilly, 2019).

1.7 Project Aims

This PhD aims to explore the current genetic landscape of AD, identifying the best methods to calculate genetic risk, improve risk prediction and propose future applications of polygenic risk scores. The thesis will cover numerous genotyping methodologies, comparing their consistency, accuracy, ease of generation, quality assurance and quality control.

The study will be primarily focussed on a cohort of individuals recruited by the Brains for Dementia Research resource, comprising of individuals across the UK, including healthy controls, patients with AD, and other dementia sub-types; deceased individuals have been post-mortem pathologically confirmed (Francis et al., 2018). The study will also use genetic data from international GWAS, international genotyping projects and the genotype data of a group of individuals diagnosed with

MCI and recruited into the Inflammation, Cognition and Stress (ICOS) longitudinal study (Sussams et al., 2013).

PRS can be primarily used to determine disease likelihood, taking the distribution of risk amongst cases, controls and a general population and determining thresholds for lifetime-risk classification of preclinical individuals. PRS will also be explored similarly as a utility for predicting conversion likelihood in individuals with MCI. PRS can also be further utilised to differentiate sub-type classification, identifying whether the genetic variants associated with sub-type are mutual or exclusive to other sub-types.

The variation of information which produces the PRS will be explored to identify if prediction can be achieved with a subset of variants based on filtering by pathway or expression platform and what conclusions or trends can be observed by these studies; further exploration of a series of subsets can determine if PRS in multiple subsets can lead to a deeper understanding of AD aetiology and how this may vary between individuals.

2 Methods

2.1 Laboratory

2.1.1 Samples

Lab-based methods outline the recruitment of individuals for genotyping studies, with priority for patients diagnosed with Alzheimer's disease or other dementia sub-types or prodromes. The individuals and their biological samples were collected from various resource and study groups. The samples primarily used throughout this thesis were provided by the Brains for Dementia Research (BDR) resource and the inflammation, cognition, and stress (ICOS) study group.

The BDR group is currently coordinated from Newcastle; they recruit individuals from five brain banks in the UK: University of Bristol, King's College London, University of Manchester, University of Newcastle, and University of Oxford. Individuals are subject to the Montreal cognitive assessment as well as providing a blood sample and donating brain tissue at death, at which point clinical features and post-mortem disease status were reported. All samples were obtained with written informed consent and approval from participants and the BDR and University of Nottingham ethics committees. As of March 2019, the resource provided the blood samples of 1164 individuals of which 760 were given a post-mortem pathological diagnosis, the majority were genotyped using the NeuroChip array and many have been whole exome sequenced.

The ICOS study group is based in Southampton, UK, for individuals with mild cognitive impairment; the study has followed the patients over time to record changes, specifically to clinically diagnose conversion to AD or other dementias. As of 2017, 124 patients provided a blood sample for NeuroChip genotyping.

2.1.2 DNA extraction

DNA was extracted from 2mL of blood using a phenol chloroform method (Sigma Aldrich and Qiagen). Blood samples were defrosted at room temperature for 30 minutes and then inverted to suspend cells. First, 2mL of whole blood was added to a 15mL falcon tube containing 20µL RNase A. Then, 200µL of Proteinase K, for protein degradation, and 2mL of AL lysis buffer, which breaks down the cellular membrane, were added and mixed well. The homogenised sample was incubated at 56°C in a thermoshaker for 10 minutes.

Two 15mL MaxTract High Density Tubes (HDT) (Qiagen) were spun at 1500g for 3 minutes. The incubated sample transferred to one of the HD tubes followed by an equal amount of Phenol:Chloroform:Isoamyl alcohol (25:24:1) pH 8.0 (Sigma-Aldrich). The HDT is centrifuged for 7 minutes at 1500g at room temperature using a PRISM benchtop microcentrifuge (Abnet). Centrifugation with phenol chloroform allows for separation of DNA, protein, and cellular debris into layers. DNA, which accumulates in the upper aqueous layer is carefully pipetted into the second pre-spun HDT, and the volume is noted. In the fume hood, equal parts Phenol:Chloroform:Isoamyl alcohol is added to re-homogenise the sample and then centrifuged at 1500g for 7 minutes. The upper phase is transferred into a new falcon tube and the volume is noted. At the bench, 3M sodium acetate (Sigma) is added to the falcon tube equivalent to 10% of the upper phase volume and inverted twice. Equal parts of cold 100% ethanol is added to the tube and inverted gently for precipitation of the DNA. The sample is either stored at -20°C overnight or kept on dry ice for 1 hour.

The falcon tube is spun at 4°C at 6000rpm for 10 minutes. Supernatant was discarded and 1mL of 70% ethanol was added to the tube to wash the pellet. The tube was again centrifuged at 4°C at 6000rpm for 10 minutes and excess alcohol was discarded. The

wash was repeated once again, and the tube was left to air dry for 45 minutes at room temperature.

Finally, 100µL of Tris-EDTA Buffer (Qiagen) was added to resuspend the pellet. The tube was then left overnight at 4°C followed by incubation on a heat block for 40 minutes at 50°C. DNA concentration was measured using the Nanodrop Fluorometer 3300 and degradation was estimated using 1% agarose alongside a 1000kb ladder. The DNA was then able to be stored at either -20°C for use within two weeks or -80°C if the sample required storing for longer.

2.1.3 DNA quality control

Samples underwent 10x dilution with nuclease-free water, this ensured the concentration levels of DNA were within suitable detection ranges for spectrophotometry. All samples were quantified with a Nanodrop Spectrophotometer 2200 (Thermo Scientific) following protocols provided with the Quant-iT dsDNA Broad Range Assay Kit (Life Technologies). The concentration of double stranded DNA was quantified from the sample as the Nanodrop 2200 used PicoGreen fluorescence; this was repeated in triplicate. The concentration of dsDNA was quantified as opposed to all DNA to provide a more accurate reading, and results were in triplicate to ensure the readings were reliable. DNA yield was calculated in 45µL; for whole exome sequencing, the samples needed to be above 10µg.

The TapeStation (BioAnalyser) was used to measure DNA integrity, 1µL of sample was added and mixed with 10µL buffer in PCR tubes and loaded onto the machine with a genomic DNA screentape. DNA is quantitatively measured by its DNA Integrity Number (DIN) score, the output of the TapeStation. Samples are required to have a DIN score ≥ 6.0 and appear as a clean band on the gel, samples with scores below 6.0

were diluted again and repeated and where the band was smeared the DNA required re-extracting from remaining tissue or blood.

2.1.4 TaqMan assays

TaqMan assays (Applied Biosystems) were performed using primers and probes provided by the manufacturer to manually genotype the two SNPs associated with *APOE* status, rs429358 and rs7412. These variants are exclusively assayed due to the GC-rich nature of the region in which they are present on chromosome 19, leading to difficulties and inaccuracies when genotyped on array platforms.

Procedures are completed in a sterile laminar flow hood, where all equipment was sterilised under ultraviolet light irradiation for at least 20 minutes before use; light sensitive reagents, the assay mix and TaqMan master mix were thawed at room temperature and covered with foil to prevent degradation. The lamp on the Mx3000P real-time quantitative PCR (qPCR) machine (Agilent) requires switching on 20 minutes before use to allow time to warm up and the machine to be primed.

To produce a master mix for the reactions, 3µL of ddH_2O was added to 4.5µL of TaqMan master mix and 0.5µL of assay mix containing primers for the SNP assays; reagent volumes were multiplied for the number of reactions taking place. A 96 well optic PCR plate was loaded with 2µL of DNA sample at a concentration of approximately 20ng/µL with the 8 µL of reagent master mix; 2µL of nuclease free water was used in lieu of a DNA sample as a no template control (NTC) whilst 2µL of DNA from Sanger sequencing confirmed positive controls for the wild type, homozygous and heterozygous mutants are included on each genotyping plate for validation. Plates are sealed with optical caps and reagents are thoroughly mixed by being vortexed. The plates were then placed in the qPCR machine, the plate plan was set up to identify the samples in each well and locations of the NTCs and positive

controls. The PCR thermal cycle was set to run at 95°C for 10 minutes, followed by 60 cycles of 92°C for 15 seconds and 60°C for 1 minute.

Alleles present at each SNP were determined by comparing Hexachloro-fluorescein (HEX) and 6-Carboxyfluorescein (FAM) fluorescence curves to positive controls. This was done by observing signals from FAM fluorescence representing T alleles and HEX representing C alleles or signals from both representing heterozygous calls. Once the two SNPs are called, *APOE* status can be determined from the alleles present at each variant and therefore individuals' haplotypes. Many individuals show a fluorescent signal for FAM for rs429358, genotyped as TT, and CC for rs7412 from a HEX fluorescent signal; their haplotypes would therefore be T/C and T/C and their *APOE* status is classified as ε3ε3, as shown in Table 1.

	rs429358		rs7412	
APOE status	FAM	HEX	FAM	HEX
ε2ε2	1	0	1	0
ε2ε3	1	0	1	1
ε3ε3	1	0	0	1
ε3ε4	1	1	0	1
ε4ε4	0	1	0	1

Table 1: APOE genotype calls by TaqMan assay

The table describes the FAM (T allele) and HEX (C allele) signals present at both *APOE* SNPs to make genotype calls, where 1 indicates the presence of a signal and 0 indicates the absence of a signal.

2.1.5 Comparison of *APOE* genotyping methods

As mentioned previously, the two *APOE* SNPs require TaqMan assays to confirm the genotypes for individuals. This is due to previous examples of genotyping arrays failing to call either or both variants even when repeated. There is also observed evidence of discrepancies between *APOE* genotypes when using alternative variant calling methods i.e., Sanger sequencing, whole-exome or whole-genome sequencing, and imputation.

2.2 Bioinformatics

2.2.1 Genetic data

Genotype data collected through various methods requires processing and storage in formats that can allow computation and manipulation by other software. As new programs and tools are still being developed, genotype data is required to be able to be stored effectively in universal file formats, which can otherwise be converted to more specific and appropriate file formats for the data analysis tools.

2.2.1.1 Variant call format

The most common method of storing individual DNA polymorphic data is in variant call format (VCF), developed during the 1000 Genomes project. The VCF file can store data for SNPs, insertions, deletions, structural variants and informs further annotative points regarding the quality of the variant and its known function and association with disease. The VCF is typically a storage and transfer format for data; data can be further compressed to save space, whilst there is no theoretical limit to the number of individuals who can be included within a VCF file (Danecek et al., 2011). VCF can be easily manipulated and used for a variety of functions, the files are easily readable and can be viewed by most text editors; software developed for analysis of genetic data is usually designed to either import or export in VCF.

2.2.1.2 Genome Analysis Toolkit

The Broad Institute, who work heavily with genetic data, developed the Genome Analysis Toolkit (GATK). GATK contains a range of programs and tools to edit data; tools are used to process the raw genotype data into formats which can be for manipulating and quality controlling data before being converted into the final VCF.

There are many formats which raw genotype data can come in, Illumina is the leading company in high throughput sequencing, and usually produces the raw forward and

reverse reads as a pair of FASTQ files per individual. FASTQ sequence files are text files containing the entire genomic base sequence with Phred quality scores for each nucleotide represented by ASCII characters.

Sequence data requires alignment to an assembly. This is completed, and the resulting data is formatted into Sequence Alignment Mapped (SAM) format. SAM files are more comprehensive than FASTQ, combining both FASTQ reads and indicating the reference genome and alignment; with quality scores, the mapping position for each nucleotide and where its paired read align are also presented. SAM files are larger in comparison to most files, and in order to be stored efficiently, the data is indexed and compressed into a binary SAM (BAM) file. Indexing allows other programs to identify and extract key information whilst the data is in its compressed format. GATK tools are also self-encompassing, so any scripts which require decompression of specific regions for analysis finish their processes by re-compressing afterwards.

2.2.1.3 PLINK

PLINK is an open-source genome association analysis tool developed by Massachusetts General Hospital and the Broad Institute. The toolkit can be used to further manipulate genetic data and complete association-based analysis. PLINK is compatible with VCF files, but analysis using the software requires conversion from VCF to other data formats for analysis. The alternative file formats to VCF can usually require less space, are more easily manipulated and some formats can be read in text editor software.

A VCF contains all the variants an individual is genotyped for, in order of chromosomal position; this can be isolated by PLINK as a MAP file. When PLINK produces the MAP file, the genotypes of all individuals present in the VCF are also exported in the order they appear in the MAP file as a PED file; the PED file may also, where provided,

include individuals' sex and phenotype, and can also store individuals maternal and paternal identifiers. The MAP and PED files are known as standard PLINK files. PLINK can also produce binary files which further reduce storage space whilst remaining accessible for analysis procedures. The PLINK binary files are like the MAP and PED files, as one contains a list of all the variants (BIM) whilst another holds the same individuals' information reported in PED files (FAM). The BIM and FAM files are separately readable but require the presence of the binary PED (BED) file to be read by PLINK for analysis and computation. With the most recent version of PLINK, the three file formats i.e., VCF, PLINK and binary files, are fully cross-compatible and can be converted from one another using PLINK tools.

2.2.2 Sequencing pipeline

2.2.2.1 Initial QC

Successfully run samples produced paired raw FASTQ files per sample, one in the forward direction and the other in the reverse direction containing paired reads. The best practice guidelines for next generation sequencing determined by Broad Institute recommended GATK v4.1.4, which includes programs and software necessary to complete their recommended pipeline. The NGS pipeline was designed to be used in a Unix operating system environment and the pipeline was completed within a Linux-based server.

Raw reads required initial assessment for quality control of the library prep and sequence run. FastQC (Babraham Bioinformatics) is a java-based program which produces a QC report of any identified issues. The report checked for contamination from external sources and adapter sequences, base composition bias from too rich GC content, sequence duplicates and overrepresentation, unexpected read lengths and per base quality scores. The accuracy at calling bases in a sequence was given using a

Phred quality scale (Q), it is vital to attain Phred scores of Q30 with its respective call error rate of 0.001 (99.9% accurate) for good quality data.

Usage:

```
$ fastqc --threads 8 sample_1.fastq -o sample_qc
```

--threads = number of processors

-o = output filename

2.2.2.2 Alignment

The first stage of the GATK pipeline was alignment of the short reads from the raw paired FASTQ files to a reference genome, producing a single file in sequence alignment mapped (SAM) format. SAM files store genomic position and quality information for each read. Alignment of all samples was completed using the Burrows-Wheeler Aligner (BWA) and aligned to the 1000 Genomes FASTA file as the reference genome; the algorithm also incorporated seeding to improve processing speeds by finding maximal exact matches (MEM) for where the sequence was most likely to align and used the affine-gap Smith-Waterman algorithm to build optimal alignments.

Usage:

```
$ bwa mem -t human_g1k_v37.fasta sample_1.fastq
sample_2.fastq > sample.sam
```

-t = number of processors

2.2.2.3 Sorting

The resulting SAM file was sorted using the SortSam Picard tool, which organises the mapped short reads in a defined order, indexes the SAM file and can be used to compress the SAM file. SAM files, which are usually between 20-30Gb in size, compress into a binary SAM (BAM) file, which are usually around 4-10Gb, saving space

whilst maintaining accessibility. Indexing allowed the BAM file to be used for downstream analyses without the requirement of decompressing the entire file when a region required loading, and the file was sorted in order of genomic position (coordinate). The resulting BAM was then checked again using FastQC to detect any issues after alignment, as a pre-processing step. FastQC also provided statistics to describe the coverage for each sample against the reference.

Usage:

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

$ fastqc --threads 8 sample.bam -o sample.align.qc
```

-Xmx4g = maximum allocation of memory to be used by java

SO = parameter to sort by (coordinate/unsorted/queryname)

CREATE_INDEX = generate an index file with output (TRUE/FALSE)

VALIDATION_STRINGENCY = bypass certain validation measures to reduce processing speed (STRICT/LENIENT/SILENT)

2.2.2.4 Read Groups

All reads which come from a single lane in a sequencing run of a sample are described as a read group. These can be identified by several read group tags which hold information about the sample in the header of the BAM file. For upstream processing stages, the read groups needed to be added or the existing information tagged for the read group had to be uniform and unique to the sample. The tags include sample name (RGSM), genotyping platform (RGPL), the barcode and lane of the flowcell (RGPU), a unique read group identifier (RGID) and the DNA preparation library identifier (RGLB). Existing read groups for the BAM file were reported in the header of the file and are checked using SAMtools view function; the information needed to be

recorded as to not replace the current read group tag information when adding or replacing tags. The Picard tool AddOrReplaceReadGroups was used to update read group information.

Usage:

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

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

-H = header

grep = search for string based on criteria ''

2.2.2.5 Validation

The Picard tool ValidateSamFile was used to confirm each sample BAM file was formatted correctly and all previous QC processes had not led to errors. The tool diagnoses any errors in processes and flags any other values which may need addressing. To pass validation, BAM files need to be aligned to a reference, sorted and with read groups; any issues reported in a summary file were re-run to resolve errors or otherwise the sample needed to be completely re-sequenced.

Usage:

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

MODE = output mode (verbose/summary)

2.2.2.6 Mark duplicates

During sequencing, raw reads can often be produced which align to the same position and contain the exact same sequence. Duplicates of the same read in sequence data

can lead to over-representation of some alleles and negatively affect variant calling. Duplicates are often produced as a result of amplification biases during PCR, the sample library prep could contain errors or low levels of starting material or enrichment methods can give rise to bias. Optical duplicates occur as a result of the sequencing instrument mistakenly producing artefacts where single amplification clusters are detected as multiple clusters by the optical sensor.

The Picard MarkDuplicates tool identified and marked these duplicates for each read so that GATK tools used in downstream analyses disregarded them from analysis. The base-quality score was used to differentiate between the primary read and its duplicates by ranking them based on the sum of their scores as part of the tool's algorithm. The output files of this process produced a BAM file with duplicates marked as well as a metrics file outlining all reported duplicates.

Usage:

```
$ 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
```

METRICS = output metrics file

2.2.2.7 Realignment around InDels

Databases of known variants are included in the GATK package; these include dbSNP, 1000 Genomes and HapMap. During initial alignment, errors may have occurred especially near the ends of reads, where indels may lead to misalignment of bases to the reference and produce sequencing artefacts; these would eventually lead to false positives to be reported during variant discovery. This was countered by using a two-step local realignment process. Firstly, the GATK tool RealignerTargetCreator used variant databases to list all potential indels present within the sequence and identified

intervals which required realignment. The next step was completed by the GATK tool IndelRealigner, which determined an optimal consensus sequence based on the intervals highlighted by the first step and performed local realignment of reads. This was completed for all samples based on the dbSNP v138 database.

Usage:

```
$ java -Xmx4g -jar GenomeAnalysisTK \
-T RealignerTargetCreator \
-R human_g1k_v37.fasta \
-I sample.sorted.rg.marked.bam \
-o indels.intervals \
--known dbSNP.vcf

$ java -Xmx4g -jar GenomeAnalysisTK \
-T IndelRealigner \
-I sample.sorted.rg.marked.bam \
-R human_g1k_v37.fasta \
-targetIntervals indels.intervals \
-o sample.sorted.rg.marked.realign.bam
```

-T = GATK tool

-R = reference fasta file

-I = path to input file

--known = VCF of known variants

-targetIntervals = list of target loci

2.2.2.8 Fix mate information

As IndelRealigner worked to adjust sequences around known indels, the process could lead to errors where fixed mates of paired-end reads of initial sequencing, which have not been realigned, are moved and are no longer the known and expected distance from one other. Mate pairs were defined during sequencing as the starting positions of the forward and reverse-complement orientated reads covering the same region. The Picard tool FixMateInformation ensures all mate pairs were re-fixed to their original positions.

Usage:

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

2.2.2.9 Base Quality Score Recalibration

Systematic errors during sequencing are one of many sources of technical errors in genotyped data, specifically base quality. The likelihood of correct base calling is quantified by quality scores, which are estimates of the error emitted by the sequencer. Accurate base calling is obtained by recalibration, which relies heavily on base quality scores. Variant calling algorithms use base quality scores as evidence for possible variant alleles at a specific site.

Base quality score recalibration (BQSR) is broken down into two stages, recalibration of scores and application of adjusted scores. Firstly, the GATK tool BaseRecalibrator produced a data table comparing the observed variants to the known sites of variation from multiple sources in VCF, any variants observed in the sequence data not previously known of are considered errors of poor base quality. The empirical probability of error was calculated at each site based on the number of mismatches and observations, with covariates for read group, reported quality score, machine cycle and nucleotide context. The GATK tool ApplyBQSR was then used to allocate new base quality scores to the sample file from the table produced by BaseRecalibrator.

Usage:

```
$ ./gatk BaseRecalibrator \
--input sample.sorted.rg.marked.realign.fixed.bam \
--reference human_g1k_v37.fasta \
--known-sites dbSNP.vcf \
--known-sites 1000G_phase1.indels.b37.vcf \
--known-sites 1000G_phase1.snps.high_confidence.b37.vcf
--known-sites
Mills_and_1000G_gold_standard.indels.b37.vcf \
--output sample_recal_data.table
```

```
$ ./gatk ApplyBQSR \
--input – sample.sorted.rg.marked.realign.fixed.bam \
--bqsr-recal-file sample_recal_data.table \
--output sample.sorted.rg.marked.realign.fixed.recal.bam
```

--input = input file

--reference = reference FASTA sequence

--known-sites = multiple VCF of known sites

--output = output file

--bqsr-recal-file = input recalibration file from BaseRecalibrator

2.2.2.10 Variant calling

Variants are called as part of a workflow which eventually combines multiple samples into a single VCF file. The first stage was completed by the tool HaplotypeCaller, which is the recommended tool for calling SNPs and indels. The tool uses multiple algorithms to redefine regions, ignoring existing mapping information, where signs of variation amongst reads occur; this leads to local de-novo assembly of haplotypes in an active region. A De Bruijn-like graph was built for each active region to reassemble the region by identifying both the haplotypes present, this was then aligned to the reference haplotype using the Smith-Waterman algorithm.

The likelihoods of different haplotypes identified using the algorithm were tested using a Pair Hidden Markov Model (PairHMM) algorithm based on the read data, the most likely genotype was then assigned for each potential site according to Bayes' ruling. The resulting file is a genomic VCF (gVCF) of raw unfiltered variants, these per-sample gVCFs were then merged with other samples using the GATK CombineGVCFs tool and joint genotyping was performed using GATK GenotypeGVCFs and produced a final VCF for all samples.

Usage:

```
$ ./gatk HaplotypeCaller \
--reference human_g1k_v37.fasta \
--input sample.sorted.rg.marked.realign.fixed.recal.bam \
--emit-ref-confidence GVCF \
```

```
        --dbsnp dbSNP.vcf \
        --output sample.g.vcf

        $ ./gatk CombineGVCFs \
        --reference human_g1k_v37.fasta \
        --variant sample.g.vcf \
        --dbsnp dbSNP.vcf \
        --output cohort.g.vcf

        $ ./gatk GenotypeGVCFs \
        --reference human_g1k_v37.fasta \
        --variant cohort.g.vcf \
        --output BDR_cohort.vcf
```

--emit-ref-confidence = mode for emitting reference confidence scores

--variant = separate inputs for each sample

2.2.2.11 Variant Quality Score Annotation and Recalibration

In order to provide context when filtering variants, properties of the sequence data specific to variants requires annotating. GATK tool VariantAnnotator was used to provide additional information for variants with regards to variant IDs according to the reference and depth of coverage. The VCF was then filtered to remove false positives and false negatives from amongst true calls by variant quality score recalibration (VQSR).

VariantRecalibrator firstly built a recalibration model for the inputted VCF and using resources from HapMap, 1000 Genomes and dbSNP of known variants and true sites of variation, trained the tool to identify true calls from false calls which had resulted from sequence or data processing artefacts. The tool scored each variant under the Gaussian mixture model with a log odds value (VQSLOD) which was included in the INFO of the variant in the VCF file.

The second stage of VQSR was filtering the variants based on their scores, which needed to be completed for SNPs and indels separately. Filtering was subject to the VQSLOD score and the target sensitivity value, GATK tool ApplyVQSR marked variants for which tranche they fell within (90%, 99%, 99.9%, 100%); if a limit of 99% was defined, the variants marked only in the 90% tranche and therefore did not fall within

the 99% or above tranches would be marked for filtering out using the GATK SelectVariants tool.

Usage:

```
$ ./gatk VariantAnnotator \
--reference human_g1k_v37.fasta \
--variant BDR_cohort.vcf \
--dbsnp dbSNP.vcf \
--output BDR_cohort_annotated.vcf

$ ./gatk VariantRecalibrator \
--reference human_g1k_v37.fasta \
--variant BDR_cohort_annotated.vcf \
--resource
dbSNP,known=true,training=false,truth=false,prior=2.0:dbS
NP.vcf
--resource
1000G,known=false,training=true,truth=false,prior=10.0:10
00G_phase1.snps.high_confidence.b37.vcf \
-an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an
SOR \
-mode SNP \
--output output.recal \
--tranches-file output.tranches \
--rscript-file output.plots.R

$ ./gatk ApplyVQSR \
--reference human_g1k_v37.fasta \
--variant BDR_cohort_annotated.vcf \
--output BDR_cohort_annotated_VQSR.vcf \
--truth-sensitivity-filter 99.0 \
--tranches-file output.tranches \
--recal-file output.recal \
-mode SNP

$ ./gatk VariantRecalibrator \
--reference human_g1k_v37.fasta \
--variant BDR_cohort_annotated_VQSR.vcf \
--resource
Millsand1000G,known=false,training=true,truth=false,prior
=10.0:1000G_gold_standard.indels.b37.vcf \
-an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an
SOR \
-mode INDEL \
--output indels_output.recal \
--tranches-file indels_output.tranches \
--rscript-file indels_output.plots.R

$ ./gatk ApplyVQSR \
--reference human_g1k_v37.fasta \
--variant BDR_cohort_annotated_VQSR.vcf \
--output BDR_cohort_annotated_VQSR2.vcf \
--truth-sensitivity-filter 99.0 \
--tranches-file indels_output.tranches \
--recal-file indels_output.recal \
-mode INDEL
```

--resource = details and file of annotation resources

-an = annotations to be included

-mode = recalibration mode to employ (SNP/INDEL/BOTH)

--tranches-file = the output tranches file to be used by ApplyRecalibration

--rscript-file = the output script to be run in R to produce plots

--truth-sensitivity-filter = sensitivity level to retain SNPs by (90.0/99.0/99.9)

--tranches-file = input tranches file

--recal-file = input recalibration file

2.2.2.12 Functional annotation

Variants with known functions can be identified within the input file and their known functions can be annotated onto the VCF file. Known functions are recorded in external data sources like gnomAD and are made available using the GATK FuncotatorDataSourceDownloader tool; the tool requires specification as to which variants are needed, whether they require validation and whether to extract after importing. The GATK Funcotator tool was then used to locate and match the variants function from the data source and applied this to the variants in the file. The output of the tool was a VCF file with the matched functional annotations included.

Other variant annotation tools exist outside of GATK, these include ANNOVAR, SnpEff and Variant Effect Predictor (VEP).

Usage:

```
$ ./gatk FuncotatorDataSourceDownloader --germline --
validate-integrity --extract-after-download

$ ./gatk Funcotator \
--reference human_g1k_v37.fasta \
--variant BDR_cohort_annotated_VQSR2.vcf \
--output BDR_cohort_Funcotator \
--output-file-format VCF \
--data-sources-path dataSourcesFolder/ \
--ref-version hg19
```

--output-file-forma = format in which results should be produced (VCF/MAF/SEG)

--data-sources-path = path to data sources from FuncotatorDataSourceDownloader

--ref-version = version of reference to use (hg19/hg38)

2.2.2.13 Quality checks

Finalised variant data was analysed, and quality metrics were obtained using multiple tools including VariantEval, TsTv-by-count (VCFtools), CalculateHsMetrics (Picard) and depth (VCFtools). VariantEval and TsTv-by-count are tools used as a data quality check by calculating the number of transitions and transversions which occur amongst different classes of variants and only bi-allelic variants, respectively. The transition/transversion (Ti/Tv or Ts/Tv) ratio describes the proportion of purine-purine or pyrimidine-pyrimidine changes compared to purine-pyrimidine or pyrimidine-purine changes and is a useful indicator for the quality of different sequencing methods; it is expected for whole-exome sequencing data to have a Ti/Tv ratio around 2.8, where lower than this would suggest an excess of false negatives.

CalculateHsMetrics computes hybrid selection specific metrics from BAM files, metrics were therefore completed on the BAM file produced before variant calling. Hybrid selection metrics included AT/GC dropout, GC content and mean coverage. The depth tool within VCFtools can also calculate average depth of coverage for all variants.

Usage:

```
$ java -Xmx128g -jar GenomeAnalysisTK \
-R human_g1k_v37.fasta \
-T VariantEval \
--eval BDR_cohort_filtered.vcf \
--dbsnp dbSNP.vcf \
-o BDR_cohort_eval.grp

$ vcftools --vcf BDR_cohort_filtered.vcf \
--TsTv-by-count --out BDR_cohort.tstvcount

$ java -Xmx128g -jar picard/BedToIntervalList \
I=SureSelect_Regions.bed \
SD=human_g1k_v37.dict \
O=SureSelect_PicardIntervals.list

$ java -Xmx128g -jar picard/CalculateHsMetrics \
R=human_g1k_v37.fasta \
I=sample.sorted.rg.marked.realigned.fixed.recal.bam \
O=BDR_cohort_HsMetrics.txt \
TARGET_INTERVALS=SureSelect_PicardIntervals.list \
BAIT_INTERVALS=SureSelect_PicardIntervals.list
```

