Regulatory genomic consequences of polygenic risk burden for Alzheimer’s disease

Submitted by

Gemma Shireby

To the University of Exeter

as a thesis for the degree of

Doctor of Philosophy in Medical Studies

in March 2021

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ………………………………………………………………..

Abstract

Dementia is an umbrella term used to describe a group of symptoms associated with global cognitive impairment and is a major contributor to the global burden of disease; currently there are over 50 million individuals affected world-wide. Due to the ageing population and lack of effective disease-modifying treatments, this number is expected to triple by 2050. Dementia encompasses a number of neurological diseases, including Alzheimer's disease (AD), which accounts for 60-80% of cases. There is a well-established genetic component to AD and genome wide-association studies have identified >75 variants robustly associated with disease. Little is known about the functional mechanisms by which risk variants mediate disease susceptibility; as the majority of these variants do not index coding variants affecting protein structure they are hypothesised to influence gene regulation, supported by the observation that they are enriched in regulatory domains including enhancers.

The primary aim of this thesis was to assess whether genetic liability for AD is associated with regulatory genomic variation (i.e. epigenetic and transcriptomic) in whole blood and the human cortex. Epigenome-wide association studies and multi-omic methods were utilised to explore the molecular mechanisms leading to disease. The results from this thesis indicate that epigenetic mechanisms are involved in AD pathogenesis and provide further support for several established AD pathways such as lipid and cholesterol metabolism, Aβ, tau and *APP* processing as well as a role for the immune system. The analyses incorporating AD genetic variation with DNA methylation infer that there are both direct *cis* genetic effects and indirect polygenic effects on regulatory processes which are involved in the aetiology of AD. Although there were consistencies at some loci across the whole blood and cortex analyses, there was also evidence for heterogeneity across tissues which might represent tissue specific effects in areas primarily affected in AD (e.g. the cortex) in comparison to peripheral tissues.

In summary, using multiple approaches, I characterised the complex relationship between genetic and epigenetic variation, enabling the exploration of molecular genomic mechanisms driving AD pathogenesis in both peripheral and brain tissues and prioritised genes which could be targeted in future functional studies.

Acknowledgements

I am grateful for the support I received while working towards my PhD. First, I would like to start by thanking my supervisors Professor Jonathan Mill and Dr Eilis Hannon. I am lucky to have had such amazing guidance over the course of the past three years and I hugely appreciate the expertise you have shared with me. You both always found the time to discuss any issues I had and encouraged me to get involved in many incredible projects. Jon, thank you for your continual support, enthusiasm and insightful feedback. Eilis, I am forever in awe of your statistical abilities - thank you for all of the advice and knowledge you have given me. You both inspire me, and I could not have asked for better supervisors.

I would like to thank all of the Complex Disease Epigenomics Group. I am fortunate to have been able to work with such an incredible group of individuals. Szi Kay and Thea, thank you for being great housemates and friends to me; having a strong support network at home helped me endlessly with completing my PhD. Isabel, I was very lucky to have been sat next to you when I first arrived - you encouraged me to get involved with numerous outreach projects and your drive never ceases to inspire me.

Charli, I would have struggled to complete my PhD without you. Thank you for always being there for me, listening to my worries, encouraging me and for all the adventures we had together – there will definitely be many more. To my amazing friends Becca, Lizzie and Rowena, our friendship is invaluable, and I appreciate everything you have done for me over the years, thank you for being by side through thick and thin. The master's girls - Isobel, Olatz, Molly, Dina and Nastashia - you have motivated me throughout the whole process and as well providing friendship, you are all inspirational scientists and I thank you for both the scientific discussions and the good times we have shared since meeting on the GED course.

I would like to thank my family – Mum, Dad and Jack – who I know are incredibly proud of me and who have always encouraged me to follow my passions. You have helped me get to where I am today and I could not have done it without your support. Cam, thank you for your patience and kindness throughout this whole process. I appreciate everything you have done to help me (including enduring listening to my presentations and helping me fix my formatting disasters) - you have enabled me to get through this with my sanity. I look forward to celebrating with everyone!

Table of Contents

Table of Figures

Table of Tables

Publications arising from this thesis

Chapter 3

(Accepted manuscript presented in Chapter 3)

Shireby, G. L., Davies, J. P., Francis, P. T., Burrage, J., Walker, E. M., Neilson, G. W., ... & Mill, J. (2020). Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex. *Brain*, *143*(12), 3763-3775.