```
$ vcftools --vcf BDR_cohort_filtered.vcf \
--depth --out BDR_cohort.depth
```

--eval	= input file to be evaluated
--vc	= input VCF file
--out	= output tstvcount file
SD	= reference sequence dictionary
TARGET_INTERVALS	= list of target locations
BAIT_INTERVALS	= list of bait locations
--depth	= calculate depth of coverage

2.2.3 Genotyping pipeline

2.2.3.1 Initial clustering

Genotyping for the NeuroChip array was completed following manufacturer's instructions on the HiSeq sequencer by collaborators at University College London, London. Quality control for the NeuroChip was completed following guidelines stipulated by the Cancer Biology research group at Vanderbilt University. The raw intensity data files (IDAT) for three batches of samples were imported onto GenomeStudio v2.0 (Illumina), and automatic clustering was completed using a cluster file provided by Blauwendraat and colleagues; a GenCall threshold was set as 0.15. GenomeStudio describes all individuals as samples and all markers for variants as SNPs.

The clustering algorithm considers calls for each SNP, it identifies the signals of the two probes for each allele (A and B) and the intensity by which the fluorescently labelled target sequences bind, which is dependent on the hybridisation conditions. The results of the fluorescence levels are scored by R and Theta values which were normalised and plotted with R on the y-axis and Theta on the x-axis. The algorithm then clusters the included samples to group those it identifies as homozygous major (AA), heterozygous (AB) and homozygous minor (BB).

Once clustering is completed, SNP and sample call frequencies and rates are calculated respectively; samples which fall below 95% call rate in the Samples Table are selected and excluded from analysis. The remainder of quality control is subject to identification of SNPs which do not follow expected cluster criteria and are zeroed in GenomeStudio; SNPs which are zeroed are excluded from the dataset when it is eventually exported from GenomeStudio and subject to downstream QC in PLINK.

2.2.3.2 Manual re-clustering

The results of the clustering algorithm may also produce incorrect genotype calls. SNPs which may have been incorrectly clustered can be identified by low GenTrain score, cluster separation score and call frequency. In order to identify and resolve these irregularities, the SNPs were filtered using the 'filter rows' function, which can filter SNPs based on the many parameters each SNP has associated to it in the SNP Table. A filter was placed on the SNP Table for each of these parameters using the logic function ("GenTrain score < 0.5") to present only those which fulfil the criteria (n=3277), ("Call Freq < 0.5") retains 2765 SNPs and ("Cluster Sep < 0.5") filters to retain 24228 SNPs.

The SNP Table was then sorted by each of the parameters and those which observably do not follow the expected format of clusters of correctly genotyped SNPs were manipulated to do so by either re-identifying an incorrectly identified cluster as AA, AB or BB or altering the cluster bounds to prevent samples from being incorrectly called.

Low GenTrain score errors can be observed with SNPs where clusters form but not all samples that would fit within the bounds are called; to resolve this, the cluster bound ovals are reshaped or moved further apart to allow calling of all samples within the

cluster. An example of this is given in Figure 1A, where rs181988691 was edited to include all samples in the AA homozygous major cluster.

Errors in call frequency were identified by a proportion of calls having low R values due to one of the probes failing or the cluster oval failing to call all samples due to a large spread of calls with a range of Theta values. This is shown in Figure 1B as the AA homozygous major oval in rs810810 were changed to include all samples in the cluster and exclude the samples with low normalised R scores.

Adjusting the cluster ovals for SNPs with low cluster separation scores can resolve many uncalled samples. The main errors found with these SNPs are found to be due to clusters spread and too close together, fitting the ovals to maintain all the calls for the correct cluster and allowing the ovals of other clusters to contain the spread of samples within its bounds increases cluster separation scores. Figure 1C shows this with the example of rs12906911, where the AA homozygous major cluster is changed to include all samples within the cluster and the ovals for the AB heterozygous and BB homozygous minor were reshaped to include all samples and remove overlap of the cluster ranges.

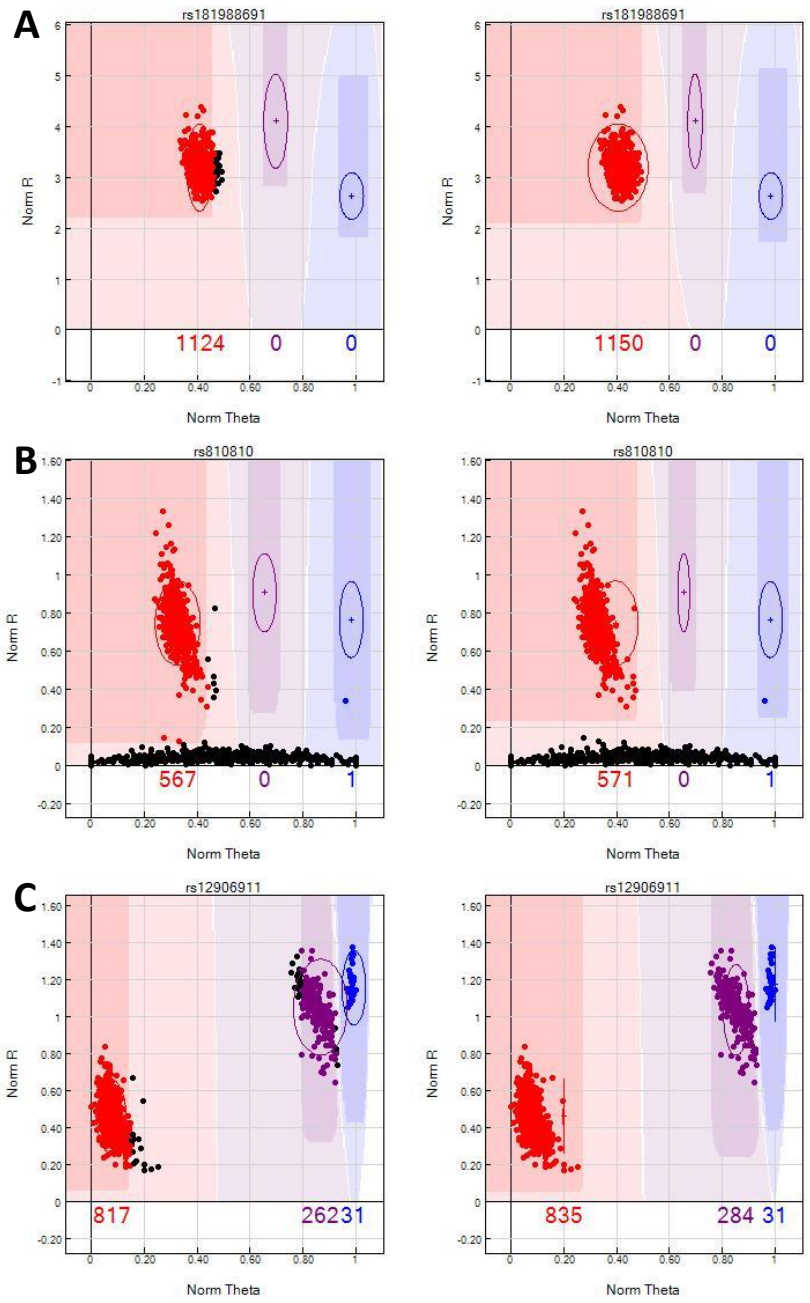


Figure 1: Examples of manual re-clustering in GenomeStudio

The figure gives examples of cases where manually adjusting the cluster ovals of SNPs improved their calling capacity beyond the cluster algorithm. **A.** The cluster oval for the AA homozygous major allele (red) is extended horizontally to capture all the uncalled samples, and the AB heterozygous allele (purple) is reduced horizontally to allow for the samples to fall within the correct cluster region. **B.** The cluster oval for the AA homozygous major allele (red) is raised vertically and expanded horizontally to capture the samples in the cluster whilst avoiding the samples with normalised R scores of <0.20. **C.** The cluster oval for the AA homozygous major allele (red) is moved horizontally to the edge of the cluster and reduced horizontally to capture all the samples in the cluster but maintaining a small cluster region respectively; the BB homozygous minor allele (blue) and AB heterozygous allele (purple) are also decreased horizontally to reduce the cluster region whilst maintaining the samples called and minimising the cluster regions.

2.2.3.3 Non-autosomal SNPs

Chromosome X and Y SNPs are assessed differently and separately to autosomal variants due to the expected nature of calls. As males have one of each chromosome, they cannot be heterozygous for either variant or as females have two X chromosomes, they are not expected to call for Y chromosome SNPs; the mitochondrial (MT) SNPs are also assessed similar to Y chromosome SNPs, as they were expected to form two clusters, with no heterozygotes. The pseudo autosomal regions (PAR) are identified and assessed as autosomal, as they are present at the ends of the X and Y chromosomes which are common to males and females.

The Samples Table holds information regarding all samples, Gender can be estimated based on the current genotypes of individuals. Samples can also be marked with an associated colour to identify in cluster plots, this allows male and female samples and the clusters they fall into.

The PAR is identified according to the Genome Reference Consortium as chrX:60,001-2,699,520, chrX:154,931,044-155,260,560, chrY:10,001-2,649,520 and chrY:59,034,050-59,363,566. Each region was filtered in the SNP Table to list all variants on the NeuroChip within these regions, the first of which was filtered using ("Chr = X" AND "Position > 60000" AND "Position < 2649521"). The list of SNPs within this region were collected using the 'Export displayed data to a file' tool and was repeated for all regions. The list of SNP Names were compiled into a table as the first column with a second column titled PAR with each SNP being given an arbitrary value of 1. The table was imported into GenomeStudio, which adds an additional column in the SNP Table marking all PAR SNPs as 1 and all other SNPs as 0; this is useful for filtering processes in later QC.

Chromosome X SNPs are filtered in the SNP Table using the ("Chr = X" AND "PAR != 1") logic function and retains 9832 SNPs. The SNPs are sorted in descending order by AB Freq, to identify males who have been called as heterozygotes, all SNPs with a large majority of male heterozygote calls are zeroed. Some SNPs may have called males at heterozygotes instead of homozygotes by the algorithm, the cluster oval were moved and reshaped to only include the females in the heterozygote cluster. Figure 2A shows all male samples marked in light blue for rs7050856, there are expected to be no male heterozygotes, so the cluster ovals were reshaped to call only females in the AB heterozygote cluster and increase the region of coverage for the BB homozygous minor cluster to include all male samples.

SNPs which tagged to chromosome Y were first filtered using GenomeStudio filter rows function, using logic functions the Chr column was set to ("Chr = Y" AND "PAR != 1) and the SNP Table was reduced to only show those variants (n=1899); all SNPs were selected, and their properties were changed so that the number of expected clusters was set to 2. The SNP Table was sorted descending by Call Freq, the SNPs with high frequency were selected and zeroed as they were calling females and should not be. The expected call frequency for chromosome Y SNPs should be the proportion of male samples in the genotyped cohort, the cluster plots were edited to show this by moving the cluster ovals to no longer call females or correctly identifying the homozygous allele clusters. Both examples are shown in Figure 2 where female samples are marked in pink. In M2B there are some females called by all three cluster groups of rs7067378, the adjustment reduces all AB heterozygote to zero and excludes females called in AA and BB homozygous clusters. M2C reflects an example of misidentification of a cluster for rs9785830, the samples with moderately high normalised Theta scores are selected and redefined as the BB homozygous minor cluster which resolves the SNP to the expected clustering.

Mitochondrial SNPs were filtered using the ("Chr = MT") filter, which presented 160 SNPs. These were selected and their properties were also set to expect 2 clusters. SNPs with heterozygote calls were zeroed; cluster ovals were moved for SNPs with clusters which may include samples that are called but not part of the genotype cluster, and more likely to be uncalled samples. Figure 2D for the SNP named 200610-105 shows two separate clusters called for AA homozygous major allele, it could be expected that the samples with low normalised Theta and R scores may have failed to be correctly called, it is more accurate to adjust the cluster oval to no longer call these samples.

2.2.3.4 Rare variants

Rare SNPs were manually checked by filtering the SNPs by minor allele frequency and call frequency to identify SNPs with low minor allele frequency and with very few uncalled samples ("MAF < 0.01" AND "Call frequency < 0.9999"). Some of the 55743 SNPs filtered by these filters may show samples with apparent heterozygotes and minor allele homozygotes which haven't been called, the oval can be moved to include the sample and therefore identify the rare call.

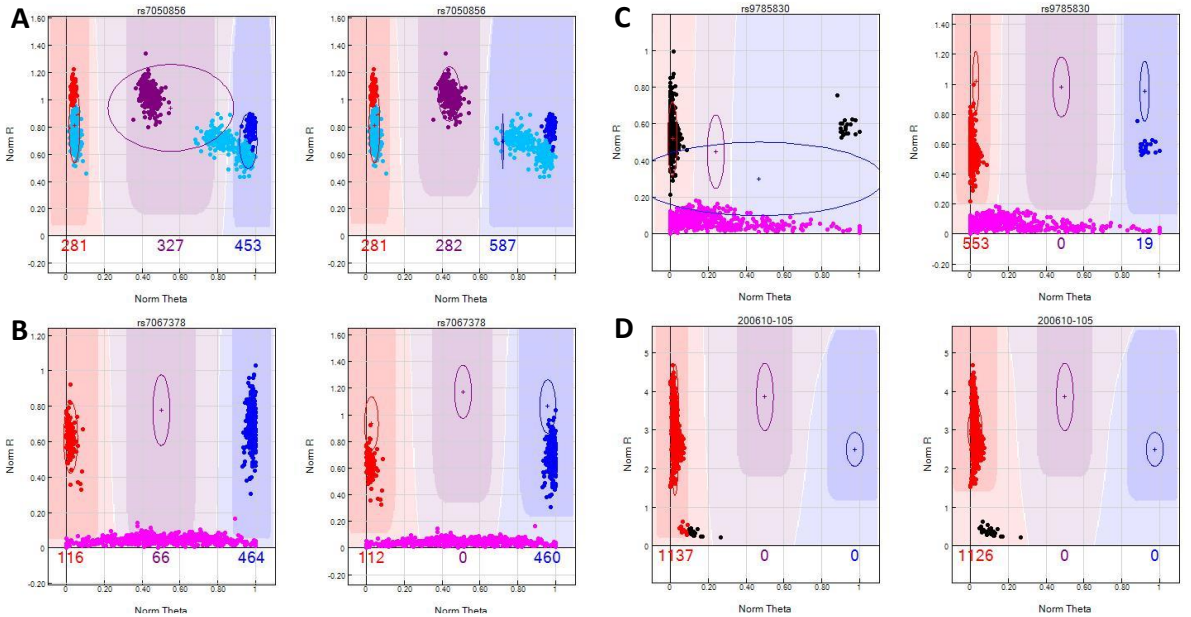


Figure 2: Examples of manual re-clustering autosomal SNPs in GenomeStudio

The figure depicts some cases where manual adjustment was required for non-autosomal SNPs. In these scenarios, the Males were marked with a sample colour of light blue and Females were marked with a sample colour of pink. **A.** The samples in the middle cluster were expected to not call males and therefore were selected and re-defined as the AB heterozygote allele (purple), the cluster oval for the BB homozygous minor allele (blue) was shifted to the left and shrunk horizontally to capture all samples whilst reducing the cluster region. **B.** In this scenario, the female samples were not expected to be called; all the cluster ovals were all raised vertically to no longer include the female samples in the cluster regions. **C.** The cluster algorithm had grossly failed to correctly define the clusters; the apparent samples within the AA homozygous major cluster were selected and redefined as such, and all cluster ovals were raised vertically to exclude the female samples from the cluster regions. **D.** The cluster algorithm has identified two clusters which have similar normalised theta scores and included them both in the cluster region; it is likely the samples with low normalised theta and R scores have failed to be correctly called and should therefore be excluded from the AA homozygous major allele (red) cluster oval, which is done by raising the oval vertically.

2.2.3.5 Alignment

The resulting dataset after quality control in GenomeStudio was exported using the PLINK report plug-in and produced the dataset in PLINK format (MAP/PED). The files were uploaded to a Linux server using Core FTP LE (version 2.2, Core FTP) for further QC using PLINK.

The PLINK files were first converted to binary files (BED/BIM/FAM) using PLINK --make-bed function. The binary ped file format allowed the dataset to be altered to reorganise genotype calls from allele zygosity to specific nucleotide change at the genotypic location.

Alignment of the binary ped file to the GRCh37/hg19 assembly was completed using the strand file CustomNeurochipHumanCore-24-v1_A-b37.strand and shell script update_build.sh, which were acquired from Rayner, W of the Wellcome Trust. The script updates the input datasets and outputs an aligned dataset in binary PLINK format.

Usage:

```
$ plink --file GenomeStudio --make-bed --out NeuroChip

$ sh update_build.sh NeuroChip CustomNeurochipHumanCore-
24-v1_A-b37.strand NeuroChip_aligned
```

--file = root of input file in PLINK format

--out = name of output file

2.2.3.6 Sample call rate

All blank controls and samples which failed and were therefore zeroed in GenomeStudio were exported within the PLINK report. These samples were removed from the dataset using the PLINK --mind function with a sample missingness set to 0.1, and the resulting binary PLINK file is produced using the --make-bed function.

Usage:

```
$ plink --bfile NeuroChip_aligned --mind 0.1 --make-bed --out NeuroChip_callfreq
```

--bfile = root of input file in binary PLINK format

2.2.3.6 Gender mismatch

PLINK uses individuals' chromosome X SNP genotypes to determine the sex of the sample. The function --check-sex tests all variants and calculates the inbreeding co-efficient, an F-statistic. The inferred sex of samples can be compared with the reported sex in clinical files or Gender estimation in GenomeStudio. Individuals and their F-statistic are reported in the output file, highlighting any discrepancies as errors. These discrepancies were further evaluated and those individuals whose sex was incorrectly reported and correctly calculated were rewritten using the --impute-sex function and a replacement dataset is produced with updates sex calls using the --make-bed function.

Usage:

```
$ plink --bfile NeuroChip_callfreq --check-sex --out NeuroChip

$ plink --bfile NeuroChip_callfreq --impute-sex --make-bed --out NeuroChip_sexcheck
```

--out = name of output sexcheck file

2.2.3.7 Relatedness

Related samples who share up to third-degree relation affect allele frequencies from heritability, which can bias the association of variants with disease status. Relatedness was calculated with PLINK using only common (MAF>0.1), independent, autosomal SNPs. These SNPs were isolated by pruning the dataset using independent pairwise analysis (kb window size=50; step size=5; linkage disequilibrium r^2=0.2). The non-

autosomal SNPs were excluded using the --exclude function from list of all corresponding variants in a non_autosomal.txt derived from the BIM file. Independent pairwise was completed using the --indep-pairwise function which produced output files containing lists of variants which would be pruned in and out according to the parameters, the SNPs which passed pruning were extracted from the dataset using the --extract function and relatedness was calculated using the --genome function.

The genome file paired individuals and calculated identity-by-decent and produced PI_HAT scores, which ranged from 1 to 0 depending on degree of relatedness. Third-degree relatives have a PI_HAT score around 0.125; under those conditions only one individual per related pair would be removed to not bias the dataset. Samples selected for removal were listed in remove_related.txt and removed from the dataset using the –remove function and a new dataset was produced using the –make-bed function.

Usage:

```
$ plink --bfile NeuroChip_sexcheck --maf 0.1 --exclude
non_autosomal.txt --indep-pairwise 50 5 0.2 --out
NeuroChip_indepSNP

$ plink --bfile NeuroChip_sexcheck --extract
NeuroChip_indepSNP.prune.in --genome --out NeuroChip

$ plink --bfile NeuroChip_sexcheck --remove
remove_related.txt --make-bed --out NeuroChip_relatedness
```

--maf	= minor allele frequency threshold
--indep-pairwise	= parameters for pruning: window, step size and r2 threshold
--extract	= input file for extracting variants for relatedness test
--remove	= input file of list of individuals to remove

2.2.3.8 Ancestry

Ancestry was calculated using the --pca function which produced eigenvector and eigenvalue files; these were used in principal component analysis. Genotype calls for

ancestry informative markers were identified from numerous online sources and those present on the NeuroChip were extracted using the --extract and –make-bed function. The resulting AIM dataset was used for ancestry testing.

The top eigenvectors were plotted against each other in a scatter plot, the samples were organised into clusters. Mean and standard deviation (SD) of each sample was calculated for each PCA, individuals with values greater than 6 SDs from the mean were listed in the failed_ancestry.txt file and removed using --remove and --make-bed function.

Usage:

```
$ plink --bfile NeuroChip_relatedness --extract AIMs.txt
--make-bed --out NeuroChip_AIMs

$ plink --bfile NeuroChip_AIMs --pca 10 --out NeuroChip

$ plink --bfile NeuroChip_relatedness --remove
failed_ancestry.txt --make-bed --out NeuroChip_ancestry
```

--extract = input list of variants for principal component analysis

--pca = maximum number of components to test the dataset on

2.2.3.9 Hardy-Weinberg equilibrium

Allele frequencies are dependent on many assumptions which would allow SNPs to fall within Hardy-Weinberg equilibrium (HWE) and therefore be correctly genotyped. These assumptions which lead to variants being within HWE include the lack of selection pressure against variants, no mutations leading to this observed variation and the tested variants existing within large populations. HWE was tested based on the allele frequencies in non-diseased samples for common SNPs (MAF>0.05) and the significance of the variant existing in this equilibrium in the population group is calculated with a given p-value. The --hardy function in PLINK is used to calculate these significances, producing an output file with significance values for cases, controls, and all individuals.

A threshold was derived using Bonferroni correction for multiple testing from an expected p-value ($p<0.05$) divided by the number of SNPs tested ($p<1.032\times10^{-7}$); SNPs with a p-value more significant than this threshold in unaffected individuals were listed in the hwe_SNPs.txt file and removed using the --exclude function and a resulting dataset was produced with the --make-bed function.

Usage:

```
$ plink --bfile NeuroChip_ancestry --maf 0.05 --hardy --
out NeuroChip

$ plink --bfile NeuroChip_ancestry --exclude hwe_SNPs.txt
--make-bed --out NeuroChip_hwe
```

2.2.3.10 Heterozygosity

Heterozygosity is the measure of genetic variability within a population, there are many conditions which could lead to excessively low or high heterozygosity being observed amongst the genotyped cohort; high levels of heterozygosity can result from cross-contamination and low heterozygosity could be an indicator of inbreeding within the population, these factors may have not been identified in earlier quality checks of relatedness. Heterozygosity can be calculated in PLINK using the --het function based on common, independent SNPs identified during the pruning stage for relatedness testing and the --extract function.

The resulting file indicated levels of heterozygosity for individuals, mean heterozygosity and SDs were calculated and those who deviated by more than 3SDs were listed in remove_het.txt and removed using the --remove function. The final output dataset of this stage was produced using the --make-bed function, the final quality-controlled dataset was to be used in further genetic testing and polygenic risk scoring.

Usage:

```
$ plink --bfile NeuroChip_hwe --extract
NeuroChip_indepSNP.prune.in --het --out NeuroChip

$ plink --bfile NeuroChipF --remove remove_het.txt --
make-bed --out final_NeuroChip
```

2.2.4 Imputation pipeline

2.2.4.1 Pre-imputation quality control

When running a job on the imputation server, input data is required to be in zipped VCF, with separate files for each chromosome; multiple samples can also be imputed at the same time. The input files were prepared in PLINK using the --chr function to isolate SNPs and the --recode function to produce VCF files. VCF files were then sorted and compressed using vcf-sort and the -c function of bgzip, respectively.

Usage:

```
$ plink --bfile final_NeuroChip --chr 1 --recode vcf --
out chr1
$ plink --bfile final_NeuroChip --chr 2 --recode vcf --
out chr2
$ plink --bfile final_NeuroChip --chr 3 --recode vcf --
out chr3
$ plink --bfile final_NeuroChip --chr 4 --recode vcf --
out chr4
$ plink --bfile final_NeuroChip --chr 5 --recode vcf --
out chr5
$ plink --bfile final_NeuroChip --chr 6 --recode vcf --
out chr6
$ plink --bfile final_NeuroChip --chr 7 --recode vcf --
out chr7
$ plink --bfile final_NeuroChip --chr 8 --recode vcf --
out chr8
$ plink --bfile final_NeuroChip --chr 9 --recode vcf --
out chr9
$ plink --bfile final_NeuroChip --chr 10 --recode vcf --
out chr10
$ plink --bfile final_NeuroChip --chr 11 --recode vcf --
out chr11
$ plink --bfile final_NeuroChip --chr 12 --recode vcf --
out chr12
$ plink --bfile final_NeuroChip --chr 13 --recode vcf --
out chr13
$ plink --bfile final_NeuroChip --chr 14 --recode vcf --
out chr14
$ plink --bfile final_NeuroChip --chr 15 --recode vcf --
out chr15
$ plink --bfile final_NeuroChip --chr 16 --recode vcf --
out chr16
$ plink --bfile final_NeuroChip --chr 17 --recode vcf --
out chr17
$ plink --bfile final_NeuroChip --chr 18 --recode vcf --
out chr18
$ plink --bfile final_NeuroChip --chr 19 --recode vcf --
```

```
out chr19
$ plink --bfile final_NeuroChip --chr 20 --recode vcf --
out chr20
$ plink --bfile final_NeuroChip --chr 21 --recode vcf --
out chr21
$ plink --bfile final_NeuroChip --chr 22 --recode vcf --
out chr22

$ vcf-sort chr1.vcf | bgzip -c > chr1.vcf.gz
$ vcf-sort chr2.vcf | bgzip -c > chr2.vcf.gz
$ vcf-sort chr3.vcf | bgzip -c > chr3.vcf.gz
$ vcf-sort chr4.vcf | bgzip -c > chr4.vcf.gz
$ vcf-sort chr5.vcf | bgzip -c > chr5.vcf.gz
$ vcf-sort chr6.vcf | bgzip -c > chr6.vcf.gz
$ vcf-sort chr7.vcf | bgzip -c > chr7.vcf.gz
$ vcf-sort chr8.vcf | bgzip -c > chr8.vcf.gz
$ vcf-sort chr9.vcf | bgzip -c > chr9.vcf.gz
$ vcf-sort chr10.vcf | bgzip -c > chr10.vcf.gz
$ vcf-sort chr11.vcf | bgzip -c > chr11.vcf.gz
$ vcf-sort chr12.vcf | bgzip -c > chr12.vcf.gz
$ vcf-sort chr13.vcf | bgzip -c > chr13.vcf.gz
$ vcf-sort chr14.vcf | bgzip -c > chr14.vcf.gz
$ vcf-sort chr15.vcf | bgzip -c > chr15.vcf.gz
$ vcf-sort chr16.vcf | bgzip -c > chr16.vcf.gz
$ vcf-sort chr17.vcf | bgzip -c > chr17.vcf.gz
$ vcf-sort chr18.vcf | bgzip -c > chr18.vcf.gz
$ vcf-sort chr19.vcf | bgzip -c > chr19.vcf.gz
$ vcf-sort chr20.vcf | bgzip -c > chr20.vcf.gz
$ vcf-sort chr21.vcf | bgzip -c > chr21.vcf.gz
$ vcf-sort chr22.vcf | bgzip -c > chr22.vcf.gz
```

--chr = chromosome number by which to filter (1-26)

--recode = output format, default is PLINK file

2.2.4.2 Imputation

Input files are uploaded to the server and selection criteria is given based on the desired reference panel, the assembly of input data, whether an LD-based r^2 filter should be used and to what level it should be set, the software which should be used for phasing, identification of the input sample population and the desired output.

Usage:

```
Run - Genotype Imputation (Minimac4)
Name: NeuroChip
Reference Panel: HRC r1.1 2016 (GRCh37/hg19)
Input Files: chr1.vcf.gz chr2.vcf.gz chr3.vcf.gz
chr4.vcf.gz chr5.vcf.gz chr6.vcf.gz chr7.vcf.gz
chr8.vcf.gz chr9.vcf.gz chr10.vcf.gz chr11.vcf.gz
chr12.vcf.gz chr13.vcf.gz chr14.vcf.gz chr15.vcf.gz
chr16.vcf.gz chr17.vcf.gz chr18.vcf.gz chr19.vcf.gz
chr20.vcf.gz chr21.vcf.gz chr22.vcf.gz
Array Build: GRCh37/hg19
rsq Filter: 0.001
Phasing: Eagle v2.4 (phased output)
Population: EUR
Mode: Quality Control & Imputation
```

2.2.4.3 Quality control

The resulting output data is downloaded as compressed VCF files which can be transferred back onto the workspace for quality control of imputation. Firstly, the single chromosome VCF files were merged into a single VCF file using the concat command of bcftools; the --allow-overlaps was used, and --rm-dups was set to all function to avoid loss of variants during merging and prevent duplicate SNPs being introduced, the --output-type function was set to z to produce an compressed VCF as the output. Imputed SNPs are filtered by INFO score, which ranges from 0-1 based on the statistical accuracy of the call; this can be completed using the filter command of bcftools.

Usage:

```
$ bcftools concat --allow-overlaps --rm-dups all --
output-type z \
chr1.vcf.gz chr2.vcf.gz chr3.vcf.gz chr4.vcf.gz
chr5.vcf.gz \
chr6.vcf.gz chr7.vcf.gz chr8.vcf.gz chr9.vcf.gz
chr10.vcf.gz \
chr11.vcf.gz chr12.vcf.gz chr13.vcf.gz chr14.vcf.gz
chr15.vcf.gz \
chr16.vcf.gz chr17.vcf.gz chr18.vcf.gz chr19.vcf.gz
chr20.vcf.gz \
chr21.vcf.gz chr22.vcf.gz > imputed_NeuroChip.vcf.gz

$ bcftools view --include 'INFO>0.8' --output-type z
imputed_NeuroChip.vcf.gz > imputedQC_NeuroChip.vcf.gz
```

--rm-dups = variant type to be managed (snps/indels/both/all/none)

--output-type = format of output file (b/u/z/v)

--include = expression by which to filter variants

2.2.5 Polygenic risk scoring

Genetic risk modelling is completed in many studies of genetic data associated with disease. Risk of disease is expected to correlate with the SNPs individuals harbour based on their association with disease phenotype; SNP effects can be measured as dosage, the number of risk alleles present or weighted by their observed association

in larger case-control cohorts. The summary statistics of GWAS highlight the SNPs most associated with disease where the effect is estimated as a beta value or odds ratio. PRS is often used to analyse a target group based on the observed effects in a discovery group, the output provides arbitrary values of risk individuals in the target cohort has for the disease phenotype.

2.2.5.1 PRSice

PRSice software was used on several occasions throughout the study, where the most recently released, stable version was using for final analysis. In this example scenario, PRSice-2 was used to identify the best model of disease amongst AD cases and cognitively healthy controls with the exclusion of a 500kb region surrounding the *APOE* gene; this was completed using the --keep function based on the samples listed in "AD_Control.txt" and --exclude range function with "APOE_locus.txt" file in PLINK, where a new dataset was produced using the --make-bed function. PRSice-2 was run using the GWAS summary statistics from the IGAP consortium, the base file used was "IGAP_stage_1.txt". PRSice-2 uses PLINK to calculate the linkage disequilibrium (LD) between SNPs, estimated from the observed linkage within a cohort of 503 European individuals genotyped as part of the 1000 Genomes project, "1000G_EUR". The clumping algorithm identifies the most significant SNP (index) according to the base dataset in a sliding window and removes SNPs in LD with the index SNP above a given r^2 value to reduce the occurrence of type I errors. The r^2 value signifies the inheritance correlated between the SNPs and false positives are introduced as SNPs in LD with a variant with known association with the disease will appear as a novel signal. The effects of clumping algorithm can be seen in Figure 3.

Usage:

```
$ plink --bfile final_NeuroChip --exclude range
APOE_locus.txt --keep AD_Control.txt --make-bed --out
NeuroChip_noAPOE_AD_Control

$ Rscript PRSice.R \
--dir . \
--prsice ./PRSice \
--base IGAP_stage_1.txt \
--target NeuroChip_noAPOE_AD_Control \
--ld 1000G_EUR \
--thread max \
--print-snp T \
--stat BETA \
--binary-target T \
--no-clump F \
--clump-kb 250\
--clump-r2 0.2 \
--clump-p 1 \
--perm 10000 \
--lower 0 \
--interval 0.000001 \
--upper 1 \
--all-score T \
--out NeuroChip_noAPOE_AD_Control
```

--exclude = exclusion format, default is individual variant name

Rscript = root to R script for PRSice-2 calculation

--dir = directory from which to work

--prsice = root to PRSice software

--base = root to base dataset

--target = root to target dataset

--ld = root to dataset for LD calculation

--thread = number of threads to use, or set as max

--print-snp = produce an output file of SNPs used in model (T/F)

--stat = identify effect statistic (BETA/OR)

--binary-target = specify use of case-control data (T/F)

--no-clump = specify exclusion of LD based clumping (T/F)

--clump-kb = specify window to be used in clumping algorithm

--clump-r2 = specify r^2 threshold for LD

--clump-p = specify p-value threshold for clumping

--perm = specify number of permutations of best model for empirical p-value

--lower = specify lower significance level to be tested

--interval = specify intervals by which to increase in modelling

--upper = specify upper significance level to be tested

--all-score = produce an output of scores for all models tested

--out = output file prefix

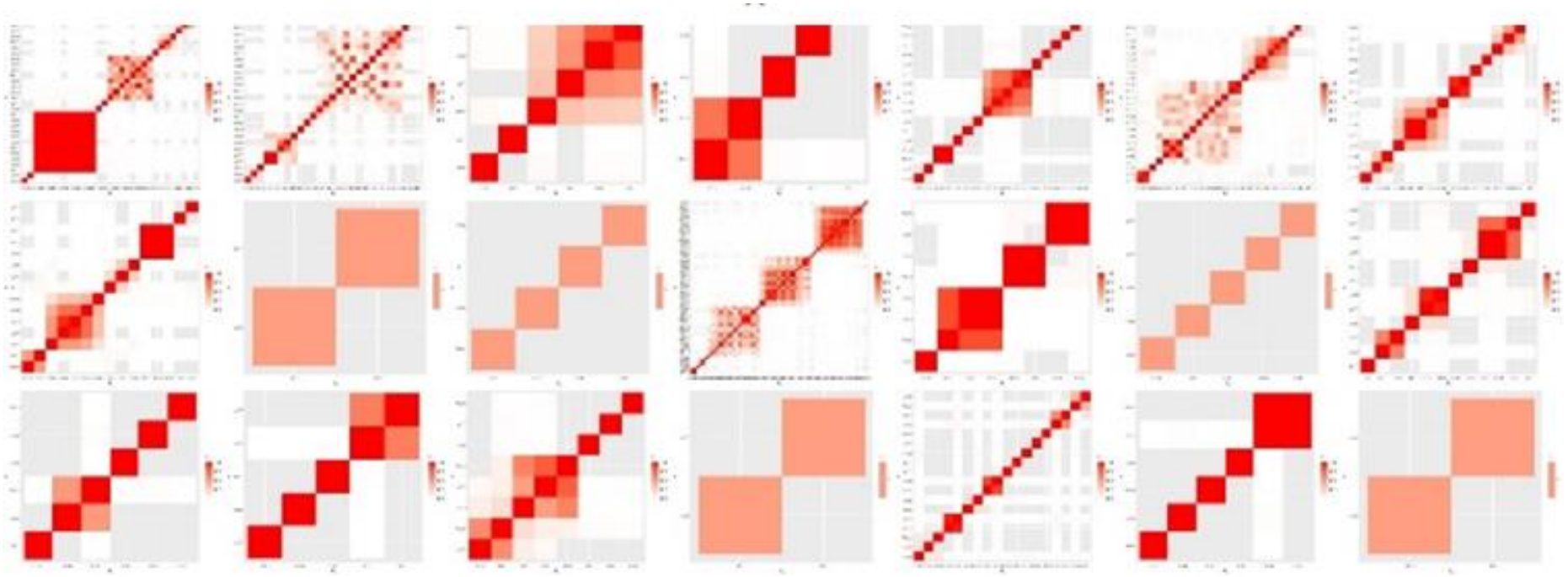


Figure 3: Linkage disequilibrium heat map of PRSice clumping algorithm

The figure identifies the observed relationship between SNPs when not subjected to LD based clumping. The red boxes represent the R^2 value of linkage between SNPs and a continuous diagonal shows the relationship a SNP has with itself ($R^2 = 1$). Each sub-figure represents each chromosome, (chr 1-20, 22). There is no figure for chromosome 21 as there were no SNPs from this chromosome present in the PRSice best model. Each extended region of red boxes shows there is linkage disequilibrium between those SNPs, and they are representative of a single locus. The PRSice clumping algorithm works by reducing the effect at each locus from being represented by multiple SNPs to just a single SNP and thus prevent over-fitting of the model.

3 Polygenic risk scoring of late-onset Alzheimer's disease

3.1 Perspective

Polygenic risk scoring was conducted on a cohort of late-onset Alzheimer's disease (LOAD) cases and neurologically healthy controls from the Brains for Dementia Research (BDR) resource (target dataset), based on summary statistics from the International Genomics of Alzheimer's disease Project (IGAP) meta-analysis (base dataset), using the most current stable version of PRSice-2 software.

PRSice-2 was used to generate PRS on samples genotyped on the NeuroChip array, the second generation of the neuro-specific array to include variants implicated in neurodegenerative diseases.

The Brains for Dementia Research (BDR) resource recruited samples; clinical information was provided with samples, including post-mortem pathological confirmation of disease status. Genotyped samples were quality controlled using Illumina GenomeStudio and PLINK and later imputed using the Michigan Imputation Server Minimac4 software.

Previous analysis of this cohort was conducted using an earlier version of PRSice software, PRSice-1.25, on samples from the first two batches of the BDR cohort. The results showed significant differences between mean PRS between LOAD cases (n=302) and controls (n=137), with a predictive ability (area under the ROC curve; AUC) for distinguishing between groups of 73% (Chaudhury et al., 2019). When samples were sorted into deciles of increasing risk and a greater proportion of cases were observed at higher deciles with the opposite being seen in controls, confirming the association between genetic risk of disease and likelihood of developing AD.

The primary aim of this study was to determine the genetic risk of AD in the entire BDR collection (to date) using updated PRS analysis software.

3.2 Samples

DNA was isolated from blood and/or brain tissue from all BDR samples (n=1172). Lab procedures were carried out by Patel, T., following methods as described (Section 2.1). All samples were genotyped using the NeuroChip array. Quality control of the genotyping data was completed using GenomeStudio v2 and PLINK (Section 2.3.3). Samples were checked for call rate, relatedness, diversion from European ancestry, and heterozygosity; SNPs were controlled for call rate and adherence to Hardy-Weinberg equilibrium. By March 2019 there were 217 cognitively healthy controls and 358 LOAD cases (Table 2), genotyped for 477,720 SNPs; additionally, 185 samples of other phenotypes and 437 living, undiagnosed samples were genotyped on the NeuroChip. Both SNPs associated with *APOE* status and additional SNPs identified to be associated with AD from recent GWAS were not included in the NeuroChip design; these variants, outlined in Table 3, were genotyped for all deceased samples using TaqMan assays, carried out by Brookes, K., following methods described (Section 2.1.4).

Phenotype	N	Age at death (SD)	Females (%)	*APOE* ε4+ (%)	*APOE* ε4ε4 (%)
Control	217	49.7 (44.2)	115 (53.0)	35 (16.1)	2 (0.9)
Late-onset AD	358	83.2 (8.5)	173 (48.3)	231 (64.5)	49 (13.7)

Table 2: Demographics of BDR samples

Individuals, categorised by disease status, were recruited from across the UK: from universities in Bristol, Manchester, Newcastle, Oxford, and King's College London. Age at death (together with standard deviation) was obtained from clinical information. The number and percentage of females in each group is also shown as well as the number and percentage of individuals with at least one *APOE* ε4 allele and additionally those who were ε4ε4 homozygotes.

SNP	Chromosome	Base Position	Alleles	GWAS gene	IGAP P-value	Genotyping Call Rate
rs429358	19	45,411,941	T>C	*APOE*	6.70×10^{-536}	99.7%
rs7412	19	45,412,079	C>T	*APOE*	1.23×10^{-22}	99.7%
rs9271192	6	32,578,530	C>A	*HLA-DRB1*	1.57×10^{-8}	97.4%
rs6656401	1	207,692,049	A>G	*CR1*	7.73×10^{-15}	95.7%
rs10948363	6	47,487,762	A>G	*CD2AP*	3.05×10^{-8}	97.1%
rs10838725	11	47,557,871	T>C	*CELF1*	6.73×10^{-6}	96.2%
rs35349669	2	234,068,476	C>T	*INPP5D*	9.58×10^{-5}	97.5%
rs28834970	8	27,195,121	T>C	*PTK2B*	3.27×10^{-9}	97.3%
rs11218343	11	121,435,587	T>C	*SORL1*	4.98×10^{-11}	96.0%
rs983392	11	59,923,508	A>G	*MS4A2*	2.76×10^{-11}	97.5%
rs9331896	8	27,467,686	G>T	*CLU*	9.63×10^{-17}	95.2%

Table 3: Additional GWAS SNPs genotyped using TaqMan assays

TaqMan assays were used to genotype all BDR samples for the two variants used to identify *APOE* ε status and additional GWAS hits absent from the NeuroChip array. The variants are listed by their rsID; chromosome, base pair and the observed alleles are given alongside the gene locus they were initially identified within. The genotyping call rate is given for each SNP from testing all BDR samples (n=1296).

Imputation using the Michigan imputation server was completed on the entire BDR cohort to identify additional variants based on the observed genotypes. The genotyped dataset underwent pre-imputation quality controls of the strand and position of alleles; their calls, assignments, and frequencies were checked to be consistent with the Haplotype Referencing Consortium. Imputation increased the number of SNPs accessible by PRSice on the BDR cohort (n=4,222,576).

3.3 Polygenic risk scoring

Primary PRS analysis (Figure 4) tested 358 LOAD cases and 217 controls using default PRSice-2 parameters and additional PRS quality control of the base and target dataset. PRSice quality control of the target dataset included standard cut-off thresholds for minor allele frequency (MAF ≤ 0.01), Hardy-Weinberg equilibrium (HWE ≤ 1×10^{-6}), genotyping call rate (GENO ≤ 0.01) and sample call rate (MIND ≤ 0.01). Ambiguous or duplicate SNPs were removed, and mismatched SNPs were resolved in both the base and target datasets; following QC, the target dataset was reduced to 232,111 SNPs. The *APOE* locus, a window of 250kb either side of *APOE* (chr19:45,160,844-45,660,844; GRCh37/hg19), was excluded from the target dataset as a well-known AD hotspot; the SNPs which identify the *APOE* ε status would be reintroduced to the risk modelling at a later stage.

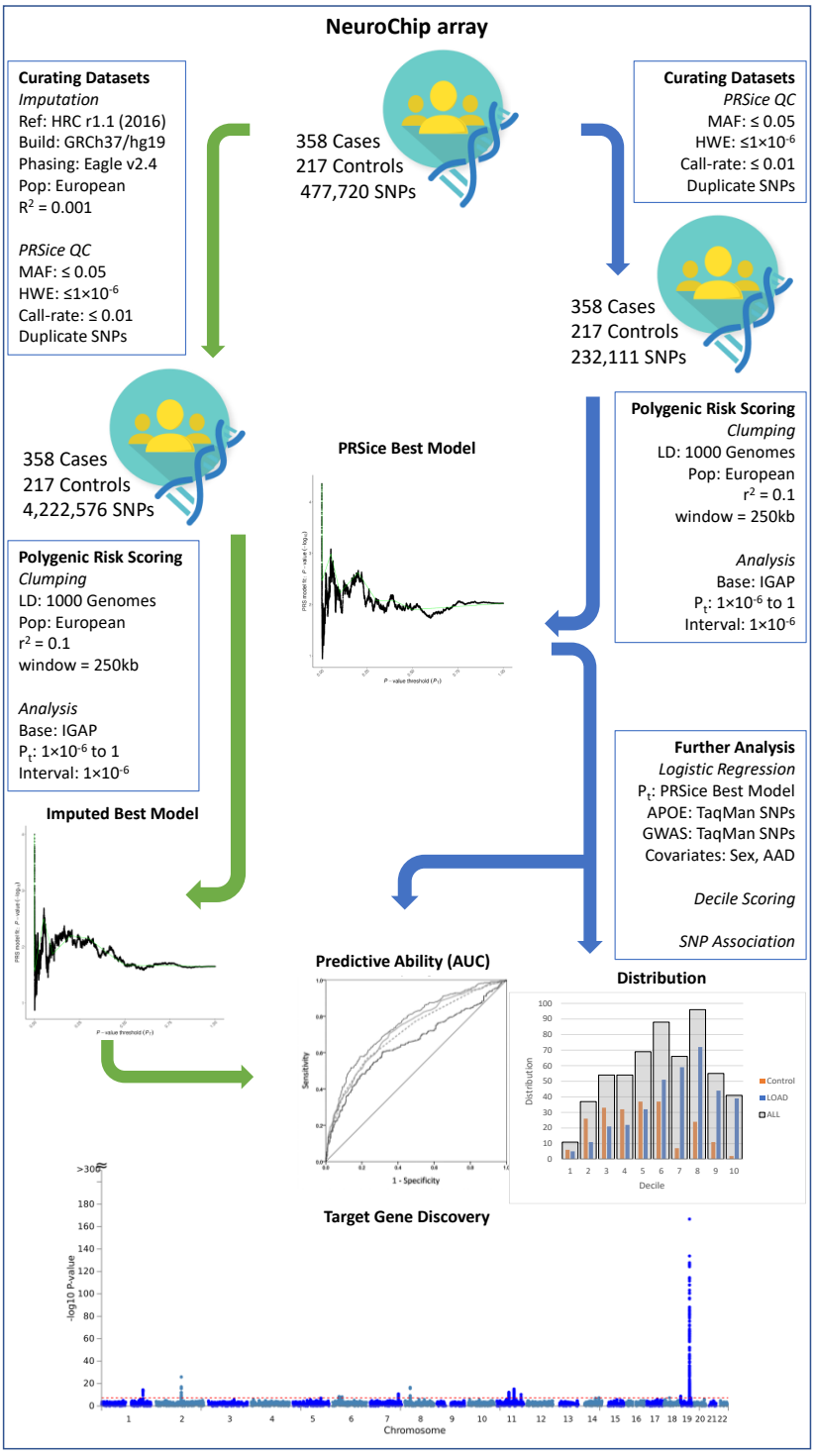


Figure 4: Analysis pipeline for PRS of LOAD with the NeuroChip

The figure outlines the process of generating polygenic risk scores from a target dataset. Primary analysis (blue) is run in parallel with secondary analysis, the use of imputation (green). Polygenic risk scores are generated, and covariates are introduced; the results are used to determine the predictive ability of the best PRS model and the measure the distribution of samples based on PRS. Further analysis includes the introduction of TaqMan assays of GWAS hits absent from the array to the genotyped model. The resulting models are checked for previously unidentified genes associated with AD.

The SNPs were then clumped based on their LD to reduce the number of potential false positives. The default clumping settings for PRSice-2 which were used included a 250kb sliding window and r^2 set as 0.1; the dataset was also tested without clumping to compare and evaluate its effect.

PRSice-2 then selects all SNPs present in the base and target dataset and runs logistic regression to model cases against controls based on the SNPs present within given significance thresholds; the threshold incrementally increases to include additional SNPs for the next model until all SNPs have been included. The results of the modelling are compared and then scored by Nagelkerke's R^2 to identify the best model fit. The lowest threshold set for PRSice-2 analysis started at SNPs with significance of association in the base dataset between $0 \leq p \leq 1\times10^{-6}$, increasing in increments of 1×10^{-6} as new SNPs are introduced up to P≤1.

3.3.1 Best model selection

PRSice-2 provided individual scores at each tested threshold and identified the best p-value threshold for modelling risk in the target cohort based on Nagelkerke's R^2 and the number of SNPs used in the model. The resulting models at various testing thresholds are shown in Table 4, with the PRSice output figures included as Figure 5. The most predictive model was identified with an Nagelkerke's R^2 value of 0.042, at a p-value threshold of 1.4×10^{-4} and consists of 134 SNPs. At this threshold, there were 152 additional SNPs which were clumped out by the PRSice algorithm.

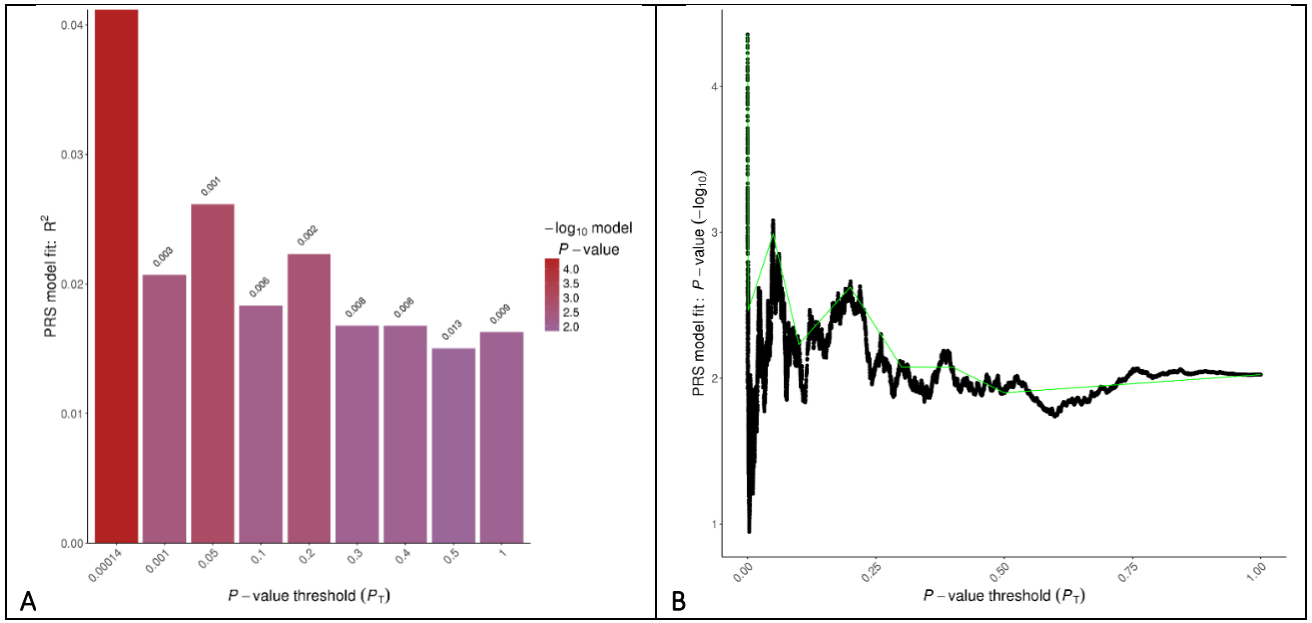


Figure 5: Output figures of PRSice modelling of LOAD with the NeuroChip

A. The bar plot presents the same thresholds in Table 4, identifying the model fitness expressed as Nagelkerke's R^2 at different stages. The results show the best model to be at 1.4×10^{-4}, which gradually decreases as more SNPs are introduced. **B.** The high-resolution plot identifies the p-value derived from each model at all tested thresholds, giving more details about the effect shown in A. The most prediction is found when using the most significantly associated variants according to the base dataset. However, there are additional peaks occurring as SNPs are introduced that are not as predictive as the first peak.

P-value threshold	N SNPs	Nagelkerke's R^2	Significance (P)
≤0.000001	16	0.0261	0.0011
≤0.00001	34	0.0182	0.00616
≤0.0001	110	0.0363	0.0363
≤0.00014	**134**	**0.0412**	**0.0000442**
≤0.001	507	0.0207	0.00344
≤0.01	2909	0.0107	0.0351
≤0.05	9212	0.0261	0.00103
≤0.1	14946	0.0183	0.00587
≤0.2	23176	0.0223	0.00238
≤0.3	30084	0.0168	0.00838
≤0.4	36219	0.0168	0.00842
≤0.5	41210	0.0150	0.0126
≤1.0	57047	0.0163	0.00946

Table 4: PRSice modelling of LOAD with the NeuroChip

The table outlines the results of PRSice modelling at p-value thresholds of significance according to the base dataset. At each threshold PRSice identifies the number of SNPs present in the target dataset at the threshold and calculates model fitness that can be determined based on the SNPs as Nagelkerke's R^2 and the associated p-value of the model. As the threshold increases, more SNPs are introduced to the model which affects the model predictiveness. The best model derived by PRSice is identified at a threshold of $P_t \leq 0.00014$, with the largest Nagelkerke's R^2 and the most significant p-value.

Mean individual PRS at the PRSice best model was calculated as 0.00139 with a standard deviation of 0.00234. Mean PRS and standard deviation for controls and LOAD cases was found as 0.000853±0.00232 and 0.00172±0.00243, respectively. Single factor ANOVA found variance for controls as 5.38×10^{-6} and LOAD cases 5.92×10^{-6}, and the significance of the variation to be 3.08×10^{-5}.

The best model was then tested with the re-introduction of SNPs associated with *APOE* ε status (rs429358 and rs7412) and running PRSice-2 with the inclusion of covariates for sex and age at death. The SNPs were introduced to the dataset using PLINK by extracting the SNPs which were present in the best model and merging with a dataset of calls for both SNPs according to TaqMan genotyping. Covariates were included in the PRSice script with the inclusion of a covariate file identifying each individual and stating their clinical sex and recorded age at death. Covariates testing was also replicated using binary logistic regression in SPSS (IBM) to validate the results and record an individual score which combined PRS with covariates. Receiver operating characteristic (ROC) curves and precision-recall curves (PRC) are produced from this value, and the area under the curve (AUC) is calculated as a measure of predictability from each. The resulting models were compared to the PRSice best model in Table 5.

Model	N SNPs	Nagelkerke's R^2	P	AUROC (%)	AUPRC (%)
PRSice best model	134	0.0418	4.42×10^{-5}	59.5%	70.8%
APOE	2	0.1954	1.28×10^{-17}	70.8%	79.9%
PRS (with *APOE*)	136	0.1995	2.76×10^{-17}	73.2%	81.3%
PRS + sex	136	0.2002	3.55×10^{-17}	73.2%	81.3%
PRS + age at death	136	0.2025	2.4×10^{-17}	73.3%	81.5%
PRS + sex + AAD	136	0.2028	2.51×10^{-16}	73.4%	81.5%

Table 5: Modelling PRS of LOAD with covariates

The best model by PRSice was isolated and tested with the inclusion of the SNPs associated with *APOE* ε status, rs429358 (ε4) and rs7412 (ε2). The *APOE* SNPs alone show greater significance and model disease status better than the best model, however, when combined the PRS model shows improved fitness than either alone. Clinical sex and age at death (AAD) were introduced to the PRS with *APOE* model independently and together to show the best model according to Nagelkerke's R2 to be PRS with sex and AAD combined. Predictive ability for all models was also calculated using area under the receiver operating characteristic curve (AUROC). Given the model is based on an unbalanced number of cases and controls, it was imperative to also test predictive ability using area under the precision-recall curve (AUPRC), which shows greater predictive ability than the former.

3.3.2 Association and risk prediction

Following the results of covariate analysis, prediction of disease status was most informative when including sex and age at death as covariates, therefore the resulting normalised scores for this model were used for downstream analyses. Mean score and standard deviation for controls was calculated as 0.529±0.171 and 0.680±0.176 for LOAD cases. Variance for controls and cases were calculated as 0.0291 and 0.0308 respectively and found to be significantly different by single factor ANOVA (p=3.87×10^{-22}).

Further statistical analysis included decile scoring, where all samples were ranked by score (PRS with sex and age at death) and distributed into ten tiers of increasing risk. The samples were then separated by disease classification to identify the proportion of cases and controls which appeared in each decile, given in Table 6, and presented in Figure 6. The results show cases and controls are present within all deciles; most controls populate the lower five deciles whilst cases are proportionally higher in the upper five deciles. Controls are observed to be highest in the fifth and sixth decile whilst the most cases are represented in the eighth decile.

Decile	1	2	3	4	5	6	7	8	9	10
All individuals	11	37	54	54	69	88	66	96	55	41
Controls	6	26	33	32	37	37	7	24	11	2
Late-onset AD	5	11	21	22	32	51	59	72	44	39

Table 6: Distribution of LOAD samples by decile scoring

The table shows the results of decile scoring on the BDR cohort LOAD cases and controls. The number of individuals, which fall within each decile, are given alongside the number of cases and controls. The entire cohort shows distribution into two peaks, with most samples observed in the sixth and then eighth deciles.

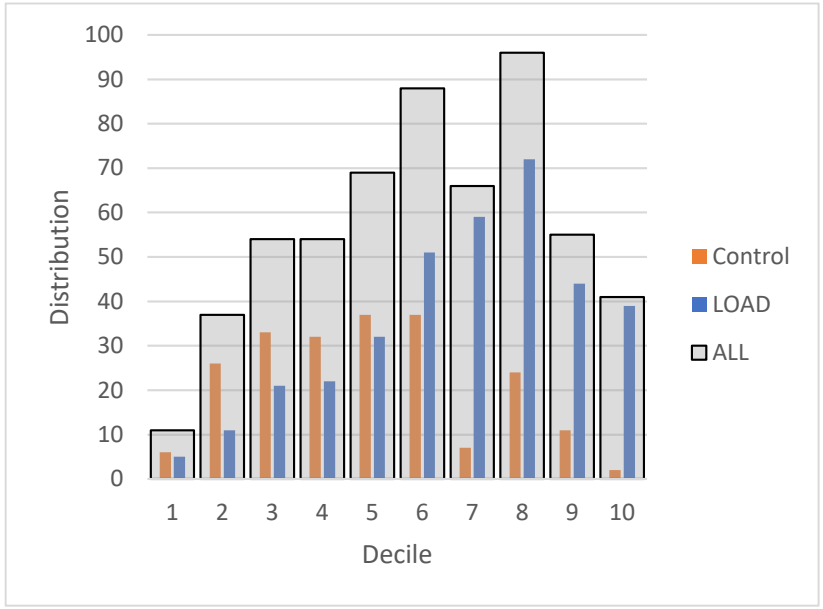


Figure 6: Distribution of LOAD PRS by decile scoring

The Figure shows the distribution of BDR cohort LOAD cases and controls when sorted into deciles. The distribution of all individuals are given as grey bars within which are the distribution of controls (orange) and LOAD cases (blue). The controls show normal distribution, with a peak around the fifth and sixth decile; the distribution of cases shows a right sided skew with more cases in higher deciles. This confirms a larger proportion of cases have higher risk classification of disease than controls and which can be discriminated using the PRS.

3.4 SNPs

PRSice variants included in the analysis (n=57,048) were ranked in order of significance in the base dataset. The rsIDs of the SNPs, which were present in the best model, were checked against online variant databases to identify their observed variation and the gene in which they are present. From amongst the 134 variants in the best model, eight were directly identified as GWAS hits on meta-analysis (J. C. Lambert et al., 2013). Additional GWAS hits were represented in the PRS, as ten of the PRS SNPs were in linkage disequilibrium with GWAS hits and would have been clumped out had the GWAS hit been genotyped on the NeuroChip. The online variant database, dbSNP (NCBI), was used to identify the genes each SNP fell within, 85 were present within 80 gene-coding regions, 12 SNPs were in RNA-coding regions or non-coding loci (LINC/LOC) whilst 37 were intergenic.

As previously mentioned, running PRSice without clumping identified an additional 151 SNPs which were removed from the model due to being in LD with another SNP of greater significance according to the base dataset. These SNP identities were derived by running analysis at the same threshold as the PRSice best model (P_T=1.4×10^{-4}) and checking the .snp file. The list of SNPs was mapped on an online LD reference tool, LDmatrix, to identify the observed LD based on a UK population. The results were exported and compiled to show the LD blocks existing within the dataset and the effectiveness of the clumping algorithm as shown in Figure 3.

3.4.1 TaqMan GWAS SNPs

In the PRSice best model it was determined that there was risk of AD associated with many regions of the genome, implicating 134 SNPs from up to 80 genes associated with developing or preventing AD. Some of the SNPs identified were observed to be within the loci of previously determined genes with significant association at a

genome-wide level according to GWAS results. Many of the previous GWAS hits however were not included on the NeuroChip array and were therefore genotyped using TaqMan assay. The effect of these GWAS hits, listed in Table 7, in lieu of the SNPs representing these associations in the model were measured by incorporating these variants into the PRSice best model and observing the effect.

Initially 9 variants were included with the 134 variants from the best model as a new target dataset and PRSice was run with the same clumping parameters to identify which variants were removed due to being within the same LD block; clumping reduced the number of variants from 143 to 137. The SNP file was used to breakdown the clumping process to find that 3 variants in the best model were replaced by GWAS hits (*CR1*, *PTK2B*, and *CLU*). In addition, 3 variants were introduced into the model as there was no previous representation from these loci (*INPP5D, HLA-DRB1*, *SORL1*), and 3 GWAS hits were less significant than the variant representing the locus according to the base dataset and were therefore not included (*CD2AP*, *CELF1*, *MS4A2*).

PRSice analysis of the clumped dataset slightly improved performance of the model when including the additional GWAS hits, Nagelkerke's R^2 increased to 0.0419 with a p-value of 3.54×10^{-5}. Further to this, the inclusion of significant SNPs with known effect has led to PRSice identifying a more predictive model at a lower threshold of significance. The best model was determined at a threshold of $p\leq1.19\times10^{-4}$ with a Nagelkerke's R^2 of 0.0421 and a p-value of 3.56×10^{-5}. With *APOE* SNPs included, both models showed significant improvement as observed when *APOE* was introduced to the PRSice best model from the initial results.

Model	N SNPs	Nagelkerke's R^2	P	AUROC	AUPRC
PRSice best model	134	0.0418	4.42×10^{-5}	59.5%	70.8%
PRS best model + *APOE* SNPs + sex + AAD	136	0.2030	2.51×10^{-16}	73.4%	81.5%
PRSice best model + GWAS	137	0.0419	3.54×10^{-5}	59.8%	71.0%
PRSice+GWAS best model	120	0.0421	3.56×10^{-5}	60.1%	70.9%
PRSice+GWAS bm + *APOE* (GWAS PRS)	122	0.2000	1.83×10^{-17}	73.3%	81.3%
GWAS PRS + sex + AAD	122	0.2040	1.78×10^{-17}	73.4%	81.4%

Table 7: Effect of GWAS SNPs on PRS of LOAD modelling

The table identifies the results from including additional GWAS hits to the PRSice best model. The previously derived best models are listed with a model with the previously unrepresented GWAS hits included, the more predictive model when GWAS hits are included, and with the further inclusion of *APOE* SNPs and covariates for sex and age at death. The table includes the number of SNPs present in each model, the Nagelkerke's R^2 and P-value derived from PRSice and SPSS, and the predictive ability derived in R.

3.4.2 Imputation

The imputed dataset was analysed under the same PRSice clumping conditions and test parameters, to compare with the PRSice results from the genotyped dataset.

The threshold identified as the best model from genotyped data was compared to imputed data to identify additional loci previously not picked up by PRSice but within the tested limits. These would be introduced to the dataset as a replacement for genotyped SNPs which were less significantly associated with disease phenotype according to the IGAP meta-analysis. At the best threshold (P_t =1.4×10^{-4}), there were 269 SNPs present in the model with only 71 common to those identified in the genotyped model; imputation introduced 198 new variants within this threshold, replacing 63 from the genotyped model and potentially introducing additional loci to the model.

3.5 Discussion

PRSice quality controls were stricter than the parameters used in earlier quality control stages of genotyping. Although they led to a reduction in the number of SNPs being incorporated into the analysis, most of the removed SNPs were duplicate or mismatched SNPs; multiallelic SNPs were unusable in PRSice analysis and would be excluded during this stage of analysis if not before. SNPs of low call rate made up a small proportion of the SNPs that were removed but may have led to type II errors in the results due to missing calls suggesting effect due to their absence in some individuals; SNPs with low minor allele frequency are classed as rare SNPs and can have strong effects on the dataset, they were removed so that analysis was based on the variation observed in common genetic variants.

The clumping parameters used in PRSice analysis were the defaults suggested in PRSice manuals and those used in other examples of PRS analysis. The rationale for

using a low r^2 value was to determine if all SNPs used in the analysis were independent and represented a single locus in the genome, the addition of variants from the same locus could inflate the results based on the significance that loci holds. Multiple SNPs of great significance from the same locus can have the ability to drive the results to suggest the locus is of greater importance than it is, therefore making the PRSice model inaccurate.

This effect is widely observed to occur in the *APOE* locus, the hotspot identifies numerous significant variants within the region which would inflate the score. The biological understanding of the gene has identified two variants within the locus which help determine which of the *APOE* subtypes individuals carry. These two variants are not the most significant within the locus based on association data and there may in fact be other variants within the locus which have pathological effects on disease progression. Hence, which of these two alleles an individual carries are a more biologically relevant representation of the locus and are therefore included at a later stage of the PRS-specific analysis.

The number of thresholds tested, and the increments of testing used were also PRSice defaults, the software is able to calculate models from smaller incremental differences with more specific threshold ranges; however, these were not used as to avoid further biasing the results based on expected or predetermined results.

The results from PRSice identify a best model using few SNPs at a low significance threshold, whilst Figure 5 also shows fluctuations in model predictability as more SNPs are introduced but never succeeding the best model determined at the p-value threshold of 1.4×10^{-4}. The resulting model shows significant differences in mean PRS between cases and controls, confirming the NeuroChip-genotyped risk is higher in cases than controls. The downstream analyses reinforced this understanding by

increasing the difference in scores between the two groups as covariates are introduced. *APOE* status is known to be a factor which affects risk of LOAD and on its own is more predictive than the PRS genetic model alone; when both are combined, the new model is more predictive as the trends from both models complement one another. This can also be seen from the incorporation of age at death and controlling for sex. AAD is highly variable for controls and lower than cases which provides insignificant predictive power on its own but can further improve the model whilst proportions of females between and within groups is fairly consistent and has little observed effect on the model but leads to overall improvement.

Decile scoring was used to further identify the trends of PRS incorporating *APOE* and controlling for sex and age at death in cases and controls. The distribution of scores in Figure 6 identifies a large overlap between individuals from both groups with fewer individuals at either extreme of low or high risk. Introducing tiers of increasing risk gives a better representation of the distribution of samples, and the benefit of modelling individuals by PRS. The presence of cases at lower thresholds controls at higher deciles, most notably in decile 8, suggests there are other factors which affect disease progression beyond genetic risk. Lifestyle choices affect likelihood of LOAD, as controls with high PRS may have prevented disease presentation whilst cases with lower PRS may have lived unhealthy lives advancing onset of AD symptoms. Additionally, the PRS from NeuroChip genotyping may not capture the entirety of genetic risk associated with LOAD.

The variants used in the model represent the associations of 134 loci across the genome, significant to at least 1.4×10^{-4} or lower. Most SNPs were found to be within gene coding regions and may implicate those genes in AD pathogenesis; some variants were intergenic and may either represent a pathologically relevant variant within that locus or potentially a variant involved in regulation of expression of a gene

associated with LOAD. Other variants may also implicate regions of DNA which are transcribed for RNA sequences which may also affect the development of AD.

The clumping algorithm (Figure 3) explains the process by which the algorithm works to reduce the number of SNPs from loci of interest. Each individual chromosome plot shows the relationships between SNPs, and areas with strong correlations of inheritance form blocks of red; there was no plot for chromosome 21 since no SNPs were present in the model from that chromosome at the chosen threshold. Each block is represented by a single SNP in the model, had these clumped out SNPs been included in the PRSice model, the resulting model would have been statistically more predictive but less accurate.