Steg, L. C., **Shireby, G. L**., Imm, J., Davies, J. P., Franklin, A., Flynn, R., ... & Hannon, E. (2021). Novel epigenetic clock for fetal brain development predicts prenatal age for cellular stem cell models and derived neurons. *Molecular Brain*, 14, 98.

Stevenson, A. J., McCartney, D. L., **Shireby, G. L**., Hillary, R. F., King, D., Tzioras, M., ... & Spires-Jones, T. L. (2020). A comparison of blood and brain-derived ageing and inflammation-related DNA methylation signatures and their association with microglial burdens. *bioRxiv*.

Grodstein, F., Lemos, B., Yu, L., Klein, H. U., Iatrou, A., Buchman, A. S., **Shireby, G. L**., ... & Bennett, D. A. (2021). The association of epigenetic clocks in brain tissue with brain pathologies and common aging phenotypes. Neurobiology of Disease, 105428.

Chapter 4

(Accepted manuscript presented in Appendix A)

Smith, R. G., Pishva, E., **Shireby, G**., Smith, A. R., Roubroeks, J. A., Hannon, E., ... & Lunnon, K. (2021). A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex. Nature communications, 12(1), 1-13.

Chapter 5

van Rheenen, W., van der Spek, R.A.A., Bakker, M.K., van Vugt, J, Hop, P., Zwamborn, R., de Klein., N , Westra, H.J., Bakker, O., Deelen , P., **Shireby G.L,** … & Veldink, J.H. (2021). Common and rare variant association analyses in Amyotrophic Lateral Sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Submitted.*

Chapter 6

(Accepted manuscript presented in Appendix A)

Hannon, E., **Shireby, G. L**., Brookes, K., Attems, J., Sims, R., Cairns, N. J., ... & Mill, J. (2020). Genetic risk for Alzheimer's disease influences neuropathology via multiple biological pathways. *Brain communications*, *2*(2), fcaa167.

Other publications arising from this thesis

Hop, P. J., Zwamborn, R. A., Hannon, E., **Shireby, G. L**., Nabais, M. F., Walker, E. M., ... & Brain MEND Consortium. (2021). Genome-wide study of DNA methylation in Amyotrophic Lateral Sclerosis identifies differentially methylated loci and implicates metabolic, inflammatory and cholesterol pathways. medRxiv.

Rovira, P., Sánchez-Mora, C., Pagerols, M., Richarte, V., Corrales, M., Fadeuilhe, C., Vilar-Ribó, L., Arribas, L., **Shireby, G**.**L**, Hannon, E. & Mill, J., (2020). Epigenome-wide association study of attention-deficit/hyperactivity disorder in adults. *Translational psychiatry*, *10*(1), 1-12.

Declarations

Brains for Dementia research (BDR)

The BDR human post-mortem tissue used in **Chapter 3-6** was obtained from the Southwest Dementia Brain Bank, London Neurodegenerative Diseases Brain Bank, Manchester Brain Bank, Newcastle Brain Tissue Resource and Oxford Brain Bank. DNA methylation data were generated by Dr Joe Burrage and Dr Jonathan Davies.

Exeter ten thousand (EXTEND)

Whole blood tissue and data collection for EXTEND was conducted by the NIHR Exeter Clinical Research Facility. DNA methylation data was generated by Dr Joe Burrage. These data were used in **Chapter 5** and **6**.

Statistical analysis

All of the bioinformatic and statistical analyses were performed by me with the exception of:

- The quality control (QC) of BDR was conducted by Emma Walker, Grant Neilson and Aisha Dahir and me.
- The QC of the first two EXTEND plates was conducted by Georgina Mansell, and the remaining plates were QC'd by me.
- The genetic analyses of neuropathology presented in **Chapter 6** were conducted by Dr Eilis Hannon.