The TaqMan genotyped SNPs were incorporated into the dataset with varying effects; 3 SNPs replaced existing SNPs in the model during the clumping stage as they represented the same loci and the TaqMan assay had a more significant p-value in the base dataset. However, 3 SNPs were not incorporated into the dataset as a more significant variant was already present within the model representing that locus. The last 3 SNPs incorporated into the model had a very small effect as they were introduced as novel signals, the effect of their inclusion was best observed by the resulting identification of a more predictive model at a lower threshold. The resulting improvements in predictive ability between the genotyped model and the inclusion of GWAS hits are not significant however, the results show the effects of most of the hits are already represented in the model in some way, there were also an additional 8 GWAS hits already present in the model before the rest were TaqMan assayed. This suggests the NeuroChip variant selection was already effective at capturing much of the variation associated with risk of AD.

The imputed dataset includes all genotyped variants which passed quality controls and additional variants imputed based on their inheritance in other individuals. The SNPs included in the model at the threshold of significant the best model was derived from includes a greater number of SNPs and therefore loci than the genotyped model. Many of the SNPs now imputed were more significant according to the base dataset than its previous representative from the LD block which explains their absence, whilst a greater number of new loci are now identified to predict risk of LOAD in the BDR cohort.

4 Polygenic risk scoring of sporadic early-onset Alzheimer's disease

4.1 Perspective

Polygenic risk scoring was conducted on a cohort of sporadic early-onset Alzheimer's disease (sEOAD) cases and neurologically healthy controls (target dataset), based on summary statistics from the International Genomics of Alzheimer's disease Project (IGAP) meta-analysis (base dataset), using the most current stable version of PRSice-2 software. PRSice-2 was used to generate PRS on samples genotyped on the NeuroX array; the first neuro-specific array to include variants implicated in neurodegenerative diseases including Alzheimer's disease and other dementias, Parkinson's disease, amyotrophic lateral sclerosis, and progressive supranuclear palsy. The NeuroX was built on the backbone of the Infinium HumanExome BeadChip consisting of 242,901 SNPs, with an additional 24,706 SNPs of custom content. The Alzheimer's Research UK (ARUK) consortium recruited samples with clinical information (post-mortem pathological confirmation of disease status and age at death). Genotyped samples were quality controlled using Illumina GenomeStudio and PLINK and later imputed using the Michigan Imputation Server Minimac4 software.

Previous analysis of this cohort was conducted using an earlier version of PRSice software on the same individuals. The results showed significant differences between mean PRS between sEOAD cases (n=408) and controls (n=437), with a predictive ability (area under the ROC curve; AUC) for distinguishing between groups of 75.5% (Chaudhury et al., 2018). When samples were sorted into deciles of increasing risk and a greater proportion of cases were observed at higher deciles with the opposite being seen in controls, confirming the association between genetic risk of disease and likelihood of developing sEOAD.

The primary aim of this study was to determine the genetic risk of AD in a cohort of early-onset cases using the PRS approach.

4.2 Samples

DNA was isolated from blood and/or brain tissue from cases (n=451) and controls (n=528). Lab procedures were carried out by I. Barber following methods as described (Section 2.1). All samples were genotyped using the NeuroX array. Quality control of the genotyping data was completed using GenomeStudio v2011.1 and PLINK (Section 2.3.3). Samples were checked for call rate, relatedness, diversion from European ancestry, and heterozygosity; SNPs were controlled for call rate and adherence to Hardy-Weinberg equilibrium. The resulting dataset consisted of 408 cases of sEOAD and 436 cognitively healthy controls (Table 8), genotyped for 265,049 SNPs. SNPs associated with *APOE* status were genotyped for all samples using TaqMan assays to identify the number of ε4 alleles individuals harbour, carried out by I. Barber following methods described (Section 2.1.4).

Imputation using the Michigan imputation server was completed on the entire cohort to identify additional variants based on the observed genotypes. The genotyped dataset underwent pre-imputation quality controls of the strand and position of alleles, and their calls, assignments, and frequencies were checked to be consistent with the Haplotype Referencing Consortium. Imputation increased the number of SNPs accessible by PRSice on the cohort (n=2,112,986).

Phenotype	N	Age at death (SD)	Females (%)	*APOE* ε4+ (%)	*APOE* ε4ε4 (%)
Control	436	77.2 (6.44)	253 (58.0)	104 (23.9)	9 (2.1)
Sporadic early-onset AD	408	64.8 (5.48)	194 (47.5)	234 (57.4)	54 (13.2)

Table 8: Demographics of sEOAD samples

Individuals, categorised by disease status, were recruited from across the UK: sEOAD cases from universities in Bristol, Manchester, Nottingham, Oxford, and controls from University College London. Age at death (together with standard deviation) was obtained from clinical information. The number and percentage of females in each group is also shown as well as the number of individuals with at least one *APOE* ε4 allele and additionally those who were ε4ε4 homozygotes.

4.3 Polygenic risk scoring

Primary PRS analysis (Figure 7) tested 408 sEOAD cases and 436 controls using default PRSice-2 parameters and additional PRS quality control of the base and target dataset. PRSice quality control of the target dataset included standard cut-off thresholds for minor allele frequency (MAF ≤ 0.01), Hardy-Weinberg equilibrium (HWE ≤ 1×10^{-6}), genotyping call rate (GENO ≤ 0.01) and sample call rate (MIND ≤ 0.01). Ambiguous or duplicate SNPs were removed, and mismatched SNPs were resolved in both the base and target datasets; following QC, the target dataset was reduced to 40,302 SNPs. The *APOE* locus, a window of 250kb either side of *APOE* (chr19:45,160,844-45,660,844; GRCh37/hg19), was excluded from the target dataset as a well-known AD hotspot; the SNPs which identify the *APOE* ε status would be reintroduced to the risk modelling at a later stage.

PRSice-2 uses PLINK to calculate the linkage disequilibrium (LD) between SNPs, estimated from a cohort of 503 European individuals genotyped within the 1000 Genomes project (LD dataset). The SNPs were then clumped based on their LD to reduce the number of potential false positives. The clumping algorithm identifies the most significant SNP (index) according to the base dataset in a sliding window and removes SNPs in LD with the index SNP above a given r^2 value, which signifies the inheritance correlated between the SNPs, to reduce the occurrence of type I errors; false positives are introduced as SNPs in LD with a variant with known association with the disease will appear as a novel signal. The default clumping settings for PRSice-2 which were used included a 250kb sliding window and r^2 set as 0.1; the dataset was also tested without clumping to compare and evaluate its effect.

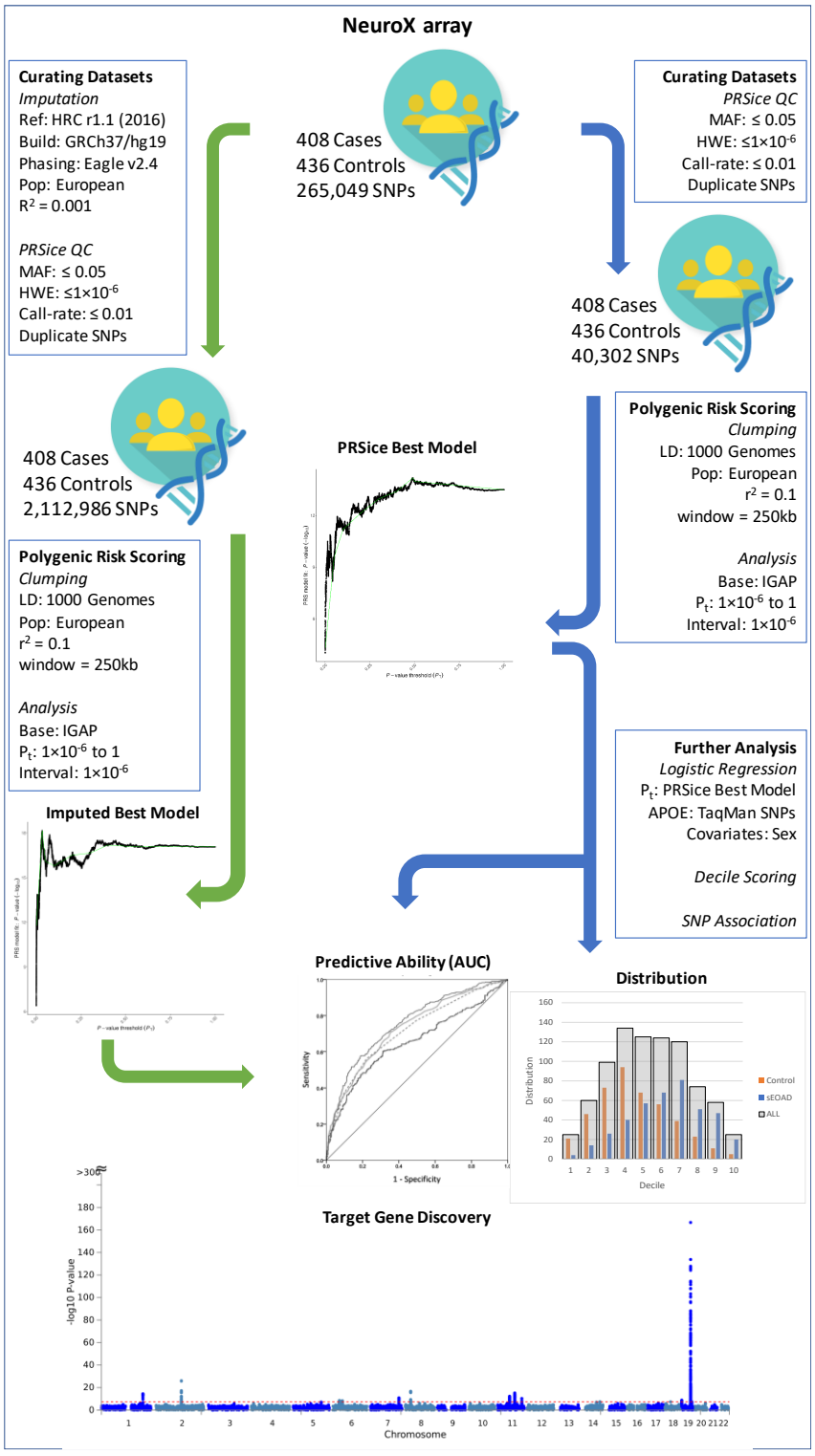


Figure 7: Analysis pipeline for PRS of sEOAD with the NeuroX

The figure outlines the process of generating polygenic risk scores from a target dataset. Primary analysis (blue) is run in parallel with secondary analysis, the use of imputation (green). Polygenic risk scores are generated, and covariates are introduced; the results are used to determine the predictive ability of the best PRS model and the measure the distribution of samples based on PRS. The resulting models are checked for previously unidentified genes associated with AD.

PRSice-2 then selects all SNPs present in the base and target dataset and runs logistic regression to model cases against controls based on the SNPs present within given significance thresholds; the threshold incrementally increases to include additional SNPs for the next model until all SNPs have been included. The results of the modelling are compared and then scored by Nagelkerke's R^2 to identify the best model fit. The lowest threshold set for PRSice-2 analysis started at SNPs with significance of association in the base dataset between $0 \leq P_t \leq 1\times10^{-6}$, increasing in increments of 1×10^{-6} as new SNPs are introduced up to $P_t \leq 1$.

4.3.1 Best model selection

PRSice-2 provided individual scores at each tested threshold and identified the best p-value threshold for modelling risk in the target cohort based on Nagelkerke's R^2 and the number of SNPs used in the model. The resulting models at various testing thresholds are shown in Table 9, with the PRSice output figures included as Figure 8. The most predictive model was identified with a Nagelkerke's R^2 value of 0.104, at a p-value threshold of 0.490401 and consists of 10,927 SNPs. At this threshold, there were 7,356 additional SNPs which were clumped out by the PRSice algorithm.

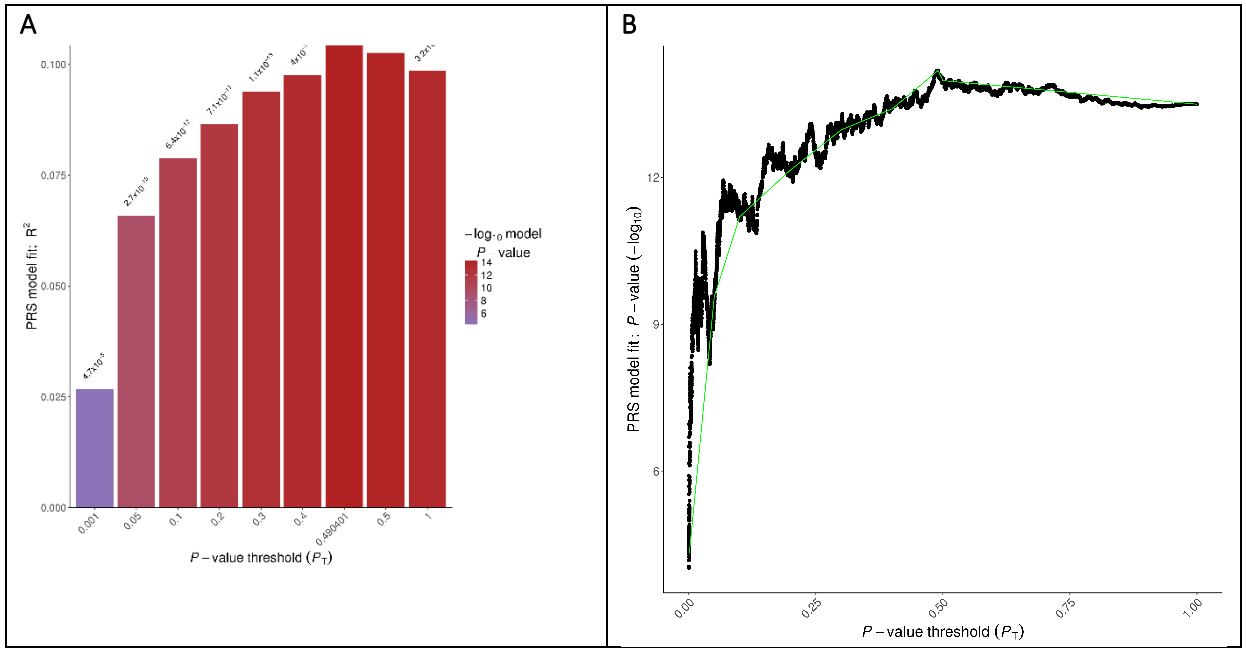


Figure 8: Output figures of PRS modelling of sEOAD with the NeuroX

The bar plot (**A**) presents some of the thresholds in Table 9, identifying the model fitness expressed as Nagelkerke's R^2 at different stages. The results show model to improve with the inclusion of more SNPs and peak at 0.490401 before falling once more. The high-resolution plot (**B**) identifies the p-value derived from each model at all tested thresholds, giving more details about the effect shown in **A**. The predictive ability fluctuates as more SNPs are introduced forming a maximum peak and plateauing steadily afterwards.

P-value threshold	N SNPs	Nagelkerke's R^2	Significance (P)
≤0.000001	10	0.0460	1.1×10^{-7}
≤0.000010	12	0.0441	2.02×10^{-5}
≤0.000100	28	0.0248	8.87×10^{-5}
≤0.001000	95	0.0267	4.67×10^{-5}
≤0.010000	430	0.0614	1.15×10^{-9}
≤0.050000	1609	0.0659	2.73×10^{-10}
≤0.100000	2856	0.0788	6.38×10^{-12}
≤0.200000	5126	0.0865	7.12×10^{-13}
≤0.300000	7206	0.0938	1.07×10^{-13}
≤0.400000	9228	0.0976	4.01×10^{-14}
≤0.490401	10927	0.1044	6.59×10^{-15}
≤0.500000	11091	0.1026	1.06×10^{-14}
≤1.000000	18853	0.0985	3.22×10^{-14}

Table 9: PRSice modelling of sEOAD with the NeuroX

The table outlines the results of PRSice modelling at p-value thresholds of significance according to the base dataset. At each threshold PRSice identifies the number of SNPs present in the target dataset at the threshold and calculates model fitness that can be determined based on the SNPs as Nagelkerke's R^2 and the associated p-value of the model. As the threshold increases, more SNPs are introduced to the model which affects the model predictiveness. The best model derived by PRSice is identified at a threshold of p≤0.490401, with the largest Nagelkerke's R^2 and the most significant p-value.

Mean individual PRS at the PRSice best model was calculated as 3.93×10^{-4} with a standard deviation of 1.06×10^{-4}. Mean PRS and standard deviation for controls and sEOAD cases was found as 3.64×10^{-4} $\pm1.04\times10^{-4}$ and 4.24×10^{-4} $\pm9.94\times10^{-5}$, respectively. Single factor ANOVA found variance for controls as 1.08×10^{-8} and sEOAD cases 9.87×10^{-9}, and the significance of the variation to be 1.36×10^{-16}.

The best model was then tested with the re-introduction of SNPs associated with *APOE* ε status (rs429358 and rs7412) and running PRSice-2 with the inclusion of covariates for sex. The SNPs were introduced to the dataset using PLINK by extracting the SNPs in the best model and merging with a dataset of calls for both *APOE* SNPs according to TaqMan genotyping. Covariates were included in the PRSice script with the inclusion of a covariate file identifying each individual and stating their clinical sex. Covariates testing was also replicated using binary logistic regression in SPSS (IBM) to validate the results and record an individual score which combined PRS with covariates. Receiver operating characteristic (ROC) curves are produced from this value, and the area under the curve (AUC) is calculated as a measure of predictability. The resulting models were compared to the PRSice best model in Table 10.

Model	N SNPs	Nagelkerke's R^2	IGAP P-value	AUROC (%)
PRSice best model	10927	0.104	6.65×10^{-15}	66.2
APOE	2	0.175	1.87×10^{-23}	69.5
PRS (with *APOE*)	10929	0.190	1.53×10^{-24}	72.1
PRS + sex	10929	0.201	2.45×10^{-24}	73.0

Table 10: Modelling PRS of sEOAD with covariates

The best model by PRSice was isolated and tested with the inclusion of the SNPs associated with *APOE* ε status, rs429358 (ε4) and rs7412 (ε2). The *APOE* SNPs alone show greater significance and model disease status better than the best model, however, when combined the PRS model shows improved fitness than either alone. Clinical sex was introduced to the PRS with *APOE* model to show the best model according to Nagelkerke's R^2 and was found to be PRS with sex considered. Predictive ability for all models was also calculated using area under the receiver operating characteristic curve (AUROC).

4.3.2 Association and risk prediction

Following the results of covariate analysis, prediction of disease status was most informative when including *APOE* and sex as a covariate, therefore the resulting normalised scores for this model were used for downstream analyses. Normalised mean score and standard deviation for controls was calculated as 0.409±0.176 and 0.563±0.184 for sEOAD cases. Variance for controls and cases were calculated as 0.0309 and 0.0337 respectively and found to be significantly different by single factor ANOVA (p=5.78×10^{-33}).

Further statistical analysis included decile scoring, where all samples were ranked by score (PRS with *APOE* + sex) and distributed into ten tiers of increasing risk. The samples were then separated by disease classification to identify the proportion of cases and controls which appeared in each decile, given in Table 11, and presented in Figure 9. The results show cases and controls are present within all deciles; most controls populate the lower five deciles whilst cases are proportionally higher in the upper five deciles. Controls are observed to be highest in the fourth decile whilst the most cases are represented in the seventh decile.

Decile	1	2	3	4	5	6	7	8	9	10
All individuals	25	60	99	134	125	124	120	74	58	25
Controls	21	46	73	94	68	56	39	23	11	5
Sporadic early-onset AD	4	14	26	40	57	68	81	51	47	20

Table 11: Distribution of sEOAD samples by decile scoring

The table shows the results of decile scoring on the cohort of sEOAD cases and controls. The number of individuals, which fall within each decile, are given alongside the number of cases and controls. The entire cohort shows normal distribution with most samples observed between the fourth and seventh deciles.

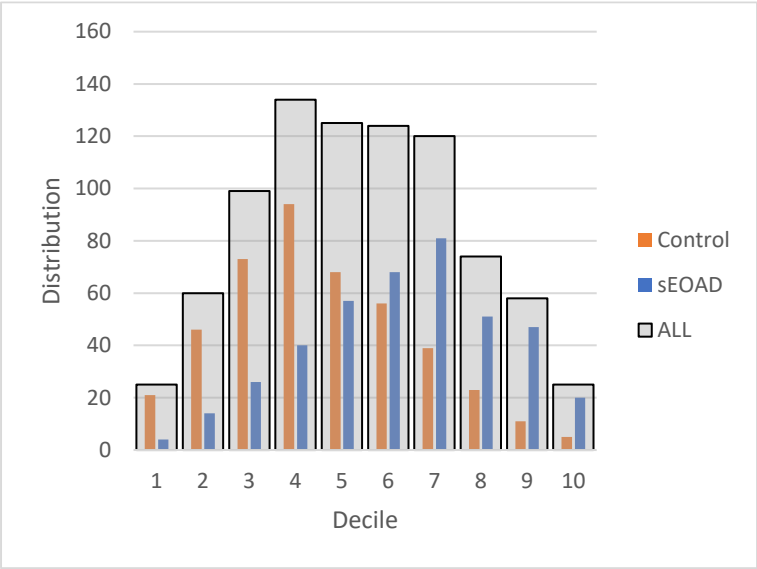


Figure 9: Distribution of sEOAD PRS by decile scoring

The figure shows the distribution of cohort of sEOAD cases and controls when sorted into deciles. The distribution of all individuals are given as grey bars within which are the distribution of controls (orange) and sEOAD cases (blue). The controls show normal distribution, with a peak around the fourth decile; the distribution of cases shows a right sided skew with more cases in higher deciles peaking in the seventh. This confirms a larger proportion of cases have higher risk classification of disease than controls and can be discriminated using PRS.

4.4 SNPs

The rsID (reference SNP identification tag) of the SNPs which were present in the best model were checked against online variant databases to identify their observed variation and the gene in which they are present. From amongst the variants in the best model, two were directly identified as GWAS hits on meta-analysis (J. C. Lambert et al., 2013). Additional GWAS hits were represented in the PRS, as ten of the PRS SNPs were in linkage disequilibrium with GWAS hits and may have been clumped out had the GWAS hit been genotyped on the NeuroX. The online variant database, dbSNP (NCBI), was used to identify the genes each SNP fell within, 7,819 were present within 5,754 gene-coding regions, 682 SNPs were in RNA-coding regions or non-coding loci (LINC/LOC) whilst 2,426 were intergenic.

As previously mentioned, running PRSice without clumping identified an additional 7,390 SNPs which were removed from the model due to being in LD with another SNP of greater significance according to the base dataset. The Nagelkerke's R^2 of a model at the same threshold (P_t=0.490401) without clumping is calculated as 0.0215, with a p-value of 2.57×10^{-4}.

4.4.1 Imputation

The imputed dataset was analysed under the same PRSice clumping conditions and test parameters, to compare with the PRSice results from the genotyped dataset. The threshold identified as the best model from genotyped data was compared to imputed data to identify additional loci previously not picked up by PRSice but within the tested limits. At the PRSice-derived best threshold (P_t =0.490401), there were 49,037 SNPs present in the model with only 155 common to those identified in the genotyped model. The model based on imputed variants at this threshold had a

Nagelkerke's R^2 of 0.128 and associated significance of $p=6.8\times10^{-18}$ and AUROC of 67.5%.

The imputed dataset was also tested without predetermining the threshold of testing, unbiased, and found a best model to be at a different threshold to the genotyped target dataset. The best model from imputed SNPs was found to be at a threshold of 0.03249, with a Nagelkerke's R^2 of 0.138 based on 8,204 SNPs and significant to a level of $p\leq6.96\times10^{-19}$. With genotypes for *APOE* SNPs considered in the model, Nagelkerke's R^2 increased to 0.197 and to 0.207 when sex was included as a covariate; additionally, a more predictive model was also presented with *APOE* SNPs introduced, at a p-value threshold of 0.000021 (Nagelkerke's $R^2=0.209$, $p=7.96\times10^{-27}$). Predictive ability was calculated using AUROC for all models at the p-value threshold of 0.03249; PRS on its own was calculated as 68.7%, increasing to 72.3% with *APOE* included and 72.9% with the addition of the covariate for sex.

In a breakdown of the 8,204 SNPs in the imputed best model there were 2,644 intergenic SNPs included and the remainder were found in 4,186 loci, of which 3,610 are gene-coding and 574 were RNA-coding regions or non-coding loci. Amongst the gene-coding loci, 1,918 genes were represented in the previous model.

4.5 Discussion

PRSice quality controls were even stricter than the parameters used in earlier quality control stages of genotyping and led to a greater reduction in usable data than with the NeuroChip. Quality controls for PRS are to ensure the variants included in analysis are commonly found in populations and thus the results would be repeatable and accurate. Most SNPs eliminated from the dataset were for having low minor allele frequency (n=212,941), which is consistent as the NeuroX hosts a range of rare variants upon its array design; additionally, 1,415 SNPs were removed for low SNP call

rate. As there may be significant associations with risk of AD from rare variants, which would be another avenue of research.

The clumping parameters used in PRSice analysis were the defaults suggested in PRSice manuals and those used in other examples of PRS analysis. This clumping approach differed to that which was used in previous analysis of the cohort, where SNPs were pruned according to LD and a r^2 of 0.8 was used to only exclude highly correlative SNPs and identify the most predictive model. This may give an explanation as to why a lower predictive ability was calculated and it may be worthwhile to explore additional clumping thresholds than the default thresholds set by PRSice-2.

The rationale for using a low r^2 value was to determine all SNPs used in the analysis were independent and represented a single locus in the genome as the addition of variants from the same locus could inflate the results based on the significance that loci holds. Multiple SNPs of great significance from the same locus can have the ability to drive the results to suggest the locus is of greater importance than it is, therefore making the PRSice model inaccurate. However, there were some observations of multiple SNPs from the same gene, possibly indicate a role in AD pathology, given that each variant is independent and still associated with risk according to IGAP.

The number of thresholds that were tested and the increments of testing used were also PRSice defaults, the software can calculate models from smaller incremental differences with more specific threshold ranges; however, these were not used as to avoid further biasing the results based on expected or predetermined results. The genotyped results showed a p-value threshold to be like that of previous analysis of the cohort, with more predictive ability introduced once all variants with some association ($p \leq 0.5$) with AD risk were included.

The results from PRSice identify a best model using many SNPs and Figure 8 also shows fluctuations in model predictability as more SNPs are introduced until reaching a peak at a p-value threshold of 0.490401. Notably, there were fewer SNPs content present on the NeuroX array with significant association with AD, which may be one factor which explains the difference in results between the NeuroX sEOAD and NeuroChip LOAD results (Results 1). After imputing the dataset to include a greater number of variants, including those more likely to be included in IGAP, there is now a best model determined at a more significant threshold.

The resulting model shows significant differences in mean PRS between cases and controls, confirming the NeuroX-genotyped risk is higher in cases than controls. The downstream analyses reinforced this understanding by increasing the difference in scores between the two groups as covariates are introduced. *APOE* status is known to be a factor which affects risk of AD and on its own is more predictive than the PRS genetic model alone; when both are combined, the new model is more predictive as the trends from both models complement one another. As sex is controlled for, there is also slight improvement of the PRS model. Age at death was not included as a covariate in this analysis as controls were selected based on cognitive health in late age whilst cases suffered and died early in life.

Decile scoring was used to further identify the trends in cases and controls. The distribution of scores in Figure 9 identifies some overlap between individuals from either group with peaks forming in different deciles for each group. The presence of cases and controls at lower and higher deciles respectively suggests there are other factors which affect disease progression beyond genetic risk, however, the greater separation between groups is indicative of a much greater risk score amongst cases than controls, leading to earlier onset of disease.

The variants used in the genotyped model represent the associations of 7,819 loci across the genome, significant to at least 0.490401 or lower. Almost 75% of SNPs were found to be within gene coding regions and may implicate those genes in AD pathogenesis; some variants were intergenic and may either represent a pathologically relevant variant within that locus or potentially a variant involved in regulation of expression of a gene associated with LOAD. Other variants may also implicate regions of DNA which are transcribed for RNA sequences which may also affect the development of AD. Additionally more gene-coding loci were introduced by the imputed model analysis, indicating heterogeneity in disease risk.

5 Polygenic risk scoring of expression networks

5.1 Perspective

Polygenic risk scoring (PRS) was completed on cohorts of late-onset Alzheimer's disease (LOAD) cases and neurologically healthy controls from the Brains for Dementia Research (BDR) resource and sporadic early-onset Alzheimer's disease (sEOAD) cases and neurologically healthy controls, based on summary statistics from the International Genomics of Alzheimer's disease Project (IGAP) meta-analysis (base dataset) (Chapters 3 & 4). PRSice-2 was then used to generate PRS from imputed data on samples genotyped on both NeuroX and NeuroChip arrays on SNPs present within a subset of genes. Gene subsets were defined from scientific literature relating to either genes expressed within a localised region or within a pathway associated with risk of AD.

Previous analysis with this approach was conducted using an earlier version of PRSice-2 software on NeuroChip-genotyped samples from the first two batches of the BDR cohort on a subset of genes identified to be expressed at the synapse (Lawingco et al., 2021; Lleó et al., 2019). The results found a 'Synaptic PRS' model using 6 SNPs in combination with *APOE* SNPs to be significantly higher in LOAD cases (n=302) than controls (n=137), with a predictive ability (area under the ROC curve; AUC) for distinguishing between groups of 72%. The predictive ability of this model was compared to a full genetic PRS model and replicated in the third batch of BDR samples with a predictive ability of 73%. The synaptic PRS model was made up of previously identified genes associated with AD through GWAS (*BIN1*, *PTK2B*, *PICALM*, *APOE*) and two novel loci in the genes *DLG2* and *MINK1*.

The primary aim of the study is to understand the relationship between synaptic genes and risk of Alzheimer's disease in two independent cohorts and to compare the results to PRS models of whole genome data.

5.2 Samples

DNA was isolated from blood and/or brain tissue from all NeuroChip (n=1172) and NeuroX (n=979) samples. Lab procedures were carried out by T. Patel and I. Barber following methods as described (Section 2.1). Quality control of the genotyping data was completed using GenomeStudio and PLINK (Section 2.3.3). Samples were checked for call rate, relatedness, diversion from European ancestry, and heterozygosity; SNPs were controlled for call rate and adherence to Hardy-Weinberg equilibrium. The resulting target datasets consisted of 358 LOAD cases and 217 controls on the NeuroChip and 408 sEOAD cases and 436 controls on the NeuroX (Table 12).

Imputation using the Michigan imputation server was completed on both datasets to identify additional variants based on the observed genotypes. The genotyped dataset underwent pre-imputation quality controls of the strand and position of alleles, and their calls, assignments, and frequencies were checked to be consistent with the Haplotype Referencing Consortium. Imputation increased the number of SNPs for both the NeuroChip (n=5,379,656) and NeuroX (n=4,284,386) cohorts.

Array	Phenotype	N	Age at death (SD)	Females (%)	*APOE* ε4+ (%)	*APOE* ε4ε4 (%)
NeuroX	Control	436	77.2 (6.44)	253 (58.0)	104 (23.9)	9 (2.1)
	sEOAD	408	64.8 (5.48)	194 (47.5)	234 (57.4)	54 (13.2)
NeuroChip	Control	217	49.7 (44.2)	115 (53.0)	35 (16.1)	2 (0.9)
	LOAD	358	83.2 (8.5)	173 (48.0)	231 (64.5)	49 (13.7)

Table 12: Demographics of samples for synaptic PRS

Individuals, categorised by disease status and array, were recruited from across the UK: NeuroX sEOAD cases from universities in Bristol, Manchester, Nottingham, Oxford, and NeuroX controls from University College London; NeuroChip samples were recruited from universities in Bristol, Manchester, Newcastle, Oxford, and King's College London. Age at death (together with standard deviation) was obtained from clinical information. The number and percentage of females in each group is also shown as well as the number of individuals with at least one *APOE* ε4 allele and additionally those who were ε4ε4 homozygotes.

5.3 Polygenic risk scoring

5.3.1 Quality controls and gene set discovery

Selective PRS analyses (Figure 10) tested both imputed target datasets, 358 LOAD cases against 217 controls and 408 sEOAD cases against 436 controls, using default PRSice-2 parameters and additional PRS quality control. PRSice quality control of the target datasets included standard cut-off thresholds for minor allele frequency (MAF ≤ 0.01), Hardy-Weinberg equilibrium (HWE ≤ 1×10^{-6}), genotyping call rate (GENO ≤ 0.01) and sample call rate (MIND ≤ 0.01). Ambiguous or duplicate SNPs were removed, and mismatched SNPs were resolved in both the base and target datasets; following QC, the NeuroChip dataset was reduced to 4,222,576 SNPs and the NeuroX dataset was reduced to 2,112,986. The *APOE* locus was excluded from the target datasets, as the SNPs which identify the *APOE* ε status would be reintroduced to the risk modelling at a later stage.

Lists of genes included in gene sets were identified through proteomic studies, where proteins with an established function and detected in enriched fractions from tissue of a proteome are considered. Synapse-enriched mouse, rat, and human brain tissue was tested to identify synaptic genes, 537 were considered synaptic (Lleó et al., 2019).

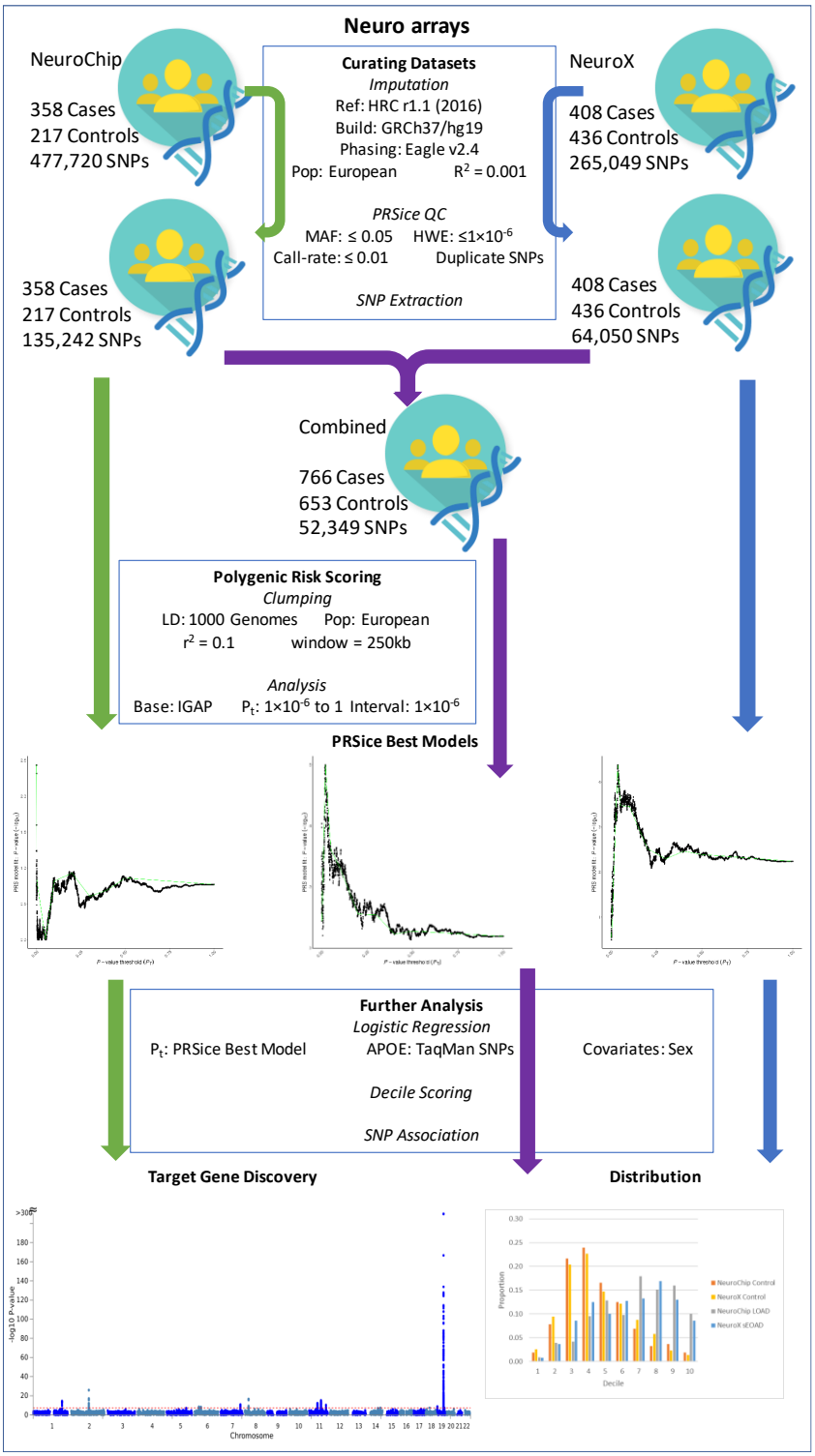


Figure 10: Analysis pipeline for synaptic PRS of AD

The figure outlines the process of generating synaptic polygenic risk scores from multiple target datasets. Analyses are run in parallel, of the NeuroChip (green) and the NeuroX (blue) dataset. Polygenic risk scores are generated, and covariates are introduced; the results of individual array models are used to determine the predictive ability of the best PRS model and SNPs implicated. The resulting models were compared to previously completed whole genome PRS results and subsequent distribution.

SNPs present within the gene ranges were extracted from the target datasets using PLINK, the genomic locations of proteins were identified according to the GRCh37/hg19 reference assembly (Kent et al., 2002); the resulting datasets were saved for PRS analysis. All samples were also combined to form a larger dataset with SNPs common to both imputed arrays. The isolation of synaptic SNPs reduced the NeuroChip dataset to 136,242 SNPs and the NeuroX dataset to 64,050 SNPs.

The default clumping settings for PRSice-2 included a 250kb sliding window and r^2 set as 0.1. The lowest threshold set for PRSice-2 analysis started at SNPs with significance of association in the base dataset between $0 \leq p \leq 1\times10^{-6}$, increasing in increments of 1×10^{-6} as new SNPs are introduced up to $P_t \leq 1$.

5.3.2 Best model selection

Clumping led to a further reduction of variants used for PRS analysis; the NeuroChip dataset consisted of 5,649 independent SNPs and the NeuroX dataset consisted of 2,961 SNPs for PRS analysis. The best model on each array was determined by Nagelkerke's R^2, individual scores were recorded for each model and SNPs which make up the best model were identified.

With the NeuroChip, the modelling between LOAD cases and controls peaked at a significance threshold of 0.000001 with Nagelkerke's R^2 calculated as 0.0203 and using 4 SNPs (p=0.0037), subsequent models were not as predictive as more SNPs were introduced as Nagelkerke's R^2 plateaued at 0.005 (Figure 11). Mean synaptic PRS was calculated as 0.0181 with a standard deviation of ±0.0192 for controls and 0.0228±0.0182 for LOAD cases; single-factor ANOVA found variance between groups to be significant to p=0.0038. Predictive ability was calculated as both area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) using R package PRROC, calculated as 58.0% and 66.8% respectively.

The best model from the NeuroX between sEOAD cases and controls was at a p-value threshold of 0.036111, where Nagelkerke's R^2 was 0.0271, significant to a p-value of 4.18×10^{-5} and consisting of 376 SNPs. Predictive ability increased as SNPs were introduced up to a peak, and steadily reduced as additional SNPs were included until Nagelkerke's R^2 plateaued above 0.01 (Figure 11). Mean synaptic PRS was calculated as 0.000497 with a standard deviation of ±0.000929 for controls and 0.000766±0.000941 for sEOAD cases. Single-factor ANOVA found variance between groups to be significant to $p=3.23\times10^{-5}$. Predictive ability was calculated using R package PRROC with an AUROC of 57.9% and AUPRC as 55.8%.

Decile scoring was used at this stage to compare the distribution of individuals between synaptic PRS to whole genome PRS from both arrays before the inclusion of covariates. The average change in decile for individuals on the NeuroChip was less than 2 (74%) with 90 individuals falling within the same decile. The NeuroX showed greater consistency of decile ranking between whole genetic and synaptic PRS with most falling within 2 deciles (82%) and 188 samples with no change (Table 13).

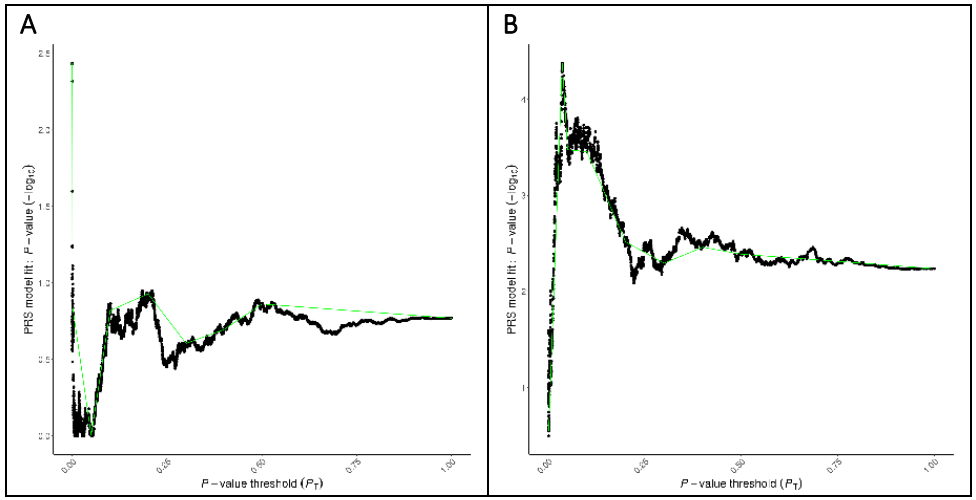


Figure 11: High-resolution plot of synaptic PRS modelling of LOAD and sEOAD with the NeuroChip and NeuroX

The figures outline the resulting p-values at each tested threshold for the NeuroChip and NeuroX target datasets. **A.** The NeuroChip shows the highest peak with the most significant variants included with later peaks of smaller effect. **B.** The NeuroX figure shows predictive models recorded using significant SNPs with a steady reduction in predictive ability as less significant variants are introduced.

The best models were then tested with the re-introduction of SNPs associated with *APOE* ε status (rs429358 and rs7412) and including sex as a covariate. Introduction of the effect of *APOE* has been known to improve models in previous results and was present on the list of genes expressed at the synapse but excluded from analysis due to the nature of the locus. The NeuroChip best model was improved by the inclusion of *APOE* SNPs, the best model remained at the same threshold with Nagelkerke's R^2 calculated as 0.306 (p=8.88×10^{-24}); with addition of sex as a covariate, Nagelkerke's R^2 was calculated as 0.308 (p=1.27×10^{-23}). Predictive probability scores were generated in SPSS from logistic regression with the inclusion of sex as a covariate; mean synaptic PRS with *APOE* and sex was calculated as 0.711±0.221 for LOAD cases and 0.478±0.173 for controls, single-factor ANOVA found the significance of variance between groups to be p=6.27×10^{-35}. Predictive ability was calculated as 78.1% (AUROC) and 85.5% (AUPRC).

The NeuroX best model also showed improvement with *APOE* SNPs: Nagelkerke's R^2 was found to be 0.187 (p=2.5×10^{-24}), however, a greater Nagelkerke's R^2 was also calculated using fewer SNPs than previously. The NeuroX best model threshold was lowered to 0.02027, the same threshold identified in the combined model, where Nagelkerke's R^2 was found to be 0.191 (p=9.05×10^{-25}). When sex was included as a covariate of the original best model at P_t=0.036111, Nagelkerke's R^2 was 0.185 (p=3.25×10^{-24}). Mean synaptic PRS with *APOE* and sex calculated as a predictive probability in SPSS was 0.409±0.164 for controls and 0.563±0.194 for sEOAD cases, variance between groups was calculated and found to be significant to p=9.76×10^{-33}. Predictive ability using both methods were calculated as an AUROC of 72.5% and AUPRC of 71.8%.

Decile	1	2	3	4	5	6	7	8	9	10
NeuroChip Control	4	17	47	52	36	27	15	7	8	4
NeuroChip LOAD	3	14	15	34	46	35	64	54	57	36
NeuroX Control	11	41	89	99	64	53	38	25	10	6
NeuroX sEOAD	3	15	35	51	41	52	54	69	53	35

Table 13: Distribution of synaptic PRS samples by decile scoring

The table shows the results of decile scoring of PRS with *APOE* SNPs and sex on the NeuroChip controls and LOAD cases, and NeuroX controls and sEOAD cases. The number of individuals which fall within each decile are given split by array and disease group. The distribution shows a greater number of control samples in the lower deciles and case groups are predominately found in the upper deciles.

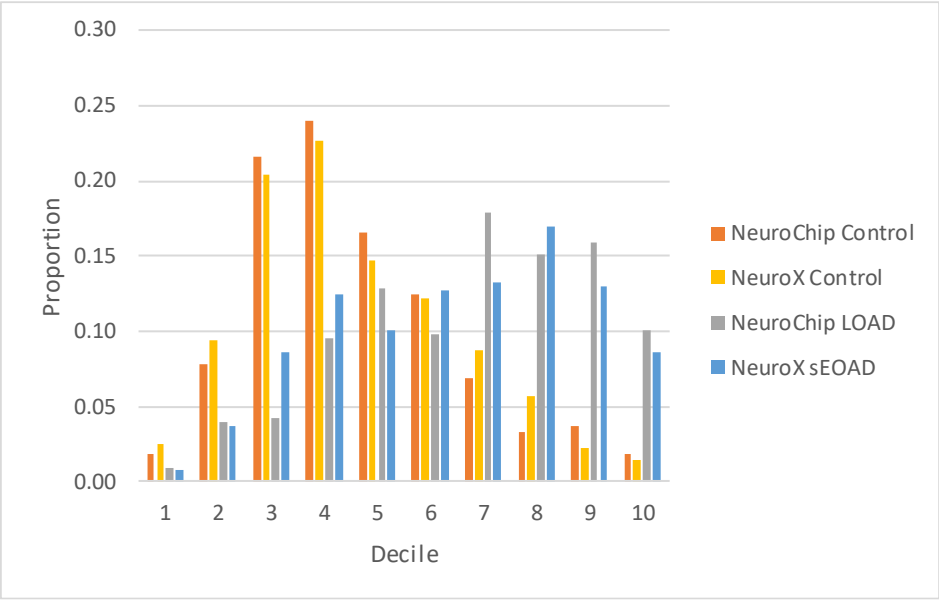


Figure 12: Distribution of synaptic PRS by decile scoring

The figure shows the proportion of cases and controls from each array based on PRS with *APOE* SNPs and controlling for sex in the combined model. Proportion was used as a replacement of distribution to counter the difference in numbers of samples per dataset; proportion was calculated as the number of samples within a decile divided by number of samples across all deciles for each test group. The NeuroChip (orange) and NeuroX (yellow) controls show similar representation at each decile with a positive skew and peak in the fourth decile. The NeuroChip (grey) cases show more negative skew than controls and a peak in the seventh decile; NeuroX (blue) cases similarly to LOAD cases show a larger distribution, with most samples proportionately in the eighth decile.

5.4 SNPs

Breakdown of the 4 SNPs identified in the NeuroChip model identified 3 synaptic genes (*BIN1*, *PICALM*, *PTK2B*) sufficient to calculate risk of LOAD, as two independent SNPs were from the same gene (*BIN1*). Comparing to the previous studies on the NeuroChip, one of the SNPs from within *BIN1* was common to a previous synaptic PRS model (rs35114168) and the SNP within *PTK2B* was also present in the imputed NeuroChip whole genome best model (rs28834970).

There were 215 genes identified from 376 SNPs identified in the NeuroX synaptic PRS model used to predict risk of sEOAD from synaptic genes as 79 genes had multiple SNPs within their loci represented. The most recurring gene had 12 independent SNPs within its gene range (*DLG2*). When compared to previous studies there were 3 SNPs from 3 genes in the best model also present in the synaptosome study (*DLG2* rs286043, *MINK1* rs8078173, *BIN1* rs35114168); 263 of the SNPs present in the synaptic PRS model were also identified in the imputed NeuroX whole genome dataset. All 4 SNPs identified in the NeuroChip best model were present within the NeuroX best model.

The combined model was made up of 199 SNPs from 138 genes, with 101 genes represented by a single SNP and the remaining 37 genes represented by 2-5 independent SNPs. The most recurring genes were also identified in the NeuroX gene list as well as the published synaptic best model list. The SNPs within the model contained 180 of the SNPs identified in the NeuroX best model which included the 4 SNPs from the NeuroChip best model; most of the remaining 19 SNPs were reported in less predictive models from the NeuroChip whilst 1 SNP was not present in either analysis initially and only remained due to the selection and clumping criteria (*NGEF* rs778357).

Association testing was completed using PLINK to determine the observed effect of variation within the combined dataset and compared to the associated effect (β) recorded in the base dataset, IGAP summary statistics. PLINK results provided an odds ratio and significance for each SNP based on differences in frequency of the effect allele between cases and controls. SNPs with an OR > 1 were associated with increased risk and < 1 were considered protective variants; in the IGAP summary statistics, β-values < 0 were considered protective and > 0 were associated with increased risk. Comparing between the datasets, 125 of the 201 SNPs (199 best model and 2 *APOE*) agreed upon whether the SNP was associated with risk or protection.

5.5 Discussion

Imputed datasets were used instead of the genotyped datasets to maximise the number of common SNPs between each target dataset and the base dataset (IGAP summary statistics) and between both datasets; SNP numbers were expectedly reduced as quality controls were carried out and when clumping was introduced. The NeuroChip was re-analysed on PRSice-2 to compare the results with a greater number of samples and the effect of imputed data to genotyped data. NeuroX analysis was included to compare the effects of synaptic genes on a concordant but separate phenotype. A combined dataset was tested to observe the correlation between PRS of sEOAD and LOAD risk.

Synaptic PRS was analysed as part of a collaboration project, the use of a selection of genes is becoming a popular method of breaking down whole genome polygenic risk. The inclusion criteria leads to a more precise score based on a certain network of genes; this approach can be replicated using different selection criteria. An example of this is the selection of genes expressed by microglia, which can provide a more specific order to the effect of genetic variation and subsequent protein expression

within microglial cells on risk of AD. Future work included PRS based on genes expressed in microglia and the pathway genes which encode proteins in the protein-lipid complex.

The NeuroChip best model was identified at a similar but lower threshold to analysis using the first two batches of the NeuroChip, the improvements in modelling may be reflective of a better distinction between cases and controls could be achieved with a greater number of samples. In a whole genome PRS, a secondary aim was to identify genes not previously researched or identified to be associated with risk of AD, the use of a selective PRS already recognises these genes and is therefore driving the identification of SNPs within these genes with the greatest association with disease risk.

The best model threshold determined from the NeuroX array is within a level of significance of (p<0.05) but not near significance levels to consider all SNPs to have a genome-wide level of significance ($p<1\times10^{-8}$). The genes identified within the model are a reduced number than those from the inclusion criteria as well as those represented in the target dataset, which suggests some genes expressed at the synapse are more associated with disease risk than others.

Decile scoring between PRS models highlighted correlation between whole genome PRS and synaptic PRS for most individuals, however, there were some individuals who ranked higher in synaptic PRS than whole genome. This observation in controls may be due to individuals carrying many of the risk variants present within the synaptic model with fewer calls in other risk SNPs or may be subject to some synaptic dysfunction with no phenotypic effect. Cases with much greater risk in synaptic PRS than whole genome PRS may indicate the disease pathology for those individuals was driven by dysfunction in the synapse. It may be noteworthy to explore the clinical

features of these individuals more specifically to identify the process by which synaptic dysfunction can lead to AD diagnosis.

Calculating predictive ability for the NeuroChip and NeuroX independent models can provide a more direct comparison of synaptic PRS models to whole genome PRS. Although both AUROC and AUPRC were calculated for all models, the use of AUPRC was specifically for NeuroChip modelling as there is a greater difference in proportion of cases to controls than with the NeuroX. The results showed the synaptic PRS with *APOE* and sex for the NeuroChip was more predictive (AUPRC=85.5%) than the results of whole genome PRS with *APOE*, sex and age at death for the NeuroChip (81.5%). The predictive ability of synaptic PRS with *APOE* and sex for the NeuroX (AUROC=72.5%) however was less predictive than whole genome PRS with *APOE* and sex for the NeuroX (73.0%). The predictive ability for the combined model PRS with *APOE* and sex was calculated using AUROC and AUPRC as the proportional difference between cases and controls is slightly reduced but the imbalance may still be a factor affecting predictive ability, as shown with a slightly greater value of AUPRC (77.5%) than AUROC (74.2%). A predictive ability of 77.5% falls within the boundaries of a useful predictive model for clinical utility.

Identification of multiple SNPs within the same gene is a stronger indicator of association of the gene with disease status; the number of variants, which may not all be present within the same individual, may identify the dysfunction of such proteins produced from these genes would influence disease pathology. Future work would include research individual SNPs from all models and how their variation affects disease pathology.

The comparison of the synaptic PRS to whole genome PRS identified that not all variants present in the synaptic PRS were present in the full model, which may have

been expected. This result may be due to the clumping algorithm removing such SNPs from the whole genome PRS due to a more significantly associated SNP being present within the tested LD window. It could therefore be considered that there may be SNPs outside the gene ranges that are affecting disease risk. The possibility of this could be further explored by adjusting the gene ranges to include SNPs up- and downstream of the gene to identify those with a regulatory effect on gene expression; although there is no distinction of how far away a regulatory SNP can be from a gene to an effect.

The trends of decile scoring of the combined model show correlation between the distribution of controls in both datasets, the combined cases also represented a greater risk than controls. There is a more apparent cross-over amongst samples in the synaptic PRS analysis than the whole genome PRS analysis of individual arrays, which is also confirmed by predictive ability calculated as an AUROC of 74.2% and AUPRC of 77.5%. Decile scoring was useful at identifying outliers, inferring in some cases synaptic dysfunction may be a standalone risk factor in AD pathogenesis.

Association testing of the best model SNPs can validate whether the associations calculated in a meta-analysis match the associations made in a smaller dataset. Although many SNPs matched associated effect, the effect sizes would not be considerably similar due to the numbers of cases and controls used in the base dataset providing a more accurate representation of the effect in the European population. In many of the SNPs which did not agree with IGAP, cases and controls in the target dataset shared similar frequencies of effect alleles; this is likely due to the SNPs used in PRS analysis representing common variation, with a minor allele frequency > 5%.

The results of this study identify a correlation between polygenic risk from synaptic genes and whole genome PRS. The results provide a starting point for further research of the previously unreported genes such as *DLG2*.

6 Polygenic risk scoring as a predictor of Alzheimer's disease

6.1 Perspective

Polygenic risk scoring was conducted on a cohort of late-onset Alzheimer's disease cases and neurologically healthy controls from the Brains for Dementia Research (BDR) resource, based on summary statistics from the International Genomics of Alzheimer's disease Project (IGAP) meta-analysis (Chapter 3). PRSice-2 was then used to additionally generate PRS on living samples genotyped on the NeuroChip array, either with an initial diagnosis of mild cognitive impairment (MCI) or undiagnosed. The samples were modelled based on the distribution of LOAD cases and controls to predict likelihood of developing AD.

Previous analysis of this cohort was conducted using an earlier version of PRSice-2 software on samples from the first two batches of the BDR cohort and individuals diagnosed with MCI recruited by the inflammation, cognition, and stress (ICOS) longitudinal study. MCI samples were tested against the best model determined for LOAD cases and controls and categorised based on conversion to AD during the period covered by the study. The results showed conversion from MCI to AD could be predicted (area under the precision-recall curve; AUC) with 61% accuracy, with a significant increase in PRS across diagnosis groups from control > non-converter > converter > LOAD (Chaudhury et al., 2019). When ICOS samples were sorted into deciles of increasing risk, a greater proportion of patients who converted to AD were observed at higher deciles with the opposite being seen in non-converting MCI patients, similar to the trends observed between LOAD cases and controls, suggesting an association between genetic risk of disease and likelihood of conversion to AD.

The primary aim of this study was to determine whether the genetic risk of AD in living patients, either undiagnosed or with MCI, based on an AD case v control model can be used to predict likelihood of developing AD using updated PRS analysis software.

6.2 Samples

DNA was isolated from blood and/or brain tissue from all BDR (n=1172) and ICOS samples (n=128). Lab procedures were carried out by T. Patel following methods as described (Section 2.1). All samples were genotyped using the NeuroChip array. Quality control of the genotyping data was completed using GenomeStudio v2 and PLINK (Section 2.3.3). Samples were checked for call rate, relatedness, diversion from European ancestry, and heterozygosity; SNPs were controlled for call rate and adherence to Hardy-Weinberg equilibrium. As a result of quality controls, there were 404 living, undiagnosed samples available for analysis from the BDR resource and 124 samples with a clinical diagnosis of either MCI or confirmed conversion to dementia from the ICOS study (Table 14), genotyped for 477,720 SNPs. SNPs associated with *APOE* status were imputed based on chromosome 19 genotype data.

Resource	Phenotype	N	Females (%)	*APOE* ε4+ (%)	*APOE* ε4ε4 (%)
BDR	Undiagnosed	404	252 (62.4)	109 (27.0)	7 (1.7)
ICOS	MCI	124	48 (38.7)	55 (44.4)	9 (7.3)

Table 14: Demographics of predictive PRS samples

BDR-recruited individuals were recruited from across the UK: from universities in Bristol, Manchester, Newcastle, Oxford, and King's College London; ICOS study samples were recruited from Southampton. The number and percentage of females in each group is also shown as well as the number of individuals with at least one *APOE* ε4 allele and additionally those who were ε4ε4 homozygotes.

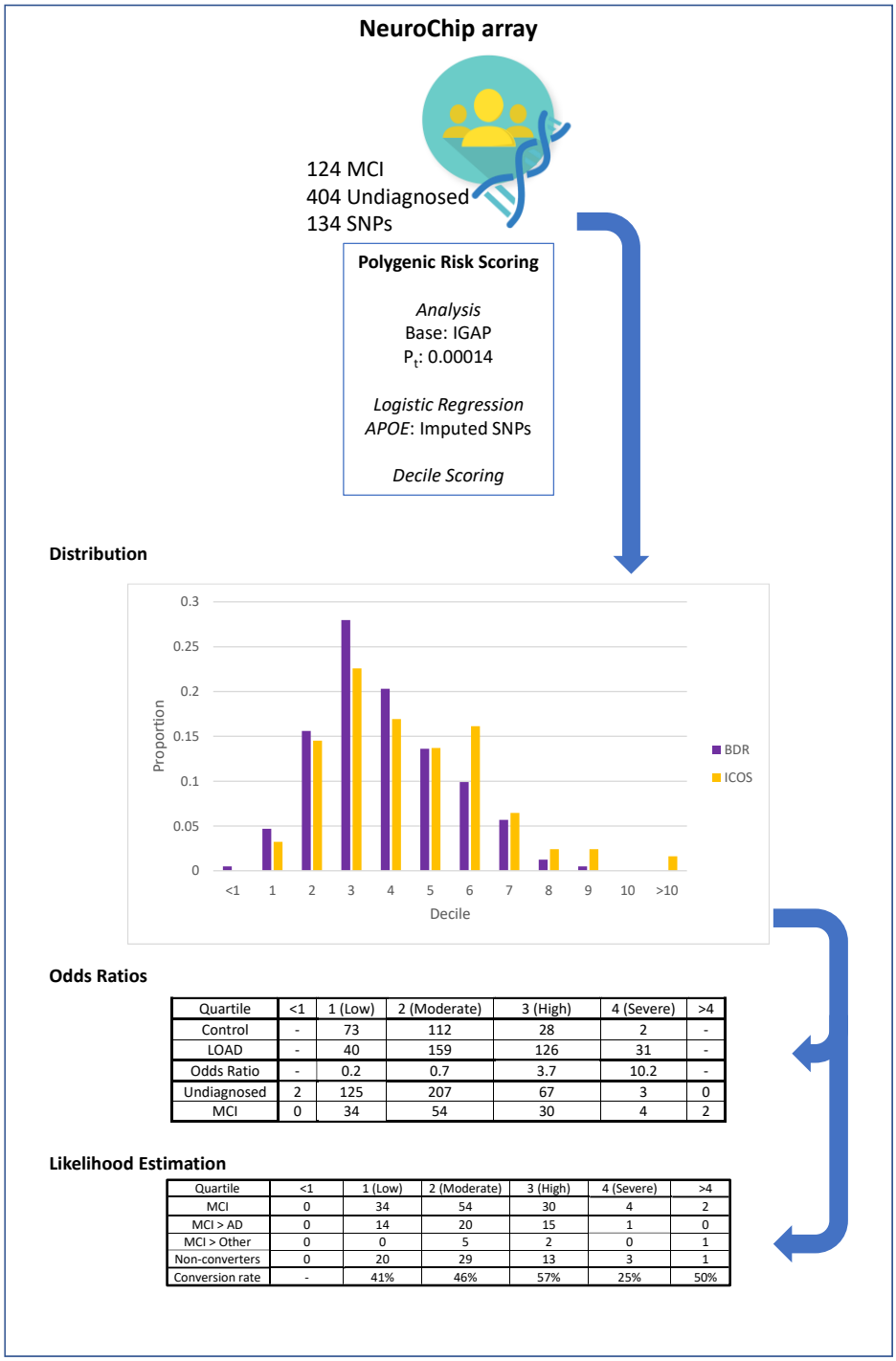
6.3 Polygenic risk scoring

Predictive PRS analysis (Figure 13) was tested on 404 undiagnosed individuals and 124 MCI cases using the SNPs present at the best model determined between pathologically confirmed 358 LOAD cases and 217 controls (Chapter 3), followed by the inclusion of *APOE* ε status SNPs. Individuals were given arbitrary phenotypes of case or control status for PRSice-2 to generate PRS and the model fitness and significance results were disregarded. As the results of model were being disregarded, covariates were not introduced to modelling through logistic regression analysis. The target dataset was formed using PLINK by extracting the SNPs which were present in the best model and *APOE* was included by later merging with a dataset of calls for both SNPs of target samples.

6.3.1 Summary

Mean individual PRS at the best model was calculated as 0.00121 with a standard deviation of ±0.00264. The model was then tested with the introduction of SNPs associated with *APOE* ε status (rs429358 and rs7412), where mean PRS with *APOE* was calculated as 0.00264±0.00399.

Further statistical analysis included decile scoring, based on the boundaries determined by the PRS of LOAD cases and controls. The samples were then separated by phenotype to distinguish between undiagnosed and MCI cases, given in Table 15, and presented in Figure 14. The results show distribution of samples similar to previous observations of distribution amongst combined groups, representing the diversity in scores which may be observed in a general population.



Quartile	<1	1 (Low)	2 (Moderate)	3 (High)	4 (Severe)	>4
Control	-	73	112	28	2	-
LOAD	-	40	159	126	31	-
Odds Ratio	-	0.2	0.7	3.7	10.2	-
Undiagnosed	2	125	207	67	3	0
MCI	0	34	54	30	4	2

Quartile	<1	1 (Low)	2 (Moderate)	3 (High)	4 (Severe)	>4
MCI	0	34	54	30	4	2
MCI > AD	0	14	20	15	1	0
MCI > Other	0	0	5	2	0	1
Non-converters	0	20	29	13	3	1
Conversion rate	-	41%	46%	57%	25%	50%

Figure 13: Analysis pipeline for predictive PRS of undiagnosed and MCI samples

The figure outlines the process of predictive polygenic risk scoring of a cohort of undiagnosed BDR-recruited samples and MCI cases from the ICOS study. Analysis was run on the NeuroChip-genotyped samples, PRS were generated based on the best model determined in analysis of LOAD cases against controls. Samples were distributed according to decile scoring used in previous analysis as well as classification into tiers of increasing risk from low, moderate, high to severe risk. Estimates were made of number of expected AD cases in undiagnosed controls; estimated AD case numbers in MCI cases were compared to observed conversion rates from the ICOS longitudinal study.

Decile	<1	1	2	3	4	5	6	7	8	9	10	>10
Controls	-	6	26	33	32	37	37	7	24	11	2	-
LOAD	-	5	11	21	22	32	51	59	72	44	39	-
Undiagnosed	2	19	63	113	82	55	40	23	5	2	0	0
MCI	0	4	18	28	21	17	20	8	3	3	0	2

Table 15: Distribution of predictive PRS samples by decile scoring

The table shows the results of decile scoring on the BDR cohort LOAD cases and controls (Results 1) alongside distribution of undiagnosed BDR samples and MCI ICOS samples. The number of individuals within each decile, determined from LOAD cases and controls, is given; outliers of the decile boundaries from the undiagnosed groups were included. The results show the greatest proportion of undiagnosed individuals and MCI cases in the third decile.

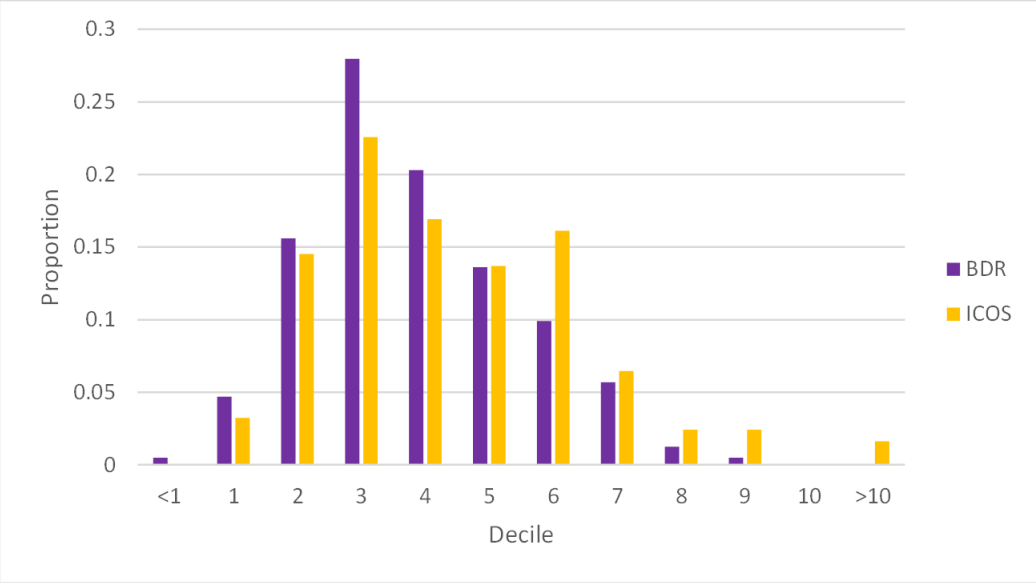


Figure 14: Distribution of predictive PRS by decile scoring

The figure shows the proportion of BDR undiagnosed (purple) and ICOS MCI (yellow) samples when sorted into deciles according to observed limits in LOAD cases and controls. Proportion was used as a replacement of distribution to counter the difference in numbers of samples per cohort; proportion was calculated as the number of samples within a decile divided by number of samples across all deciles for each test group. Undiagnosed BDR samples show normal distribution with a positive skew which peaks at the third decile. The ICOS MCI samples show bimodal distribution with the first peak in the third decile and another in the sixth decile. Although no samples were present in the tenth decile, there are MCI samples with a higher score than any BDR LOAD case.

6.3.2 Risk prediction

Prediction of disease status in undiagnosed individuals may be achieved by considering the odds ratio observed in LOAD v Control. Odds ratio is calculated by identifying the number of cases and controls which fall within and outside a selection criterion. With the number of samples present within each decile being too few, the alternative was to categorise samples by quartiles which may be indicative of low, moderate, high, and severe risk according to polygenic risk (Table 16). The results show individuals whose PRS falls within the first quartile have a 1 in 5 chance of developing AD whilst individuals whose scores fall within the fourth quartile have more than 10× greater risk of developing AD.

Quartile	<1	1 (Low)	2 (Moderate)	3 (High)	4 (Severe)	>4
Control	-	73	112	28	2	-
LOAD	-	40	159	126	31	-
Odds Ratio	-	0.2	0.7	3.7	10.2	-
Undiagnosed	2	125	207	67	3	0
MCI	0	34	54	30	4	2

Table 16: Distribution of samples by quartile scoring

The table shows the results of the BDR cohort LOAD cases and controls (Results 1) alongside distribution of undiagnosed BDR samples and MCI ICOS samples into quartiles of increased risk. An odds ratio is calculated for each quartile based on the number of LOAD cases and controls present within. The number of individuals within each quartile, determined from LOAD cases and controls, is given; outliers of the boundaries from the undiagnosed groups are included. Undiagnosed BDR individuals who score below the 'low' risk category have a further decreased likelihood of developing AD and MCI cases with a PRS greater than the 'severe' risk category have an even greater likelihood of developing AD.

Predicting probability of AD or other dementia sub-types in sample groups from odds ratio was attempted by categorising into tiers of risk. From the observed odds ratio it was calculated that up to 21 of the 125 undiagnosed BDR samples with 'low' risk, around 85 of the 207 with 'moderate' risk, up to 53 of the 67 with 'high' risk, and all 3 individuals with 'severe' risk would expect to develop AD. With MCI cases, the expected outcome according to odds ratios would be up to 6 of the 34 cases with 'low' risk, around 22 of the 54 cases with 'moderate' risk, up to 24 of the 30 cases with 'high' risk, and all 6 cases with 'severe' or higher risk would develop AD.

Based on available data on conversion to AD and other dementia sub-types throughout the ICOS study, comparisons could be made regarding the predicted probability of and resulting diagnosis. It was reported that within 36 months, 50 the 124 MCI cases converted to AD and 8 converted to other dementia sub-types (dementia with Lewy bodies=2, vascular dementia=2, mixed dementia=4). The rate of conversion was found to be 47% overall; the highest conversion rate was observed in those within a 'high' risk category (57%) with those in 'low' and 'moderate' risk categories converted below the overall conversion rate (Table 17).

Quartile	<1	1 (Low)	2 (Moderate)	3 (High)	4 (Severe)	>4
All MCI	0	34	54	30	4	2
MCI > AD	0	14	20	15	1	0
MCI > Other	0	0	5	2	0	1
Non-converters	0	20	29	13	3	1
Conversion rate	-	41%	46%	57%	25%	50%

Table 17: Distribution of MCI conversion by quartile scoring

The table shows the results of MCI ICOS samples spread into quartiles of increased risk, further categorised by clinical diagnosis 36 months from recruitment. The number of individuals within each quartile, determined from LOAD cases and controls, is given; outliers of the boundaries from the MCI samples are included. Conversion rate was calculated based on the proportion within each quartile with reported conversion. Non-converters are considered those who continue to suffer from MCI or are yet to convert to dementia.

6.4 Discussion

PRSice is typically used to model the difference between binary phenotypes and with absence of this in undiagnosed samples, the software has limited function and use for predicting risk. In previous work on the ICOS MCI cohort, the difference in PRS between converters and non-converters was tested but the predictive ability was much lower compared to LOAD cases v controls; this was primarily due to all MCI cases having symptoms of neurodegeneration and could not be classed as neurologically healthy controls; sample numbers for this analysis are low and there may not be sufficient power to predict. Risk categorisation was an alternative approach to identifying more at-risk individuals than identifying genetic variation associated with conversion.

The use of case v control can be informative in determining where individuals lie on a scale of risk, and previous analysis which confirms the correlation between score and risk of LOAD provides the ability to make predictions. The results are, however, based on a limited number of samples and therefore the confidence of these predictions cannot be entirely accurate, given that there are more undiagnosed individuals in the BDR (n=404) than either LOAD cases (n=358) or controls (n=217).

As the scale by which undiagnosed samples were compared was limited by the range of scores observed in the BDR cases and controls, there were instances where individuals fell outside the predetermined boundaries. As more LOAD cases and controls are genotyped and tested in the future, the boundaries specific to this model and the SNPs involved may change. An alternative approach to setting the limits would be to generate the PRS of a positive and negative control for the model; a positive control would harbour all the risk variants identified in a model with no protective variants whilst the negative control would represent the opposite. These

boundary samples would then provide the definitive limits for scores individuals could have, however unlikely an individual is to harbour all risk or protective variants, if that were biologically possible.

Although the expected number of AD cases from MCI samples based on odds ratios and the observed number of conversions to dementia in the cohort was identical (n=58), there was a discrepancy with the expected conversions from each risk group. The study showed more individuals with 'low' and 'moderate' risk converted to dementia than expected whilst fewer MCI cases with 'moderate' or greater risk converted. It could be suggested that genetic risk of LOAD would not be translatable for determining likelihood of disease for MCI cases and could be due to other factors which led to cognitive impairment which were not identifiable in this model, such as environment and lifestyle; alternatively, further study may be required to identify genetic risk of MCI by comparing with controls directly. Conversely, a greater number of individuals from this cohort may go on to develop AD within their lifetime as those with scores similar to AD cases have sufficient measurable risk; the potential benefit of therapies or preventative measures for this cohort needs to be realised.

Categorising individuals into tiers of risk was more suitable than using decile scoring especially when calculating odds ratios, given the sample sizes. In the ICOS MCI cohort, it was observed that those within the higher tiers had already converted; predictions could still be made based on the remaining non-converters for who would still be likely to convert in the future. As the conversion rate increases with risk between 'low' and 'high' risk, the trend would suggest more conversions may be expected in the 'severe' risk category as well as the non-converting individual with a PRS higher than any LOAD cases reported.

The resulting estimates for undiagnosed BDR samples suggests up to 162 individuals may develop AD within their lifetime (40.1%), higher than the lifetime risk estimates. Based on the distribution of these samples in Figure 14, the observed spread was like that seen in controls; this may suggest the number of potential AD cases was over-estimated. It would be impossible to identify which specific individuals within each risk tier would develop AD, as the estimations were based on odds ratios of sub-groups. Despite this, the benefit of calculating risk and categorising individual into risk tiers would help direct future preventative measures and treatments before the onset of symptoms for the most at-risk individuals. With the models by which individuals are measured likely to improve with more case and control data, the precision of identifying at-risk individuals would continue to improve.

7 Discussion

7.1 Summary

The objective of the project was to investigate polygenic risk scoring in AD using available genotyped data of the BDR cohort, the ICOS study and the sporadic early-onset AD study samples. BDR is a growing initiative aiming to develop a comprehensive cohort of individuals for research in AD and other dementia sub-types, providing additional clinical observations from post-mortem examination. ICOS was a longitudinal study observing the effects of diagnosis of mild cognitive impairment on the likelihood of conversion to AD and other dementia sub-types. The sEOAD study dataset was compiled by institutions across the UK to identify genetic variation which led to AD beyond that observed in the early-onset familial form. PRS was generated to model the risk and protective variants in case-control data based on the observed associations in a larger cohort to identify and validate the common variation involved in AD susceptibility.

Polygenic risk scoring was previously tested on a finalised dataset of NeuroX-genotyped sEOAD cases and controls to determine the effectiveness of the PRSice software and methods for discerning between pathologically confirmed case-control data. Following the establishment of the next-generation of Neuro- arrays, genotyping and quality control was performed on the NeuroChip-genotyped BDR cohort and ICOS study samples, made up of three 'batches' of recruitment phases. Quality control and dataset curation was completed after each phase to generate preliminary data as well as testing the entire ICOS study cohort after its conclusion.

Using finalised datasets and tested methods, PRS was carried out on 358 LOAD cases and 217 controls from the BDR cohort genotyped on the NeuroChip array (Chapter 3), 408 sEOAD cases and 436 controls genotyped on the NeuroX array (Chapter 4), 358

LOAD cases, 408 sEOAD cases and 653 controls using combined imputed genotype data for a subset of genes expressed at the synapse (Chapter 5), and 124 MCI cases and 404 undiagnosed samples genotyped on the NeuroChip array based on results from LOAD v Control PRS (Chapter 6).

The first batch of the BDR cohort consisted of 248 LOAD cases and 105 controls, which increased to 302 LOAD cases and 128 controls with the second batch and finalised with 358 LOAD cases and 217 controls. The results in Chapter 3 show a predictive model for distinguishing between cases and controls was achieved from genotyped data at a p-value threshold of 1.4×10^{-4} using 134 SNPs; inclusion of *APOE* SNPs and covariates for sex and age at death improved to the model whilst TaqMan assayed GWAS hits had less effect as most SNPs were represented in the model by proxy SNPs. The greatest predictive ability calculated as AUPRC was 81.5%, sufficiently predictive to be used in a clinical setting.

The NeuroX dataset, consisting of 408 sEOAD cases and 436 controls, was tested using an improved version of PRSice and alternative methods to compare results to previous studies of the cohort. The results in Chapter 4 found the most predictive genotyped model to be identical to the identified model of PRS alone, subject to further improvement with inclusion of *APOE* SNPs and sex as a covariate. Imputation of the NeuroX dataset implicated additional loci in the model but showed a similar level of predictive ability to the genotyped dataset. The most predictive model was identified as the genotyped model with the inclusion of *APOE* SNPs and sex as covariate, calculated as AUROC to be 73.0%; this may not be sufficient for use in a clinical setting, however the additional genes implicated in the model provide avenues for further research into the AD sub-type.

Synaptic PRS analysis was completed to identify predictive models in the imputed NeuroChip and NeuroX datasets independently, to predict risk of AD from variation in genes expressed at the synapse. An additional model was formed from combining both cohorts by SNPs common to both imputed datasets. The results in Chapter 5 showed a predictive model was found for both independent target datasets, neither was as predictive as whole genome PRS. The combined dataset also showed predictive ability, the distribution of samples by this measure also showed some divergences to the trends seen in whole genome PRS, which may be of note. Additionally, genes previously not considered to be associated with AD risk have been identified and require more research.

The best model determined in Chapter 3 was then used to calculate PRS in individuals without AD or control diagnosis to predict likelihood of AD, this included 404 undiagnosed, BDR-recruited individuals and 124 individuals with a prior diagnosis of mild cognitive impairment. The results of Chapter 6 found a large distribution of both sample groups when split into deciles and subsequently categorised into quartile tiers of risk. Odds ratio of developing AD was calculated from LOAD case v control and used to estimate probability of AD in undiagnosed groups, estimates for the MCI cohort were also compared to observed clinical diagnosis of conversion to AD or other dementia sub-types which suggested and under-estimation with the possibility of other potential later conversions.

7.2 Genotyping

The quality control of NeuroChip genotyped data was the most time-consuming stage of the study; from the point of receiving the raw IDAT data files to having a fully curated dataset for analysis required many months of work. This stage was repeated each time a new batch was received as additional samples would produce clearer

clusters for SNPs and improve overall quality from the previous dataset. By the third batch, many of the SNP allocations were clearer and the process required less time. The improvements from repeating quality controls with previous batches when the final batch was received led to the re-introduction of samples which were previously observed to have low call rate and excluded in the first two batches. The cluster file produced from successful quality control of over 1000 samples on the NeuroChip can be useful for any other studies using the genotyped NeuroChip data, saving time and improving the ability to assign calls.

The genotyped data for the NeuroX was already produced and curated and all analysis was completed based on the assumption all SNPs calls were accurate and under the same level of scrutiny as with NeuroChip genotyping. Imputation provided the option to find calls for a greater number of SNPs, based on the genotyping, some of which were useful for generating PRS and many more present in both imputed NeuroChip and NeuroX datasets. This highlights the utility of research approaches to include imputation of genotyped data for analysis to get the most out of arrays data and gives rise to the potential of merging datasets to form larger cohorts for meta-analyses.

7.3 PRSice

PRSice was used throughout the study as polygenic risk scoring software, despite other methods and software being available. The general method for deriving any PRS comes from using the PLINK score function, however, the PRSice script is written to test multiple thresholds using computational processing speed to generate and test more data than testing preselected thresholds or SNPs. Other PRS analysis methods include testing a set number of SNPs such as only GWAS hits, which is often referred to in literature as a polygenic hazard score. Although the primary aim is to develop a functional predictive model using this approach, the use of PRSice can lead to greater

understanding of the effects of variation across the genome, independent of the significance seen in more researched and understood SNPs.

There are some major caveats to using PRS which may be addressed in future updates as the field of genomics continues to grow. Most GWAS look at the effects and associations seen in autosomal chromosomes (1-22) and do not account for how variation in allosomes or mitochondria may affect risk of disease. The developers of PRSice are continuing to work on the inclusion of non-diploid chromosomes into PRS analysis; this also remains impeded by the absence of association data for these chromosomes in meta-analysis summary statistics, which are used as base datasets. It is reasonable to predict that as the field of bioinformatics and computational statistics develops, this issue will be resolved; for now, it remains best practice to develop genetics models and control for sex with logistic regression.

As genomics further developed, it had become more apparent that some genetic variation is not bi-allelic in nature. This caused issues with modelling risk as the current approach to computing risk involves identifying a polymorphism, the major allele and then calculating effect and association based on the frequency of the effect allele. In cases where a polymorphism is multi-allelic, the effect the variation has on overall disease risk may vary depending on the nucleotide change, which is why PRSice actively identifies and removes SNPs it observes to be multi-allelic in nature. PRSice does this by comparing the record A1 (major) and A2 (effect) alleles in the base, target, and LD datasets to confirm a match. This works in most instances however arrays like the NeuroChip seek to identify the nucleotide change for some known multi-allelic SNPs by genotyping the point of variation for the presence of whichever allele the individual harbours, only for PRSice to disregard these SNPs from analysis entirely. Online SNP databases like dbSNP identify many of the SNPs used in modelling to be multi-allelic, which may suggest they should be disregarded from the

study post-hoc, but it can also be justified that these SNPs were introduced with confidence that the observed effect alleles match the effect allele of the other datasets and the recorded association is bi-allelic.

The default settings for PRSice-2 included intervals of testing to a million models and testing between the p-value thresholds (P_t) of 1×10^{-6} to 1. PRSice-2 can however test at smaller intervals than this too, at intervals of 10^{-10}, and produce 10 billion models of PRS starting from 1×10^{-10} to compare between. This effect may have led to alternative models being identified for any results, as it would primarily affect the models in the lower boundaries of testing (between 1×10^{-10} and 1×10^{-6}). The defaults were used across all analysis to follow procedures advised by PRSice, and the presence of few SNPs in any quality-controlled target datasets with significance ≤ 1×10^{-6} (NeuroX = 10, imputed NeuroX = 25, NeuroChip = 16, imputed NeuroChip = 27). Further expansion of analysis in future work may address the potential of generating risk in more significant models, one suggestion may be to only model SNPs which are considered to have reached genome-wide significance ($p\leq1\times10^{-8}$).

7.3.1 Decile scoring

There were many appropriate approaches to categorise sample groups by risk and the most used was distribution of samples into deciles. This approach can also vary by separating the number of samples into tenths instead of the dividing the score into tenths, so the proportion within each cohort is the same and the average score of each decile would then not be equidistant. This approach may give additional information about the nature of PRS in the cohort, and could be tested on case proportion, control proportion and proportion of all individuals. However, the approach to split by score and observe the relationship of decile to proportion of cases or controls proved effective.

The changes to methods seen with using PRS as a predictor (Chapter 6) saw the use of quartiles instead of deciles or even percentiles to break down the cohort into smaller risk groups which can give risk by score, odds ratio, with greater confidence. This could have still been measured using deciles, but the number of samples within each decile was not sufficient to calculate an odds ratio which would appropriately reflect the risk observed by individuals within.

It was previously mentioned that it may be more useful to compare PRS of individuals for a specific model using a positive control with maximum risk and minimum protection and a negative control with minimum risk and maximum protection (according to the base dataset) to set the upper and lower boundaries of decile scoring, respectively.

7.4 Outcomes

The results from Chapter 3 identify a predictive model for determining between LOAD cases and controls when controlling for sex and age at death. With other factors included in modelling, this may improve further but without the consideration of covariates, a model of genetic risk including *APOE* could be utilised for predicting risk in individuals. This model is likely to change in future as more samples are recruited and improvements to the PRS methods continue, as seen with the results published in 2019 differ from those available now. The eventual aim is to develop a model which can be derived from across multiple arrays or from a limited number of assays to determine a consistent and accurate risk.

7.4.1 Gene discovery

The SNPs present in imputed models for both LOAD and sEOAD were identified, the genes of exonic SNPs and reported in Table 18. Many of the genes identified have previously been implicated in AD risk; newly reported genes contain SNPs with a p-

value of significance ≤ 2.1×10^{-5}. These genes warrant further study to understand their involvement in AD pathology. Intergenic SNPs were also found in both models, alongside SNPs in long intergenic non-protein coding RNA (LINC) and uncharacterised locus (LOC) regions. Whilst these variants are not present in gene-coding regions, their presence suggests involvement in AD pathology.

Chromosome	Genes
1	VAV3, RGS5, KIF21B, IL19, CR1, CR1L, SMYD3
2	BIN1, PPP1R1C, MYT1L, KANSL1L, INPP5D, NDUFAF7, NRXN1, ZNF638
3	OSBPL10
4	COL25A1, RAPGEF2, FRG1-DT, FAM114A1, SLC4A4
5	TMED7, MEGF10, PSD2, NRG2, SPRY4-AS1, GRIA1, PRLR, MEF2C-AS1
6	AFG1L, CCDC162P, PRKN
7	MOGAT3, BMT2, ZYX, EPHA1, EPHA1-AS1, DPP6, PMS2P1
8	SNX31, PTK2B, EPHX2, CLU, CSMD1, XKR9, NDUFAF6
9	ABCA1
10	TCERG1L, PIP4K2A, IPMK, TLL2
11	SNX19, NAV2, MADD, CELF1, NDUFS3, MS4A4A, MS4A6E, MARK2, SHANK2, GUCY2EP, PICALM, AP2A2, CNTN5
12	OAS1, CACNA1C, POC1B
13	MCF2L, RASA3
14	EML5, SLC24A4
15	SPPL2A, EFL1, ZNF710, RGMA
16	CLEC16A, DNAH3, RAB11FIP3, FTO, MTSS2
17	ITGB3, ABI3, ZNF652, MINK1, SCIMP, SDCBP2, TSPOAP1, GAS7
18	L3MBTL4, CCDC102B
19	CNN2, ABCA7, SSBP4, ACP7, CEACAM22P, PVR, TRAPPC6A, EXOC3L2, OPA3, QPCTL, FBXO46, CD33
20	SLC9A8, CASS4

Table 18: Genes identified in PRS analyses

The table lists the genes, categorised by chromosome (Chr) with variants present in both LOAD and sEOAD imputed models.

7.5 Limitations

As time and access became a limiting factor during the end stages of analysis, there were many aspects of research which were available but not able to be considered in the results. Some factors may have affected modelling post-PRS, as well as additional features to compare PRS results to. Examples of modelling factors included principal component analysis of genetic diversity and effects from which batch the sample was collected in and centre from which samples were recruited. As some analyses involve living samples, confirmation of clinical diagnoses can only be achieved with postmortem examination; whilst all analyses were completed with the most accurate data, incorrect diagnoses will impact the sensitivity and specificity of modelling.

Additional clinical information was available with the BDR resource, measuring lifestyle and environmental factors. As these factors affect risk of developing AD, they would provide beneficial insight towards understanding disease classification of AD cases with low PRS and controls with high PRS. Researching the clinical information with greatest utility and setting standards in practice when recruiting participants will improve the ability to conduct these analyses.

7.6 Future work

The continuation of this work would entail genotyping new samples on the NeuroBooster array; PRS analyses to predict likelihood of other dementia sub-types and samples of other ethnicity groups; and utilising artificial intelligence (AI) and machine learning to improve PRS methods.

7.6.1 BDR resource

The BDR cohort is utilised by many dementia research groups, benefitting from the availability of clinical features, blood and brain tissue and samples from several

dementia sub-types. Access to genetic data when conducting research can support candidate selection and reporting association.

7.6.2 NeuroBooster array

The third iteration of neuro-specific arrays is the NeuroBooster array; the array hosts SNPs associated with neurological disorders discovered since the NeuroChip was developed. Newly recruited BDR samples are now genotyped on this array, where imputation can support backwards compatibility with the NeuroX and NeuroChip genotyped samples.

7.6.3 Targets for analysis

During this study, there were insufficient sample numbers for PRS analysis of cohorts genotyped on the BDR with non-AD dementia sub-type classification, such as FTD and DLB. Increased recruitment of participants with these forms of dementia and large-scale GWAS of these dementia sub-types can provide sufficient power for analysis. From an equality, diversity and inclusion perspective, the historic non-participation of individuals of non-European ancestry in research has led to limited dementia studies of these populations; this has been addressed in other countries through actively engaging these communities and this approach needs to be taken for research in the UK to reduce the potential of bias existing within genetic research.

7.6.4 Machine learning and AI

Machine learning has already been utilised in imputation methodologies, furthering the use of this technology to select SNPs for PRS analysis can improve accuracy of PRS models. The significant advancements of generative AI in recent years holds potential utility for PRS analysis. Developing this technology for extensive variant discovery and analysis, which is currently unachievable due to limited time and resources to conduct

research, can accelerate the scale of genetic research to the same effect as seen with genotyping and sequencing technology advancements.

7.7 Conclusion

This PhD has improved the understanding and utility of analysing PRS in AD, comparing early and late-onset and as a toll for prediction. The published articles during this period have supported development of the research area, cited in various other papers and providing effective methods for analysis. The outputs delivered from genotyping have enhanced the current and future research into these cohorts by other dementia research groups.

The continuation of this research is necessary for driving our understanding of the role genetic variation plays in AD risk and pathology. The potential that recent advancements in technology provide create opportunities for conducting more comprehensive analysis over the next few years. Therefore, we can be optimistic about the improvements in diagnosing and treating dementia and ultimately quality of life of sufferers.

References

Lewis, F., Karlsberg Schaffer, S., Sussex, J., Cockcroft, L., & Sussex jsussex, J. (2014). The Trajectory of Dementia in the UK-Making a Difference. In *ARUK*.

Wittenberg, R., Hu, B., Barraza-Araiza, L., & Funder, A. R. (2019). *Projections of older people living with dementia and costs of dementia care in the CPEC Working Paper 5 The projections were produced using an updated version of a model developed by CPEC at LSE for the Modelling Outcome and Cost Impacts of Interventions for Dementia (MODEM) study (www.modem-dementia.org.uk). DISCLAIMER*. www.modem-dementia.org.uk

Public Health England. (2021). *Statistical commentary: dementia profile, March 2021 update - GOV.UK*. https://www.gov.uk/government/statistics/dementia-profile-updates/statistical-commentary-dementia-profile-march-2021-update

Burns, A., & Illife, S. (2009). Alzheimer's disease. *BMJ*, *338*, 467–471. https://doi.org/10.1136/BMJ.B1349

Budson, A. E. (2016). *Memory loss, alzheimer's disease, and dementia : a practical guide for clinicians* (P. R. Solomon & R. Britton, Eds.; Second edition.) [Book]. Elsevier.

Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., Fiske, A., & Pedersen, N. L. (2006). Role of genes and environments for explaining Alzheimer disease. *Archives of General Psychiatry*, *63*(2), 168–174. https://doi.org/10.1001/ARCHPSYC.63.2.168

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. *Annals of Neurology*, *67*(1), 114. https://doi.org/10.1002/ANA.21915

Prince, M., Wimo, A., Guerchet, M., Ali, G.-C., Wu, Y.-T., & Prina, M. (2015). *World Alzheimer Report 2015: The Global Impact of Dementia - An Analysis of Prevalence, Incidence, Cost and Trends*. www.alz.co.uk/worldreport2015corrections

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, *34*(7), 939–944. https://doi.org/10.1212/WNL.34.7.939

Goedert, M., & Spillantini, M. G. (2006). A century of Alzheimer's disease. *Science (New York, N.Y.)*, *314*(5800), 777–781. https://doi.org/10.1126/SCIENCE.1132814

Li, W. W., Wang, Z., Fan, D. Y., Shen, Y. Y., Chen, D. W., Li, H. Y., Li, L., Yang, H., Liu, Y. H., Bu, X. Le, Jin, W. S., Zeng, F., Xu, Z. Q., Yu, J. T., Chen, L. Y., &

Wang, Y. J. (2020). Association of Polygenic Risk Score with Age at Onset and Cerebrospinal Fluid Biomarkers of Alzheimer's Disease in a Chinese Cohort. *Neuroscience Bulletin*, *36*(7), 696–704. https://doi.org/10.1007/s12264-020-00469-8

Jorm, A. F., & Jolley, D. (1998). The incidence of dementia: a meta-analysis. *Neurology*, *51*(3), 728–733. https://doi.org/10.1212/WNL.51.3.728

Seshadri, S., & Wolf, P. A. (2007). Lifetime risk of stroke and dementia: current concepts, and estimates from the Framingham Study. *The Lancet. Neurology*, *6*(12), 1106–1114. https://doi.org/10.1016/S1474-4422(07)70291-0

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. *Archives of Neurology*, *56*(3), 303–308. https://doi.org/10.1001/ARCHNEUR.56.3.303

Zanetti, O., Solerte, S. B., & Cantoni, F. (2009). Life expectancy in Alzheimer's disease (AD). *Archives of Gerontology and Geriatrics*, *49 Suppl 1*, 237–243. https://doi.org/10.1016/J.ARCHGER.2009.09.035

Rodríguez-Rodríguez, E., Sánchez-Juan, P., Vázquez-Higuera, J. L., Mateo, I., Pozueta, A., Berciano, J., Cervantes, S., Alcolea, D., Martínez-Lage, P., Clarimón, J., Lleó, A., Pastor, P., & Combarros, O. (2013). Genetic risk score predicting accelerated progression from mild cognitive impairment to Alzheimer's disease. *Journal of Neural Transmission*, *120*(5), 807–812. https://doi.org/10.1007/s00702-012-0920-x

Adams, H. H. H., De Bruijn, R. F. A. G., Hofman, A., Uitterlinden, A. G., Van Duijn, C. M., Vernooij, M. W., Koudstaal, P. J., & Ikram, M. A. (2015). Genetic risk of neurodegenerative diseases is associated with mild cognitive impairment and conversion to dementia. *Alzheimer's and Dementia*, *11*(11), 1277–1285. https://doi.org/10.1016/j.jalz.2014.12.008

Pyun, J. M., Park, Y. H., Lee, K. J., Kim, S. Y., Saykin, A. J., & Nho, K. (2021). Predictability of polygenic risk score for progression to dementia and its interaction with APOE ε4 in mild cognitive impairment. *Translational Neurodegeneration*, *10*(1). https://doi.org/10.1186/s40035-021-00259-w

Hansson, O., Zetterberg, H., Vanmechelen, E., Vanderstichele, H., Andreasson, U., Londos, E., Wallin, A., Minthon, L., & Blennow, K. (2010). Evaluation of plasma Aβ40 and Aβ42 as predictors of conversion to Alzheimer's disease in patients with mild cognitive impairment. *Neurobiology of Aging*, *31*(3), 357–367. https://doi.org/10.1016/J.NEUROBIOLAGING.2008.03.027

Li, X. L., Hu, N., Tan, M. S., Yu, J. T., & Tan, L. (2014). Behavioral and Psychological Symptoms in Alzheimer's Disease. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/927804

Förstl, H., & Kurz, A. (1999). Clinical features of Alzheimer's disease. *European Archives of Psychiatry and Clinical Neuroscience*, *249*(6), 288–290. https://doi.org/10.1007/S004060050101

Whitwell, J. L. (2010). Progression of atrophy in Alzheimer's disease and related disorders. *Neurotoxicity Research*, *18*(3–4), 339–346. https://doi.org/10.1007/S12640-010-9175-1

Wenk, G. L. (2003). Neuropathologic Changes in Alzheimer's Disease. *The Journal of Clinical Psychiatry*, *64*(suppl 9), 12701. https://www.psychiatrist.com/jcp/neurologic/dementia/neuropathologic-changes-alzheimers-disease

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 : A Journal of Neurology*, *130*(Pt 7), 1777–1786. https://doi.org/10.1093/BRAIN/AWM112

Thal, R. D., Ghebremedhin, E., Rub, U., Yamaguchi, H., Del Tredici, K., & Braak, H. (2002). Two Types of Sporadic Cerebral Amyloid Angiopathy. In *Journal of Neuropathology and Experimental Neurology* (Vol. 61, Issue 3). https://academic.oup.com/jnen/article-abstract/61/3/282/2610063

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. *Molecular Neurobiology*, *38*(1), 1–15. https://doi.org/10.1007/S12035-008-8019-Y/TABLES/3

Pozueta, J., Lefort, R., & Shelanski, M. L. (2013). Synaptic changes in Alzheimer's disease and its models. *Neuroscience*, *251*, 51–65. https://doi.org/10.1016/J.NEUROSCIENCE.2012.05.050

Postina, R. (2012). Activation of α-secretase cleavage. *Journal of Neurochemistry*, *120 Suppl 1*(SUPPL. 1), 46–54. https://doi.org/10.1111/J.1471-4159.2011.07459.X

Lichtenthaler, S. F. (2012). Alpha-secretase cleavage of the amyloid precursor protein: proteolysis regulated by signaling pathways and protein trafficking. *Current Alzheimer Research*, *9*(2), 165–177. https://doi.org/10.2174/156720512799361655

Hampel, H., Hardy, J., Blennow, K., Chen, C., Perry, G., Kim, S. H., Villemagne, V. L., Aisen, P., Vendruscolo, M., Iwatsubo, T., Masters, C. L., Cho, M., Lannfelt, L., Cummings, J. L., & Vergallo, A. (2021). The Amyloid-β Pathway in Alzheimer's Disease. *Molecular Psychiatry 2021 26:10*, *26*(10), 5481–5503. https://doi.org/10.1038/s41380-021-01249-0

Bird, T. D. (2018). Alzheimer Disease Overview. *GeneReviews®*. https://www.ncbi.nlm.nih.gov/books/NBK1161/

Marsden, I. T., Laurie, L. S., & Bamburg, J. R. (2011). Amyloid-β-induced amyloid-β secretion: A possible feed-forward mechanism in Alzheimer

disease. *Journal of Alzheimer's Disease : JAD*, *24*(4), 681. https://doi.org/10.3233/JAD-2011-101899

Jan, A., Gokce, O., Luthi-Carter, R., & Lashuel, H. A. (2008). The ratio of monomeric to aggregated forms of Abeta40 and Abeta42 is an important determinant of amyloid-beta aggregation, fibrillogenesis, and toxicity. *The Journal of Biological Chemistry*, *283*(42), 28176–28189. https://doi.org/10.1074/JBC.M803159200

Gu, L., & Guo, Z. (2013). Alzheimer's Aβ42 and Aβ40 peptides form interlaced amyloid fibrils. *Journal of Neurochemistry*, *126*(3), 305. https://doi.org/10.1111/JNC.12202

Klein, W. L. (2013). Synaptotoxic amyloid-β oligomers: a molecular basis for the cause, diagnosis, and treatment of Alzheimer's disease? *Journal of Alzheimer's Disease : JAD*, *33 Suppl 1*(SUPPL. 1). https://doi.org/10.3233/JAD-2012-129039

Hardy, J., & Allsop, D. (1991). Amyloid deposition as the central event in the aetiology of Alzheimer's disease. *Trends in Pharmacological Sciences*, *12*(10), 383–388. https://doi.org/10.1016/0165-6147(91)90609-V

Karran, E., Mercken, M., & Strooper, B. De. (2011). The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nature Reviews. Drug Discovery*, *10*(9), 698–712. https://doi.org/10.1038/NRD3505

Spillantini, M. G., & Goedert, M. (2013). Tau pathology and neurodegeneration. *The Lancet. Neurology*, *12*(6), 609–622. https://doi.org/10.1016/S1474-4422(13)70090-5

Ballatore, C., Lee, V. M. Y., & Trojanowski, J. Q. (2007). Tau-mediated neurodegeneration in Alzheimer's disease and related disorders. *Nature Reviews. Neuroscience*, *8*(9), 663–672. https://doi.org/10.1038/nrn2194

Churcher, I. (2006). Tau therapeutic strategies for the treatment of Alzheimer's disease. *Current Topics in Medicinal Chemistry*, *6*(6), 579–595. https://doi.org/10.2174/156802606776743057

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 Structure and Function*, *21*(5), 283–295. https://doi.org/10.1247/CSF.21.283

Goedert, M., Jakes, R., & Vanmechelen, E. (1995). Monoclonal antibody AT8 recognises tau protein phosphorylated at both serine 202 and threonine 205. *Neuroscience Letters*, *189*(3), 167–170. https://doi.org/10.1016/0304-3940(95)11484-E

Roy, S., Zhang, B., Lee, V. M. Y., & Trojanowski, J. Q. (2005). Axonal transport defects: a common theme in neurodegenerative diseases. *Acta Neuropathologica*, *109*(1), 5–13. https://doi.org/10.1007/S00401-004-0952-X

Nussbaum, J. M., Seward, M. E., & Bloom, G. S. (2013). Alzheimer disease: a tale of two prions. *Prion*, *7*(1), 14–19. https://doi.org/10.4161/PRI.22118

Neumann, M., Tolnay, M., & Mackenzie, I. R. A. (2009). The molecular basis of frontotemporal dementia. *Expert Reviews in Molecular Medicine*, *11*. https://doi.org/10.1017/S1462399409001136

Hutton, M., Lendon, C. L., Rizzu, P., Baker, M., Froelich, S., Houlden, H. H., Pickering-Brown, S., Chakraverty, S., Isaacs, A., Grover, A., Hackett, J., Adamson, J., Lincoln, S., Dickson, D., Davies, P., Petersen, R. C., Stevena, M., De Graaff, E., Wauters, E., … Heutink, P. (1998). Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature*, *393*(6686), 702–704. https://doi.org/10.1038/31508

Spires-Jones, T. L., & Hyman, B. T. (2014). The intersection of amyloid beta and tau at synapses in Alzheimer's disease. *Neuron*, *82*(4), 756–771. https://doi.org/10.1016/J.NEURON.2014.05.004

Braak, H., & Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, *82*(4), 239–259. https://doi.org/10.1007/BF00308809

Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H., & Tredici, K. (2006). Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathologica*, *112*(4), 389–404. https://doi.org/10.1007/S00401-006-0127-Z

Knopman, D. S., Parisi, J. E., Salviati, A., Floriach-Robert, M., Boeve, B. F., Ivnik, R. J., Smith, G. E., Dickson, D. W., Johnson, K. A., Petersen, L. E., McDonald, W. C., Braak, H., & Petersen, R. C. (2003). Neuropathology of cognitively normal elderly. *Journal of Neuropathology and Experimental Neurology*, *62*(11), 1087–1095. https://doi.org/10.1093/JNEN/62.11.1087

Lipton, S. A. (2006). Paradigm shift in neuroprotection by NMDA receptor blockade: memantine and beyond. *Nature Reviews. Drug Discovery*, *5*(2), 160–170. https://doi.org/10.1038/NRD1958

Francis, P. T., Palmer, A. M., Snape, M., & Wilcock, G. K. (1999). The cholinergic hypothesis of Alzheimer's disease: a review of progress. *Journal of Neurology, Neurosurgery, and Psychiatry*, *66*(2), 137–147. https://doi.org/10.1136/JNNP.66.2.137

Cummings, J. L., Morstorf, T., & Zhong, K. (2014). Alzheimer's disease drug-development pipeline: Few candidates, frequent failures. *Alzheimer's Research and Therapy*, *6*(4), 1–7. https://doi.org/10.1186/ALZRT269/TABLES/3

Doody, R. S., Thomas, R. G., Farlow, M., Iwatsubo, T., Vellas, B., Joffe, S., Kieburtz, K., Raman, R., Sun, X., Aisen, P. S., Siemers, E., Liu-Seifert, H., & Mohs, R. (2014). Phase 3 Trials of Solanezumab for Mild-to-Moderate Alzheimer's Disease. *New England Journal of Medicine*, *370*(4), 311–321.

https://doi.org/10.1056/NEJMOA1312889/SUPPL_FILE/NEJMOA1312889_DISCLOSURES.PDF

Novak, P., Schmidt, R., Kontsekova, E., Zilka, N., Kovacech, B., Skrabana, R., Vince-Kazmerova, Z., Katina, S., Fialova, L., Prcina, M., Parrak, V., Dal-Bianco, P., Brunner, M., Staffen, W., Rainer, M., Ondrus, M., Ropele, S., Smisek, M., Sivak, R., … Novak, M. (2017). Safety and immunogenicity of the tau vaccine AADvac1 in patients with Alzheimer's disease: a randomised, double-blind, placebo-controlled, phase 1 trial. *The Lancet. Neurology*, *16*(2), 123–134. https://doi.org/10.1016/S1474-4422(16)30331-3

National Institute for Health and Clinical Excellence. (2011). *Donepezil, galantamine, rivastigmine and memantine for the treatment of Alzheimer's disease Your responsibility*. www.nice.org.uk/guidance/ta217

Wilkinson, D. (2012). A review of the effects of memantine on clinical progression in Alzheimer's disease. *International Journal of Geriatric Psychiatry*, *27*(8), 769–776. https://doi.org/10.1002/GPS.2788

Ramanan, V. K., & Day, G. S. (2023). Anti-amyloid therapies for Alzheimer disease: finally, good news for patients. *Molecular Neurodegeneration*, *18*(1), 1–3. https://doi.org/10.1186/S13024-023-00637-0/FIGURES/1

Alzheimer's Society. (2023). *Three promising drugs for treating Alzheimer's disease bring fresh hope | Alzheimer's Society*. https://www.alzheimers.org.uk/blog/three-promising-drugs-for-treating-alzheimers-disease-bring-fresh-hope

Blanc, F., Philippi, N., Cretin, B., Kleitz, C., Berly, L., Jung, B., Kremer, S., Namer, I. J., Sellal, F., Jaulhac, B., & De Seze, J. (2014). Lyme neuroborreliosis and dementia. *Journal of Alzheimer's Disease : JAD*, *41*(4), 1087–1093. https://doi.org/10.3233/JAD-130446

Alzheimer's Society. (2021). *HIV-associated neurocognitive disorder (HAND)*. https://www.alzheimers.org.uk/about-dementia/types-dementia/hiv-cognitive-impairment

Abbott, A. (2020). Are infections seeding some cases of Alzheimer's disease? *Nature*, *587*(7832), 22–25. https://doi.org/10.1038/D41586-020-03084-9

Seaksid, C. E., & Wilcockid, D. M. (2020). *Infectious hypothesis of Alzheimer disease*. https://doi.org/10.1371/journal.ppat.1008596

Itzhaki, R. F., Lin, W. R., Shang, D., Wilcock, G. K., Faragher, B., & Jamieson, G. A. (1997). Herpes simplex virus type 1 in brain and risk of Alzheimer's disease. *Lancet (London, England)*, *349*(9047), 241–244. https://doi.org/10.1016/S0140-6736(96)10149-5

Chen, C. K., Wu, Y. T., & Chang, Y. C. (2017). Association between chronic periodontitis and the risk of Alzheimer's disease: a retrospective, population-based, matched-cohort study. *Alzheimer's Research & Therapy*, *9*(1). https://doi.org/10.1186/S13195-017-0282-6

Skrobot, O. A., Black, S. E., Chen, C., DeCarli, C., Erkinjuntti, T., Ford, G. A., Kalaria, R. N., O'Brien, J., Pantoni, L., Pasquier, F., Roman, G. C., Wallin, A., Sachdev, P., Skoog, I., Taragano, F. E., Kril, J., Cavalieri, M., Jellinger, K. A., Kovacs, G. G., … Kehoe, P. G. (2018). Progress toward standardized diagnosis of vascular cognitive impairment: Guidelines from the Vascular Impairment of Cognition Classification Consensus Study. *Alzheimer's & Dementia*, *14*(3), 280–292. https://doi.org/10.1016/J.JALZ.2017.09.007

Hofman, A., Ott, A., Breteler, M. M. B., Bots, M. L., Slooter, A. J. C., Van Harskamp, F., Van Duijn, C. N., Van Broeckhoven, C., & Grobbee, D. E. (1997). Atherosclerosis, apolipoprotein E, and prevalence of dementia and Alzheimer's disease in the Rotterdam Study. *Lancet (London, England)*, *349*(9046), 151–154. https://doi.org/10.1016/S0140-6736(96)09328-2

Andrews, S. J., Fulton-Howard, B., O'Reilly, P., Marcora, E., Goate, A. M., Farrer, L. A., Haines, J. L., Mayeux, R., Naj, A. C., Pericak-Vance, M. A., Schellenberg, G. D., & Wang, L. S. (2021). Causal Associations Between Modifiable Risk Factors and the Alzheimer's Phenome. *Annals of Neurology*, *89*(1), 54–65. https://doi.org/10.1002/ana.25918

Chandler, H. L., Wise, R. G., Murphy, K., Tansey, K. E., Linden, D. E. J., & Lancaster, T. M. (2019). Polygenic impact of common genetic risk loci for Alzheimer's disease on cerebral blood flow in young individuals. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-018-36820-3

Kivipelto, M., & Solomon, A. (2006). Cholesterol as a risk factor for Alzheimer's disease - epidemiological evidence. *Acta Neurologica Scandinavica. Supplementum*, *185*(SUPPL. 185), 50–57. https://doi.org/10.1111/J.1600-0404.2006.00685.X

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. *Neurobiology of Aging*, *21*(1), 49–55. https://doi.org/10.1016/S0197-4580(00)00096-8

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*(18), 1568–1574. https://doi.org/10.1212/WNL.0B013E3182190D09

Panza, F., Frisardi, V., Capurso, C., D'Introno, A., Colacicco, A. M., Imbimbo, B. P., Santamato, A., Vendemiale, G., Seripa, D., Pilotto, A., Capurso, A., & Solfrizzi, V. (2010). Late-Life Depression, Mild Cognitive Impairment, and Dementia: Possible Continuum? *The American Journal of Geriatric Psychiatry*, *18*(2), 98–116. https://doi.org/10.1097/JGP.0B013E3181B0FA13

Kokmen, E., Beard, C. M., O'Brien, P. C., & Kurland, L. T. (1996). Epidemiology of dementia in Rochester, Minnesota. *Mayo Clinic Proceedings*, *71*(3), 275–282. https://doi.org/10.4065/71.3.275

Matloff, W. J., Zhao, L., Ning, K., Conti, D. V., & Toga, A. W. (2020). Interaction effect of alcohol consumption and Alzheimer disease polygenic risk score on the brain cortical thickness of cognitively normal subjects. *Alcohol*, *85*, 1–12. https://doi.org/10.1016/j.alcohol.2019.11.002

Dosunmu, R., Wu, J., Basha, M. R., & Zawia, N. H. (2007). Environmental and dietary risk factors in Alzheimer's disease. *Expert Review of Neurotherapeutics*, *7*(7), 887–900. https://doi.org/10.1586/14737175.7.7.887

Andrews, S. J., Fulton-Howard, B., Patterson, C., McFall, G. P., Gross, A., Michaelis, E. K., Goate, A., Swerdlow, R. H., & Pa, J. (2020). Mitonuclear interactions influence Alzheimer's disease risk. *Neurobiology of Aging*, *87*, 138.e7-138.e14. https://doi.org/10.1016/j.neurobiolaging.2019.09.007

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*(9), 1707–1712. https://doi.org/10.1212/WNL.45.9.1707

Zhou, D. F., Wu, C. S., Qi, H., Fan, J. H., Sun, X. D., Como, P., Qiao, Y. L., Zhang, L., & Kieburtz, K. (2006). Prevalence of dementia in rural China: impact of age, gender and education. *Acta Neurologica Scandinavica*, *114*(4), 273–280. https://doi.org/10.1111/J.1600-0404.2006.00641.X

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. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *60*(5). https://doi.org/10.1093/GERONB/60.5.P251

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's Disease Centers, 2005–2010. *Journal of Neuropathology and Experimental Neurology*, *71*(4), 266. https://doi.org/10.1097/NEN.0B013E31824B211B

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Feldman, H. H., Frisoni, G. B., Hampel, H., Jagust, W. J., Johnson, K. A., Knopman, D. S., Petersen, R. C., Scheltens, P., Sperling, R. A., & Dubois, B. (2016). *VIEWS & REVIEWS A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers*.

Dubois, B., Feldman, H. H., Jacova, C., DeKosky, S. T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., Meguro, K., O'Brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Stern, Y., Visser, P. J., & Scheltens, P. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *The Lancet Neurology*, *6*(8), 734–746. https://doi.org/10.1016/S1474-4422(07)70178-3

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., Mohs, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillo, M. C., Thies, B., Weintraub, S., & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, *7*(3), 263–269. https://doi.org/10.1016/J.JALZ.2011.03.005

Blacker, D., Albert, M. S., Bassett, S. S., Rodney, C. P., Harrell, L. E., & Folstein, M. F. (1994). Reliability and validity of NINCDS-ADRDA criteria for Alzheimer's disease. The National Institute of Mental Health Genetics Initiative. *Archives of Neurology*, *51*(12), 1198–1204. https://doi.org/10.1001/ARCHNEUR.1994.00540240042014

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. *Journal of Psychiatric Research*, *12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Pangman, V. C., Sloan, J., & Guse, L. (2000). An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: implications for clinical practice. *Applied Nursing Research : ANR*, *13*(4), 209–213. https://doi.org/10.1053/APNR.2000.9231

Mungas, D. (1991). In-office mental status testing: a practical guide. *Geriatrics*.

Yaari, R., & Corey-Bloom, J. (2007). Alzheimer's disease. *Seminars in Neurology*, *27*(1), 32–41. https://doi.org/10.1055/S-2006-956753

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. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *14*(3), 311–320. https://doi.org/10.1176/JNP.14.3.311

Johnson, K. A., Fox, N. C., Sperling, R. A., & Klunk, W. E. (2012). Brain Imaging in Alzheimer Disease. *Cold Spring Harbor Perspectives in Medicine*, *2*(4). https://doi.org/10.1101/CSHPERSPECT.A006213

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. Oxford Project to Investigate Memory and Aging. *International Psychogeriatrics*, *10*(3), 271–302. https://doi.org/10.1017/S1041610298005389

Minati, L., Edginton, T., Grazia Bruzzone, M., & Giaccone, G. (2009). Current concepts in Alzheimer's disease: a multidisciplinary review. *American*

Journal of Alzheimer's Disease and Other Dementias, *24*(2), 95–121. https://doi.org/10.1177/1533317508328602

Vlassenko, A. G., Benzinger, T. L. S., & Morris, J. C. (2012). PET Amyloid-Beta Imaging in Preclinical Alzheimer's Disease. *Biochimica et Biophysica Acta*, *1822*(3), 370. https://doi.org/10.1016/J.BBADIS.2011.11.005

Mattsson, N. (2011). CSF biomarkers in neurodegenerative diseases. *Clinical Chemistry and Laboratory Medicine*, *49*(3), 345–352. https://doi.org/10.1515/CCLM.2011.082

Ewers, M., Mattsson, N., Minthon, L., Molinuevo, J. L., Antonell, A., Popp, J., Jessen, F., Herukka, S. K., Soininen, H., Maetzler, W., Leyhe, T., Bürger, K., Taniguchi, M., Urakami, K., Lista, S., Dubois, B., Blennow, K., & Hampel, H. (2015). CSF biomarkers for the differential diagnosis of Alzheimer's disease: A large-scale international multicenter study. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, *11*(11), 1306–1315. https://doi.org/10.1016/J.JALZ.2014.12.006

Humpel, C. (2011). Identifying and validating biomarkers for Alzheimer's disease. *Trends in Biotechnology*, *29*(1), 26–32. https://doi.org/10.1016/J.TIBTECH.2010.09.007

Scheltens, P., Fox, N., Barkhof, F., & De Carli, C. (2002). Structural magnetic resonance imaging in the practical assessment of dementia: Beyond exclusion. *Lancet Neurology*, *1*(1), 13–21. https://doi.org/10.1016/S1474-4422(02)00002-9

Nobili, F., & Morbelli, S. (2010). [18F]FDG-PET as a Biomarker for Early Alzheimer's Disease. *The Open Nuclear Medicine Journal*, *2*, 46–52.

Mistur, R., Mosconi, L., de Santi, S., 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. *Journal of Clinical Neurology (Seoul, Korea)*, *5*(4), 153. https://doi.org/10.3988/JCN.2009.5.4.153

Herholz, K., Westwood, S., Haense, C., & Dunn, G. (2011). Evaluation of a calibrated (18)F-FDG PET score as a biomarker for progression in Alzheimer disease and mild cognitive impairment. *Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine*, *52*(8), 1218–1226. https://doi.org/10.2967/JNUMED.111.090902

Blennow, K., & Hampel, H. (2003). CSF markers for incipient Alzheimer's disease. *Lancet Neurology*, *2*(10), 605–613. https://doi.org/10.1016/S1474-4422(03)00530-1

Llorens, F., Schmitz, M., Ferrer, I., & Zerr, I. (2016). CSF biomarkers in neurodegenerative and vascular dementias. *Progress in Neurobiology*, *138–140*, 36–53. https://doi.org/10.1016/J.PNEUROBIO.2016.03.003

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 & Dementia (Amsterdam, Netherlands), *1*(4), 455–463. https://doi.org/10.1016/J.DADM.2015.09.003

Motter, R., Vigo-Pelfrey, C., Kholodenko, D., Barbour, R., Johnson-Wood, K., Galasko, D., Chang, L., Miller, B., Clark, C., Green, R., Olson, D., Southwick, P., Wolfert, R., Munroe, B., Lieberburg, I., Seubert, P., & Schenk, D. (1995). Reduction of beta-amyloid peptide42 in the cerebrospinal fluid of patients with Alzheimer's disease. *Annals of Neurology*, *38*(4), 643–648. https://doi.org/10.1002/ANA.410380413

Suárez-Calvet, M., Kleinberger, G., Araque Caballero, M. Á., Brendel, M., Rominger, A., Alcolea, D., Fortea, J., Lleó, A., Blesa, R., Gispert, J. D., Sánchez-Valle, R., Antonell, A., Rami, L., Molinuevo, J. L., Brosseron, F., Traschütz, A., Heneka, M. T., Struyfs, H., Engelborghs, S., … Haass, C. (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 Molecular Medicine*, *8*(5), 466–476. https://doi.org/10.15252/EMMM.201506123

Mattsson, N., Zetterberg, H., Hansson, O., Andreasen, N., Parnetti, L., Jonsson, M., Herukka, S. K., Van Der Flier, W. M., Blankenstein, M. A., Ewers, M., Rich, K., Kaiser, E., Verbeek, M., Tsolaki, M., Mulugeta, E., Rosén, E., Aarsland, D., Jelle Visser, P., Schröder, J., … Blennow, K. (2009). CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. *JAMA*, *302*(4), 385–393. https://doi.org/10.1001/JAMA.2009.1064

van Oijen, M., Hofman, A., Soares, H. D., Koudstaal, P. J., & Breteler, M. M. (2006). Plasma Abeta(1-40) and Abeta(1-42) and the risk of dementia: a prospective case-cohort study. *The Lancet. Neurology*, *5*(8), 655–660. https://doi.org/10.1016/S1474-4422(06)70501-4

Mehta, P. D., Pirttilä, T., Mehta, S. P., Sersen, E. A., Aisen, P. S., & Wisniewski, H. M. (2000). Plasma and cerebrospinal fluid levels of amyloid beta proteins 1-40 and 1-42 in Alzheimer disease. *Archives of Neurology*, *57*(1), 100–105. https://doi.org/10.1001/ARCHNEUR.57.1.100

Thambisetty, M., & Lovestone, S. (2010). Blood-based biomarkers of Alzheimer's disease: challenging but feasible. *Biomarkers in Medicine*, *4*(1), 65. https://doi.org/10.2217/BMM.09.84

Ray, S., Britschgi, M., Herbert, C., Takeda-Uchimura, Y., Boxer, A., Blennow, K., Friedman, L. F., Galasko, D. R., Jutel, M., Karydas, A., Kaye, J. A., Leszek, J., Miller, B. L., Minthon, L., Quinn, J. F., Rabinovici, G. D., Robinson, W. H., Sabbagh, M. N., So, Y. T., … Wyss-Coray, T. (2007). Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. *Nature Medicine*, *13*(11), 1359–1362. https://doi.org/10.1038/NM1653

Giacomucci, G., Mazzeo, S., Bagnoli, S., Ingannato, A., Leccese, D., Berti, V., Padiglioni, S., Galdo, G., Ferrari, C., Sorbi, S., Bessi, V., & Nacmias, B.

(2022). Plasma neurofilament light chain as a biomarker of Alzheimer's disease in Subjective Cognitive Decline and Mild Cognitive Impairment. *Journal of Neurology*, *269*(8), 4270. https://doi.org/10.1007/S00415-022-11055-5

Piaceri, I., Nacmias, B., & Sorbi, S. (2013). Genetics of familial and sporadic Alzheimer's disease. *Frontiers in Bioscience (Elite Edition)*, *5*(1), 167–177. https://doi.org/10.2741/E605

Wilson, R. S., Barral, S., Lee, J. H., Leurgans, S. E., Foroud, T. M., Sweet, R. A., Graff-Radford, N., Bird, T. D., Mayeux, R., & Bennett, D. A. (2011). Heritability of different forms of memory in the Late Onset Alzheimer's Disease Family Study. *Journal of Alzheimer's Disease : JAD*, *23*(2), 249–255. https://doi.org/10.3233/JAD-2010-101515

Mendez, M. F. (2012). Early-onset Alzheimer's disease: nonamnestic subtypes and type 2 AD. *Archives of Medical Research*, *43*(8), 677–685. https://doi.org/10.1016/J.ARCMED.2012.11.009

Antonell, A., Lladó, A., Altirriba, J., Botta-Orfila, T., Balasa, M., Fernández, M., Ferrer, I., Sánchez-Valle, R., & Molinuevo, J. L. (2013). A preliminary study of the whole-genome expression profile of sporadic and monogenic early-onset Alzheimer's disease. *Neurobiology of Aging*, *34*(7), 1772–1778. https://doi.org/10.1016/J.NEUROBIOLAGING.2012.12.026

Ertekin-Taner, N. (2007). Genetics of Alzheimer's disease: a centennial review. *Neurologic Clinics*, *25*(3), 611–667. https://doi.org/10.1016/J.NCL.2007.03.009

Bertram, L., Lill, C. M., & Tanzi, R. E. (2010). The genetics of Alzheimer disease: back to the future. *Neuron*, *68*(2), 270–281. https://doi.org/10.1016/J.NEURON.2010.10.013

Guerreiro, R., & Hardy, J. (2014). Genetics of Alzheimer's disease. *Neurotherapeutics : The Journal of the American Society for Experimental NeuroTherapeutics*, *11*(4), 732–737. https://doi.org/10.1007/S13311-014-0295-9

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., & Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(45), 19096–19101. https://doi.org/10.1073/PNAS.0910672106

Hampel, H., & Lista, S. (2012). From inherited to sporadic AD—crossing the biomarker bridge. *Nature Reviews Neurology 2012 8:11*, *8*(11), 598–600. https://doi.org/10.1038/nrneurol.2012.202

Tanzi, R. E., & Bertram, L. (2005). Twenty years of the Alzheimer's disease amyloid hypothesis: a genetic perspective. *Cell*, *120*(4), 545–555. https://doi.org/10.1016/J.CELL.2005.02.008

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *The New England Journal of Medicine*, *363*(2), 166–176. https://doi.org/10.1056/NEJMRA0905980

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology*, *8*(12), e1002822. https://doi.org/10.1371/JOURNAL.PCBI.1002822

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. *Nature Protocols*, *6*(2), 121–133. https://doi.org/10.1038/NPROT.2010.182

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, *5*(9), 1564–1573. https://doi.org/10.1038/NPROT.2010.116

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., Grenier-Boley, B., Russo, G., Thornton-Wells, T. A., Jones, N., Smith, A. V., Chouraki, V., Thomas, C., Ikram, M. A., Zelenika, D., … Seshadri, S. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics 2013 45:12*, *45*(12), 1452–1458. https://doi.org/10.1038/ng.2802

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. *American Journal of Human Genetics*, *84*(1), 35–43. https://doi.org/10.1016/J.AJHG.2008.12.008

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., & Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature Genetics*, *39*(1), 17–23. https://doi.org/10.1038/NG1934

Seshadri, S., Fitzpatrick, A. L., Ikram, M. A., DeStefano, A. L., Gudnason, V., Boada, M., Bis, J. C., Smith, A. V., Carassquillo, M. M., Lambert, J. C., Harold, D., Schrijvers, E. M. C., Ramirez-Lorca, R., Debette, S., Longstreth, W. T., Janssens, A. C. J. W., Pankratz, V. S., Dartigues, J. F., Hollingworth, P., … Breteler, M. M. B. (2010). Genome-wide analysis of genetic loci associated with Alzheimer disease. *JAMA*, *303*(18), 1832–1840. https://doi.org/10.1001/JAMA.2010.574

Coon, K. D., Myers, A. J., Craig, D. W., Webster, J. A., Pearson, J. V., Lince, D. H., Zismann, V. L., Beach, T. G., Leung, D., Bryden, L., Halperin, R. F., Marlowe, L., Kaleem, M., Walker, D. G., Ravid, R., Heward, C. B., Rogers, J., Papassotiropoulos, A., Reiman, E. M., … Stephan, D. A. (2007). A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *The Journal*

of Clinical Psychiatry, *68*(4), 613–618. https://doi.org/10.4088/JCP.V68N0419

Grupe, A., Abraham, R., Li, Y., Rowland, C., Hollingworth, P., Morgan, A., Jehu, L., Segurado, R., Stone, D., Schadt, E., Karnoub, M., Nowotny, P., Tacey, K., Catanese, J., Sninsky, J., Brayne, C., Rubinsztein, D., Gill, M., Lawlor, B., … Williams, J. (2007). Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. *Human Molecular Genetics*, *16*(8), 865–873. https://doi.org/10.1093/HMG/DDM031

Li, Y., Rowland, C., Catanese, J., Morris, J., Lovestone, S., O'Donovan, M. C., Goate, A., Owen, M., Williams, J., & Grupe, A. (2008). SORL1 variants and risk of late-onset Alzheimer's disease. *Neurobiology of Disease*, *29*(2), 293. https://doi.org/10.1016/J.NBD.2007.09.001

Reiman, E. M., Webster, J. A., Myers, A. J., Hardy, J., Dunckley, T., Zismann, V. L., Joshipura, K. D., Pearson, J. V., Hu-Lince, D., Huentelman, M. J. J., Craig, D. W., Coon, K. D., Liang, W. S., Herbert, R. L. H., Beach, T., Rohrer, K. C., Zhao, A. S., Leung, D., Bryden, L., … Stephan, D. A. (2007). GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. *Neuron*, *54*(5), 713–720. https://doi.org/10.1016/J.NEURON.2007.05.022

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M. L., Pahwa, J. S., Moskvina, V., Dowzell, K., Williams, A., Jones, N., Thomas, C., Stretton, A., Morgan, A. R., Lovestone, S., Powell, J., Proitsi, P., Lupton, M. K., Brayne, C., … Williams, J. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature Genetics*, *41*(10), 1088–1093. https://doi.org/10.1038/NG.440

Lambert, J. C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., Combarros, O., Zelenika, D., Bullido, M. J., Tavernier, B., Letenneur, L., Bettens, K., Berr, C., Pasquier, F., Fiévet, N., Barberger-Gateau, P., Engelborghs, S., De Deyn, P., Mateo, I., … Amouyel, P. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature Genetics*, *41*(10), 1094–1099. https://doi.org/10.1038/NG.439

Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J. C., Carrasquillo, M. M., Abraham, R., Hamshere, M. L., Pahwa, J. S., Moskvina, V., Dowzell, K., Jones, N., Stretton, A., Thomas, C., Richards, A., Ivanov, D., Widdowson, C., Chapman, J., Lovestone, S., … Williams, J. (2011). Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nature Genetics*, *43*(5), 429–436. https://doi.org/10.1038/NG.803

Bellenguez, C., Küçükali, F., Jansen, I. E., Kleineidam, L., Moreno-Grau, S., Amin, N., Naj, A. C., Campos-Martin, R., Grenier-Boley, B., Andrade, V., Holmans, P. A., Boland, A., Damotte, V., van der Lee, S. J., Costa, M. R., Kuulasmaa, T., Yang, Q., de Rojas, I., Bis, J. C., … Lambert, J. C. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias.

Nature Genetics 2022 54:4, *54*(4), 412–436. https://doi.org/10.1038/s41588-022-01024-z

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature 2009 461:7265*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L., & Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science (New York, N.Y.)*, *261*(5123), 921–923. https://doi.org/10.1126/SCIENCE.8346443

Liao, F., Yoon, H., & Kim, J. (2017). Apolipoprotein E metabolism and functions in brain and its role in Alzheimer's disease. *Current Opinion in Lipidology*, *28*(1), 60–67. https://doi.org/10.1097/MOL.0000000000000383

Farrer, L. A., Cupples, L. A., Haines, J. L., Hyman, B., Kukull, W. A., Mayeux, R., Myers, R. H., Pericak-Vance, M. A., Risch, N., & van Duijn, C. M. (1997). Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium - PubMed. *JAMA*, *278*(16), 1349–1356. https://pubmed.ncbi.nlm.nih.gov/9343467/

Corder, E. H., Saunders, A. M., Risch, N. J., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Rimmler, J. B., Locke, P. A., Conneally, P. M., Schmader, K. E., Small, G. W., Roses, A. D., Haines, J. L., & Pericak-Vance, M. A. (1994). Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nature Genetics*, *7*(2), 180–184. https://doi.org/10.1038/NG0694-180

Holtzman, D. M., Bales, K. R., Tenkova, T., Fagan, A. M., Parsadanian, M., Sartorius, L. J., Mackey, B., Olney, J., McKeel, D., Wozniak, D., & Paul, S. M. (2000). Apolipoprotein E isoform-dependent amyloid deposition and neuritic degeneration in a mouse model of Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(6), 2892–2897. https://doi.org/10.1073/PNAS.050004797

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 in Genetics : TIG*, *26*(2), 84–93. https://doi.org/10.1016/J.TIG.2009.12.004

Ware, E. B., Faul, J. D., Mitchell, C. M., & Bakulski, K. M. (2020). Considering the APOE locus in Alzheimer's disease polygenic scores in the Health and Retirement Study: a longitudinal panel study. *BMC Medical Genomics*, *13*(1), 1–13. https://doi.org/10.1186/S12920-020-00815-9/TABLES/2

Kulminski, A. M., Philipp, I., Shu, L., & Kulminskaya, I. (2021). The ε4-bearing TOMM40-APOE-APOC1 haplotype but not the ε4 allele confers an exceptionally high risk of Alzheimer's disease. *Alzheimer's & Dementia*, *17*, e050617. https://doi.org/10.1002/ALZ.050617

Bekris, L. M., Yu, C. E., Bird, T. D., & Tsuang, D. W. (2010). Genetics of Alzheimer Disease. *Journal of Geriatric Psychiatry and Neurology*, *23*(4), 213. https://doi.org/10.1177/0891988710383571

Taddei, K., Fisher, C., Laws, S. M., Martins, G., Paton, A., Clarnette, R. M., Chung, C., Brooks, W. S., Hallmayer, J., Miklossy, J., Relkin, N., St George-Hyslop, P. H., Gandy, S. E., & Martins, R. N. (2002). Association between presenilin-1 Glu318Gly mutation and familial Alzheimer's disease in the Australian population. *Molecular Psychiatry*, *7*(7), 776–781. https://doi.org/10.1038/SJ.MP.4001072

De Strooper, B., Saftig, P., Craessaerts, K., Vanderstichele, H., Guhde, G., Annaert, W., Von Figura, K., & Van Leuven, F. (1998). Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. *Nature*, *391*(6665), 387–390. https://doi.org/10.1038/34910

Levy-Lahad, E., Poorkaj, P., Wang, K., Ying Hui Fu, Oshima, J., Mulligan, J., & Schellenberg, G. D. (1996). Genomic structure and expression of STM2, the chromosome 1 familial Alzheimer disease gene. *Genomics*, *34*(2), 198–204. https://doi.org/10.1006/GENO.1996.0266

Tomita, T., Maruyama, K., Saido, T. C., Kume, H., Shinozaki, K., Tokuhiro, S., Capell, A., Walter, J., Grünberg, J., Haass, C., Iwatsubo, T., & Obata, K. (1997). The presenilin 2 mutation (N141I) linked to familial Alzheimer disease (Volga German families) increases the secretion of amyloid β protein ending at the 42nd (or 43rd) residue. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(5), 2025–2030. https://doi.org/10.1073/PNAS.94.5.2025/ASSET/7A15DC22-93C0-4F35-9D33-48BE06905F0C/ASSETS/GRAPHIC/PQ0473582003.JPEG

Lleó, A., Blesa, R., Gendre, J., Castellví, 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*(10), 1926–1928. https://doi.org/10.1212/WNL.57.10.1926

Desikan, R. S., Schork, A. J., Wang, Y., Witoelar, A., Sharma, M., McEvoy, L. K., Holland, D., Brewer, J. B., Chen, C. H., Thompson, W. K., Harold, D., Williams, J., Owen, M. J., O'Donovan, M. C., Pericak-Vance, M. A., Mayeux, R., Haines, J. L., Farrer, L. A., Schellenberg, G. D., … Dale, A. M. (2015). Genetic overlap between Alzheimer's disease and Parkinson's disease at the MAPT locus. *Molecular Psychiatry*, *20*(12), 1588–1595. https://doi.org/10.1038/MP.2015.6

Guerreiro, R., Wojtas, A., Bras, J., Carrasquillo, M., Rogaeva, E., Majounie, E., Cruchaga, C., Sassi, C., Kauwe, J. S. K., Younkin, S., Hazrati, L., Collinge, J., Pocock, J., Lashley, T., Williams, J., Lambert, J.-C., Amouyel, P., Goate, A.,

Rademakers, R., … Hardy, J. (2013). TREM2 variants in Alzheimer's disease. *The New England Journal of Medicine*, *368*(2), 117–127. https://doi.org/10.1056/NEJMOA1211851

Jin, S. C., Benitez, B. A., Karch, C. M., Cooper, B., Skorupa, T., Carrell, D., Norton, J. B., Hsu, S., Harari, O., Cai, Y., Bertelsen, S., Goate, A. M., & Cruchaga, C. (2014). Coding variants in TREM2 increase risk for Alzheimer's disease. *Human Molecular Genetics*, *23*(21), 5838. https://doi.org/10.1093/HMG/DDU277

Jonsson, T., Stefansson, H., Steinberg, S., Jonsdottir, I., Jonsson, P. V., Snaedal, J., Bjornsson, S., Huttenlocher, J., Levey, A. I., Lah, J. J., Rujescu, D., Hampel, H., Giegling, I., Andreassen, O. A., Engedal, K., Ulstein, I., Djurovic, S., Ibrahim-Verbaas, C., Hofman, A., … Stefansson, K. (2013). Variant of TREM2 associated with the risk of Alzheimer's disease. *The New England Journal of Medicine*, *368*(2), 107–116. https://doi.org/10.1056/NEJMOA1211103

Abduljaleel, Z., Al-Allaf, F. A., Khan, W., Athar, M., Shahzad, N., Taher, M. M., Elrobh, M., Alanazi, M. S., & El-Huneidi, W. (2014). Evidence of Trem2 Variant Associated with Triple Risk of Alzheimer's Disease. *PLOS ONE*, *9*(3), e92648. https://doi.org/10.1371/JOURNAL.PONE.0092648

Pottier, C., Wallon, D., Rousseau, S., Rovelet-Lecrux, A., Richard, A. C., Rollin-Sillaire, A., Frebourg, T., Campion, D., & Hannequin, D. (2013). TREM2 R47H variant as a risk factor for early-onset Alzheimer's disease. *Journal of Alzheimer's Disease : JAD*, *35*(1), 45–49. https://doi.org/10.3233/JAD-122311

Guerreiro, R. J., Lohmann, E., Brás, J. M., Gibbs, J. R., Rohrer, J. D., Gurunlian, N., Dursun, B., Bilgic, B., Hanagasi, H., Gurvit, H., Emre, M., Singleton, A., & Hardy, J. (2013). Using exome sequencing to reveal mutations in TREM2 presenting as a frontotemporal dementia-like syndrome without bone involvement. *JAMA Neurology*, *70*(1), 78–84. https://doi.org/10.1001/JAMANEUROL.2013.579

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24. https://doi.org/10.1016/J.AJHG.2011.11.029

Morgan, K. (2011). The three new pathways leading to Alzheimer's disease. *Neuropathology and Applied Neurobiology*, *37*(4), 353–357. https://doi.org/10.1111/J.1365-2990.2011.01181.X

Tosto, G., & Reitz, C. (2013). Genome-wide association studies in Alzheimer's disease: a review. *Current Neurology and Neuroscience Reports*, *13*(10). https://doi.org/10.1007/S11910-013-0381-0

Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development*, *19*(3), 212–219. https://doi.org/10.1016/J.GDE.2009.04.010

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, *15*(2), 256–278. https://doi.org/10.1093/BIB/BBS086

Vardarajan, B. N., Zhang, Y., Lee, J. H., Cheng, R., Bohm, C., Ghani, M., Reitz, C., Reyes-Dumeyer, D., Shen, Y., Rogaeva, E., St George-Hyslop, P., & Mayeux, R. (2015). Coding mutations in SORL1 and Alzheimer disease. *Annals of Neurology*, *77*(2), 215–227. https://doi.org/10.1002/ANA.24305

Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends in Genetics : TIG*, *17*(9), 502–510. https://doi.org/10.1016/S0168-9525(01)02410-6

Pottier, C., Hannequin, D., Coutant, S., Rovelet-Lecrux, A., Wallon, D., Rousseau, S., Legallic, S., Paquet, C., Bombois, S., Pariente, J., Thomas-Anterion, C., Michon, A., Croisile, B., Etcharry-Bouyx, F., Berr, C., Dartigues, J. F., Amouyel, P., Dauchel, H., Boutoleau-Bretonnière, C., … Campion, D. (2012). High frequency of potentially pathogenic SORL1 mutations in autosomal dominant early-onset Alzheimer disease. *Molecular Psychiatry*, *17*(9), 875–879. https://doi.org/10.1038/MP.2012.15

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., & Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, *42*(1), 30–35. https://doi.org/10.1038/ng.499

Grada, A., & Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *The Journal of Investigative Dermatology*, *133*(8). https://doi.org/10.1038/JID.2013.248

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/PNAS.74.12.5463

Goh, G., & Choi, M. (2012). Application of Whole Exome Sequencing to Identify Disease-Causing Variants in Inherited Human Diseases. *Genomics & Informatics*, *10*(4), 214. https://doi.org/10.5808/GI.2012.10.4.214

Rosenthal, S. L., & Kamboh, M. I. (2014). Late-Onset Alzheimer's Disease Genes and the Potentially Implicated Pathways. *Current Genetic Medicine Reports*, *2*(2), 85–101. https://doi.org/10.1007/S40142-014-0034-X

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. *Nature Reviews. Genetics*, *12*(11), 745–755. https://doi.org/10.1038/nrg3031

Sims, R., Van Der Lee, S. J., Naj, A. C., Bellenguez, C., Badarinarayan, N., Jakobsdottir, J., Kunkle, B. W., Boland, A., Raybould, R., Bis, J. C., Martin, E. R., Grenier-Boley, B., Heilmann-Heimbach, S., Chouraki, V., Kuzma, A. B., Sleegers, K., Vronskaya, M., Ruiz, A., Graham, R. R., … Schellenberg, G. D. (2017). Rare coding variants in PLCG2, ABI3 and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nature Genetics*, *49*(9), 1373. https://doi.org/10.1038/NG.3916

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics 2014 15:2*, *15*(2), 121–132. https://doi.org/10.1038/nrg3642

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*(1), 27. https://doi.org/10.1016/J.CELL.2013.09.006

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics 2011 43:5*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome Sequencing: Current and Future Perspectives. *G3 (Bethesda, Md.)*, *5*(8), 1543–1550. https://doi.org/10.1534/G3.115.018564

Guo, Y., Long, J., He, J., Li, C. I., Cai, Q., Shu, X. O., Zheng, W., & Li, C. (2012). Exome sequencing generates high quality data in non-target regions. *BMC Genomics*, *13*(1), 1–10. https://doi.org/10.1186/1471-2164-13-194/TABLES/4

Nalls, M. A., Bras, J., Hernandez, D. G., Keller, M. F., Majounie, E., Renton, A. E., Saad, M., Jansen, I., Guerreiro, R., Lubbe, S., Plagnol, V., Gibbs, J. R., Schulte, C., Pankratz, N., Sutherland, M., Bertram, L., Lill, C. M., Destefano, A. L., Faroud, T., … Tzourio, C. (2015). NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiology of Aging*, *36*(3), 1605.e7-1605.e12. https://doi.org/10.1016/j.neurobiolaging.2014.07.028

Blauwendraat, C., Faghri, F., Pihlstrom, L., Geiger, J. T., Elbaz, A., Lesage, S., Corvol, J. C., May, P., Nicolas, A., Abramzon, Y., Murphy, N. A., Gibbs, J. R., Ryten, M., Ferrari, R., Bras, J., Guerreiro, R., Williams, J., Sims, R., Lubbe, S., … Heutink, P. (2017). NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. *Neurobiology of Aging*, *57*, 247.e9-247.e13. https://doi.org/10.1016/j.neurobiolaging.2017.05.009

Barber, I. S., Braae, A., Clement, N., Patel, T., Guetta-Baranes, T., Brookes, K., Medway, C., Chappell, S., Guerreiro, R., Bras, J., Hernandez, D., Singleton,

A., Hardy, J., Mann, D. M., Passmore, P., Craig, D., Johnston, J., McGuinness, B., Todd, S., … Morgan, K. (2017). Mutation analysis of sporadic early-onset Alzheimer's disease using the NeuroX array. *Neurobiology of Aging*, *49*, 215.e1-215.e8. https://doi.org/10.1016/J.NEUROBIOLAGING.2016.09.008

Illumina. (2011). *HumanExome BeadChips*. http://www.illumina.com/

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/NG.3656

Lambert, S. A., Abraham, G., & Inouye, M. (2019). Towards clinical utility of polygenic risk scores. *Human Molecular Genetics*, *28*(2), 133–142. https://doi.org/10.1093/hmg/ddz187

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic Risk Score software. *Bioinformatics (Oxford, England)*, *31*(9), 1466–1468. https://doi.org/10.1093/BIOINFORMATICS/BTU848

Escott-Price, V., Myers, A. J., Huentelman, M., & Hardy, J. (2017). Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Annals of Neurology*, *82*(2), 311–314. https://doi.org/10.1002/ANA.24999

Porter, T., Burnham, S. C., Milicic, L., Savage, G., Maruff, P., Lim, Y. Y., Li, Q. X., Ames, D., Masters, C. L., Rainey-Smith, S., Rowe, C. C., Salvado, O., Groth, D., Verdile, G., Villemagne, V. L., & Laws, S. M. (2018). Utility of an Alzheimer's Disease Risk-Weighted Polygenic Risk Score for Predicting Rates of Cognitive Decline in Preclinical Alzheimer's Disease: A Prospective Longitudinal Study. *Journal of Alzheimer's Disease*, *66*(3), 1193–1211. https://doi.org/10.3233/JAD-180713

Foo, H., Thalamuthu, A., Jiang, J., Koch, F., Mather, K. A., Wen, W., & Sachdev, P. S. (2021). Associations between Alzheimer's disease polygenic risk scores and hippocampal subfield volumes in 17,161 UK Biobank participants. *Neurobiology of Aging*, *98*, 108–115. https://doi.org/10.1016/j.neurobiolaging.2020.11.002

Darst, B. F., Koscik, R. L., Racine, A. M., Oh, J. M., Krause, R. A., Carlsson, C. M., Zetterberg, H., Blennow, K., Christian, B. T., Bendlin, B. B., Okonkwo, O. C., Hogan, K. J., Hermann, B. P., Sager, M. A., Asthana, S., Johnson, S. C., & Engelman, C. D. (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. *Journal of Alzheimer's Disease*, *55*(2), 473–484. https://doi.org/10.3233/JAD-160195

Logue, M. W., Panizzon, M. S., Elman, J. A., Gillespie, N. A., Hatton, S. N., Gustavson, D. E., Andreassen, O. A., Dale, A. M., Franz, C. E., Lyons, M. J., Neale, M. C., Reynolds, C. A., Tu, X., & Kremen, W. S. (2019). Use of an

Alzheimer's disease polygenic risk score to identify mild cognitive impairment in adults in their 50s. *Molecular Psychiatry*, *24*(3), 421–430. https://doi.org/10.1038/s41380-018-0030-8

Chaudhury, S., Brookes, K. J., Patel, T., Fallows, A., Guetta-Baranes, T., Turton, J. C., Guerreiro, R., Bras, J., Hardy, J., Francis, P. T., Croucher, R., Holmes, C., & Morgan, K. (2019). Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment. *Translational Psychiatry*, *9*(1). https://doi.org/10.1038/s41398-019-0485-7

Euesden, J., Gowrisankar, S., Qu, A. X., Jean, P. S., Hughes, A. R., & Pulford, D. J. (2020). Cognitive decline in alzheimer's disease: Limited clinical utility for gwas or polygenic risk scores in a clinical trial setting. *Genes*, *11*(5). https://doi.org/10.3390/genes11050501

Lawingco, T., Chaudhury, S., Brookes, K. J., Guetta-Baranes, T., Guerreiro, R., Bras, J., Hardy, J., Francis, P., Thomas, A., Belbin, O., & Morgan, K. (2021). Genetic variants in glutamate-, Aβ−, and tau-related pathways determine polygenic risk for Alzheimer's disease. *Neurobiology of Aging*, *101*, 299.e13-299.e21. https://doi.org/10.1016/j.neurobiolaging.2020.11.009

Hu, Y. S., Xin, J., Hu, Y., Zhang, L., & Wang, J. (2017). Analyzing the genes related to Alzheimer's disease via a network and pathway-based approach. *Alzheimer's Research & Therapy*, *9*(1). https://doi.org/10.1186/S13195-017-0252-Z

Femminella, G. D., Harold, D., Scott, J., Williams, J., & Edison, P. (2021). The Differential Influence of Immune, Endocytotic, and Lipid Metabolism Genes on Amyloid Deposition and Neurodegeneration in Subjects at Risk of Alzheimer's Disease. *Journal of Alzheimer's Disease*, *79*(1), 127–139. https://doi.org/10.3233/JAD-200578

Bellou, E., Stevenson-Hoare, J., & Escott-Price, V. (2020). Polygenic risk and pleiotropy in neurodegenerative diseases. In *Neurobiology of Disease* (Vol. 142). Academic Press Inc. https://doi.org/10.1016/j.nbd.2020.104953

Axelrud, L. K., Sato, J. R., Santoro, M. L., Talarico, F., Pine, D. S., Rohde, L. A., Zugman, A., Junior, E. A., Bressan, R. A., Grassi-Oliveira, R., Pan, P. M., Hoffmann, M. S., Simioni, A. R., Guinjoan, S. M., Hakonarson, H., Brietzke, E., Gadelha, A., Pellegrino da Silva, R., Hoexter, M. Q., … Salum, G. A. (2019). Genetic risk for Alzheimer's disease and functional brain connectivity in children and adolescents. *Neurobiology of Aging*, *82*, 10–17. https://doi.org/10.1016/j.neurobiolaging.2019.06.011

Chaudhury, S., Patel, T., Barber, I. S., Guetta-Baranes, T., Brookes, K. J., Chappell, S., Turton, J., Guerreiro, R., Bras, J., Hernandez, D., Singleton, A., Hardy, J., Mann, D., Passmore, P., Craig, D., Johnston, J., McGuinness, B., Todd, S., Heun, R., … Morgan, K. (2018). Polygenic risk score in postmortem diagnosed sporadic early-onset Alzheimer's disease. *Neurobiology of Aging*, *62*, 244.e1-244.e8. https://doi.org/10.1016/j.neurobiolaging.2017.09.035

Cruchaga, C., Del-Aguila, J. L., Saef, B., Black, K., Fernandez, M. V., Budde, J., Ibanez, L., Deming, Y., Kapoor, M., Tosto, G., Mayeux, R. P., Holtzman, D. M., Fagan, A. M., Morris, J. C., Bateman, R. J., Goate, A. M., & Harari, O. (2018). Polygenic risk score of sporadic late-onset Alzheimer's disease reveals a shared architecture with the familial and early-onset forms. *Alzheimer's and Dementia*, *14*(2), 205–214. https://doi.org/10.1016/j.jalz.2017.08.013

Del-Aguila, J. L., Fernández, M. V., Schindler, S., Ibanez, L., Deming, Y., Ma, S., Saef, B., Black, K., Budde, J., Norton, J., Chasse, R., Harari, O., Goate, A., Xiong, C., Morris, J. C., & Cruchaga, C. (2018). Assessment of the Genetic Architecture of Alzheimer's Disease Risk in Rate of Memory Decline. *Journal of Alzheimer's Disease*, *62*(2), 745–756. https://doi.org/10.3233/JAD-170834

Ebenau, J. L., van der Lee, S. J., Hulsman, M., Tesi, N., Jansen, I. E., Verberk, I. M. W., van Leeuwenstijn, M., Teunissen, C. E., Barkhof, F., Prins, N. D., Scheltens, P., Holstege, H., van Berckel, B. N. M., & van der Flier, W. M. (2021). Risk of dementia in APOE ε4 carriers is mitigated by a polygenic risk score. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *13*(1). https://doi.org/10.1002/dad2.12229

Elman, J. A., Vuoksimaa, E., Franz, C. E., & Kremen, W. S. (2020). Degree of cognitive impairment does not signify early versus late mild cognitive impairment: confirmation based on Alzheimer's disease polygenic risk. *Neurobiology of Aging*, *94*, 149–153. https://doi.org/10.1016/j.neurobiolaging.2020.05.015

Escott-Price, V., & Schmidt, K. M. (2021). Probability of Alzheimer's disease based on common and rare genetic variants. *Alzheimer's Research and Therapy*, *13*(1). https://doi.org/10.1186/s13195-021-00884-7

Escott-Price, V., Myers, A., Huentelman, M., Shoai, M., & Hardy, J. (2019). Polygenic Risk Score Analysis of Alzheimer's Disease in Cases without APOE4 or APOE2 Alleles. *The Journal of Prevention of Alzheimer's Disease*, *6*(1), 16–19. https://doi.org/10.14283/jpad.2018.46

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. *Neurobiology of Aging*, *49*, 214.e7. https://doi.org/10.1016/J.NEUROBIOLAGING.2016.07.018

Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., Badarinarayan, N., Morgan, K., Passmore, P., Holmes, C., Powell, J., Brayne, C., Gill, M., Mead, S., Goate, A., Cruchaga, C., Lambert, J. C., Van Duijn, C., Maier, W., … Williams, J. (2015). Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain : A Journal of Neurology*, *138*(Pt 12), 3673–3684. https://doi.org/10.1093/BRAIN/AWV268

Fulton-Howard, B., Goate, A. M., Adelson, R. P., Koppel, J., Gordon, M. L., Barzilai, N., Atzmon, G., Davies, P., & Freudenberg-Hua, Y. (2021). Greater effect of polygenic risk score for Alzheimer's disease among younger cases who are apolipoprotein E-ε4 carriers. *Neurobiology of Aging*, *99*, 101.e1-101.e9. https://doi.org/10.1016/j.neurobiolaging.2020.09.014

Ge, T., Sabuncu, M. R., Smoller, J. W., Sperling, R. A., & Mormino, E. C. (2018). Dissociable influences of APOE «4 and polygenic risk of AD dementia on amyloid and cognition. *Neurology*, *90*(18), E1605–E1612. https://doi.org/10.1212/WNL.0000000000005415

Gibson, J., Russ, T. C., Adams, M. J., Clarke, T. K., Howard, D. M., Hall, L. S., Fernandez-Pujals, A. M., Wigmore, E. M., Hayward, C., Davies, G., Murray, A. D., Smith, B. H., Porteous, D. J., Deary, I. J., & McIntosh, A. M. (2017). Assessing the presence of shared genetic architecture between Alzheimer's disease and major depressive disorder using genome-wide association data. *Translational Psychiatry*, *7*(4). https://doi.org/10.1038/tp.2017.49

Hong, S., Prokopenko, D., Dobricic, V., Kilpert, F., Bos, I., Vos, S. J. B., Tijms, B. M., Andreasson, U., Blennow, K., Vandenberghe, R., Cleynen, I., Gabel, S., Schaeverbeke, J., Scheltens, P., Teunissen, C. E., Niemantsverdriet, E., Engelborghs, S., Frisoni, G., Blin, O., … Bertram, L. (2020). Genome-wide association study of Alzheimer's disease CSF biomarkers in the EMIF-AD Multimodal Biomarker Discovery dataset. *Translational Psychiatry*, *10*(1). https://doi.org/10.1038/s41398-020-01074-z

Huq, A. J., Fulton-Howard, B., Riaz, M., Laws, S., Sebra, R., Ryan, J., Renton, A. E., Goate, A. M., Masters, C. L., Storey, E., Shah, R. C., Murray, A., McNeil, J., Winship, I., James, P. A., & Lacaze, P. (2021). Polygenic score modifies risk for Alzheimer's disease in APOE ε4 homozygotes at phenotypic extremes. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *13*(1). https://doi.org/10.1002/dad2.12226

Kauppi, K., Rönnlund, M., Nordin Adolfsson, A., Pudas, S., & Adolfsson, R. (2020). Effects of polygenic risk for Alzheimer's disease on rate of cognitive decline in normal aging. *Translational Psychiatry*, *10*(1). https://doi.org/10.1038/s41398-020-00934-y

Korologou-Linden, R., Anderson, E. L., Jones, H. J., Davey Smith, G., Howe, L. D., & Stergiakouli, E. (2019). Polygenic risk scores for Alzheimer's disease, and academic achievement, cognitive and behavioural measures in children from the general population. *International Journal of Epidemiology*, *48*(6), 1972–1980. https://doi.org/10.1093/ije/dyz080

Kremen, W. S., Panizzon, M. S., Elman, J. A., Granholm, E. L., Andreassen, O. A., Dale, A. M., Gillespie, N. A., Gustavson, D. E., Logue, M. W., Lyons, M. J., Neale, M. C., Reynolds, C. A., Whitsel, N., & Franz, C. E. (2019). Pupillary dilation responses as a midlife indicator of risk for Alzheimer's disease: association with Alzheimer's disease polygenic risk. *Neurobiology of*

Aging, *83*, 114–121. https://doi.org/10.1016/j.neurobiolaging.2019.09.001

Leonenko, G., Shoai, M., Bellou, E., Sims, R., Williams, J., Hardy, J., & Escott-Price, V. (2019). Genetic risk for alzheimer disease is distinct from genetic risk for amyloid deposition. *Annals of Neurology*, *86*(3), 427–435. https://doi.org/10.1002/ana.25530

Leonenko, G., Baker, E., Stevenson-Hoare, J., Sierksma, A., Fiers, M., Williams, J., de Strooper, B., & Escott-Price, V. (2021). Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-24082-z

Leonenko, G., Sims, R., Shoai, M., Frizzati, A., Bossù, P., Spalletta, G., Fox, N. C., Williams, J., Hardy, J., & Escott-Price, V. (2019). Polygenic risk and hazard scores for Alzheimer's disease prediction. *Annals of Clinical and Translational Neurology*, *6*(3), 456–465. https://doi.org/10.1002/acn3.716

Tasaki, S., Gaiteri, C., Mostafavi, S., De Jager, P. L., & Bennett, D. A. (2018). The molecular and neuropathological consequences of genetic risk for Alzheimer's dementia. *Frontiers in Neuroscience*, *12*(OCT). https://doi.org/10.3389/fnins.2018.00699

Tasaki, S., Gaiteri, C., Petyuk, V. A., Blizinsky, K. D., De Jager, P. L., Buchman, A. S., & Bennett, D. A. (2019). Genetic risk for Alzheimer's dementia predicts motor deficits through multi-omic systems in older adults. *Translational Psychiatry*, *9*(1). https://doi.org/10.1038/s41398-019-0577-4

Wehby, G. L., Domingue, B. W., & Wolinsky, F. D. (2018). Genetic Risks for Chronic Conditions: Implications for Long-term Wellbeing. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *73*(4), 477–483. https://doi.org/10.1093/gerona/glx154

Yesavage, J. A., Noda, A., Heath, A., McNerney, M. W., Domingue, B. W., Hernandez, Y., Benson, G., Hallmayer, J., O'Hara, R., Williams, L. M., Goldstein-Piekarski, A. N., Zeitzer, J. M., & Fairchild, J. K. (2020). Sleep-wake disorders in Alzheimer's disease: Further genetic analyses in relation to objective sleep measures. *International Psychogeriatrics*, *32*(7), 807–813. https://doi.org/10.1017/S1041610219001777

Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., Boland, A., Vronskaya, M., van der Lee, S. J., Amlie-Wolf, A., Bellenguez, C., Frizatti, A., Chouraki, V., Martin, E. R., Sleegers, K., Badarinarayan, N., Jakobsdottir, J., Hamilton-Nelson, K. L., Moreno-Grau, S., … Pericak-Vance, M. A. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nature Genetics 2019 51:9*, *51*(9), 1423–1424. https://doi.org/10.1038/s41588-019-0495-7

Stocker, H., Perna, L., Weigl, K., Möllers, T., Schöttker, B., Thomsen, H., Holleczek, B., Rujescu, D., & Brenner, H. (2021). Prediction of clinical diagnosis of Alzheimer's disease, vascular, mixed, and all-cause dementia

by a polygenic risk score and APOE status in a community-based cohort prospectively followed over 17 years. *Molecular Psychiatry*, *26*(10), 5812–5822. https://doi.org/10.1038/s41380-020-0764-y

Skoog, I., Kern, S., Najar, J., Guerreiro, R., Bras, J., Waern, M., Zetterberg, H., Blennow, K., & Zettergren, A. (2021). A non-APOEPolygenic risk score foralzheimer's disease is associated with cerebrospinal fluid neurofilament light in a representative sample of cognitively unimpaired 70-year olds. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *76*(6), 983–990. https://doi.org/10.1093/gerona/glab030

Najar, J., van der Lee, S. J., Joas, E., Wetterberg, H., Hardy, J., Guerreiro, R., Bras, J., Waern, M., Kern, S., Zetterberg, H., Blennow, K., Skoog, I., & Zettergren, A. (2021). Polygenic risk scores for alzheimer's disease are related to dementia risk in apoe ε4 negatives. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, *13*(1). https://doi.org/10.1002/dad2.12142

Wightman, D. P., Jansen, I. E., Savage, J. E., Shadrin, A. A., Bahrami, S., Holland, D., Rongve, A., Børte, S., Winsvold, B. S., Drange, O. K., Martinsen, A. E., Skogholt, A. H., Willer, C., Bråthen, G., Bosnes, I., Nielsen, J. B., Fritsche, L. G., Thomas, L. F., Pedersen, L. M., … Posthuma, D. (2021). A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nature Genetics*, *53*(9), 1276–1282. https://doi.org/10.1038/S41588-021-00921-Z

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, *9*(3). https://doi.org/10.1371/JOURNAL.PGEN.1003348

Jin, X., Shu, C., Zeng, Y., Liang, L., & Ji, J. S. (2021). Interaction of greenness and polygenic risk score of Alzheimer's disease on risk of cognitive impairment. *Science of the Total Environment*, *796*. https://doi.org/10.1016/j.scitotenv.2021.148767

Desikan, R. S., Fan, C. C., Wang, Y., Schork, A. J., Cabral, H. J., Cupples, L. A., Thompson, W. K., Besser, L., Kukull, W. A., Holland, D., Chen, C. H., Brewer, J. B., Karow, D. S., Kauppi, K., Witoelar, A., Karch, C. M., Bonham, L. W., Yokoyama, J. S., Rosen, H. J., … Dale, A. M. (2017). Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLOS Medicine*, *14*(3), e1002258. https://doi.org/10.1371/JOURNAL.PMED.1002258

Lleó, A., Núñez-Llaves, R., Alcolea, D., Chiva, C., Balateu-Paños, D., Colom-Cadena, M., Gomez-Giro, G., Muñoz, L., Querol-Vilaseca, M., Pegueroles, J., Rami, L., Lladó, A., Molinuevo, J. L., Tainta, M., Clarimón, J., Spires-Jones, T., Blesa, R., Fortea, J., Martínez-Lage, P., … Belbin, O. (2019). Changes in Synaptic Proteins Precede Neurodegeneration Markers in Preclinical Alzheimer's Disease Cerebrospinal Fluid. *Molecular & Cellular Proteomics : MCP*, *18*(3), 546–560. https://doi.org/10.1074/MCP.RA118.001290

Choi, S. W., & O'Reilly, P. F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*, *8*(7), 1–6. https://doi.org/10.1093/GIGASCIENCE/GIZ082

Francis, P. T., Costello, H., & Hayes, G. M. (2018). Brains for Dementia Research: Evolution in a Longitudinal Brain Donation Cohort to Maximize Current and Future Value. *Journal of Alzheimer's Disease*, *66*(4), 1635–1644. https://doi.org/10.3233/JAD-180699

Sussams, R., Schlotz, W., Perry, H., Viv, H., Lynn, D., Rayner, C., Lewzey, I., Christodoulou, A., MacFarlene, B., Sharples, R., & Holmes, C. (2013). P4–191: Systemic inflammatory responses to stress and its impact on cognition in people with mild cognitive impairment. *Alzheimer's & Dementia*, *9*(4S_Part_19), P775–P775. https://doi.org/10.1016/J.JALZ.2013.05.1582

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Kent, J. W., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102. Article published online before print in May 2002