Abbreviations

5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
AD	Alzheimer's disease
APP	Amyloid precursor protein
AUC	Area under the curve
Aβ	Amyloid-beta
bp	Base pair(s)
CpG	Cytosine-guanine dinucleotide
CSF	Cerebrospinal fluid
DMP	Differentially methylated position
DMR	Differentially methylated region
DN	Dystrophic neurites
DNAm	DNA methylation
EOAD	Early onset Alzheimer's disease
eQTL	expression quantitative trait loci
EWAS	Epigenome-wide association study/ studies
fEOAD	Familial early onset Alzheimer's disease
FTD	Frontotemporal dementia
GO	Gene ontology
GWAS	Genome-wide association study/ studies
kb	Kilobases
LB	Lewy body
LBD	Lewy body dementia
LD	Linkage disequilibrium
LOAD	Late onset Alzheimer's disease
mQTL	Methylation quantitative trait loci
NCI	Neuronal cytoplasmic inclusions
NFT	Neurofibrillary tangle
NII	Intra-neuronal inclusions
OCC	Occipital cortex
PD	Parkinson's disease
PFC	Prefrontal Cortex

PRS	Polygenic risk score(s)
PRS-mQTL	Methylation quantitative trait loci included in a polygenic risk score
PRS-Jansen	Polygenic risk score calculated using the Jansen *at al.* (2019) GWAS excluding *APOE*
PRS-Jansen$_{\mathbf{APOE}}$	Polygenic risk score calculated using the Jansen *at al.* (2019) GWAS including *APOE*
PRS-Kunkle	Polygenic risk score calculated using the Kunkle *at al.* (2019) GWAS excluding *APOE*
PRS-Kunkle$_{\mathbf{APOE}}$	Polygenic risk score calculated using the Kunkle *at al.* (2019) GWAS including *APOE*
PSEN	Presenilin
QC	Quality control
QQ	Quantile-quantile
sEOAD	Sporadic early onset Alzheimer's disease
SNP	Single nucleotide polymorphism
TET	Ten-eleven translocation
VaD	Vascular Dementia
WGCNA	Weighted correlation network analysis

1 General Introduction

1.1 Alzheimer's disease and Other Dementias

Dementia is an umbrella term used to describe a group of symptoms associated with global cognitive impairment and encompasses a number of neurological diseases, including Alzheimer's disease (AD), vascular dementia (VaD), dementia with Lewy bodies (DLB), Parkinson's disease (PD) and Frontotemporal dementia (FTD) (Lobo *at al.*, 2000).

Dementia is a major contributor to the global burden of disease and currently there are over 50 million individuals affected world-wide (Alzheimer's Association, 2019). Due to the ageing population and lack of effective disease-modifying treatments, this number is expected to triple by 2050 (Alzheimer's Association, 2019). Dementia is estimated to cost the economy over one trillion dollars (£760 billion) annually (Alzheimer's Association, 2019; Wimo *at al.*, 2017), and based on the current trajectory, this number is predicted to surpass two trillion dollars by 2030 (Alzheimer's Association, 2019; Wimo *at al.*, 2017).

1.1.1 Clinical Manifestation

In the International Classification of Diseases version 11 (ICD-11) dementia is categorised within 'Mental and behavioural disorders' (categories 6D8[X]) (World Health Organization, 2020). The ICD-11 clinically characterises dementia as a decline in cognitive functioning beyond that of normal ageing, with impairment affecting at least two of the following cognitive domains: memory, executive function, attention, language, judgement, social cognition, psychomotor speed and visual perception/ spatial abilities (World Health Organization, 2020). The symptoms vary for different dementia subtypes (see **Table 1.1**).

AD is the most common form of dementia, accounting for ~60-80% of cases (Alzheimer's Association, 2019). The primary symptom of AD is memory impairment and as the disease progresses there is a steady decline in cognitive functioning affecting cognitive domains such as executive functioning, attention, language, social cognition and judgement (World Health Organization, 2020). In the earlier stages, dementia due to AD is often accompanied by both psychiatric and behavioural symptoms such as apathy and depression. In the later stages individual's may present with psychotic symptoms, aggression, confusion, gait abnormalities and mobility issues (World Health Organization, 2020).

In VaD, the second most common form of dementia (8-15% of cases) (Goodman *at al.*, 2017), the onset of cognitive deficits is related to vascular events (i.e. a result of poor circulation to the brain) (World Health Organization, 2020). The most affected cognitive domains are those involved in processing speed, attention and fontal-executive functioning (World Health Organization, 2020).

DLB is the third most common form of dementia (4.6% of cases) (Kane *at al.*, 2018) and is initially characterised by cognitive deficits to attentional and executive functioning which can be accompanied by hallucinations, sleep disturbances, depression and delusions (World Health Organization, 2020). Within a year of cognitive symptoms first presenting there is usually onset of Parkinsonism (movement abnormalities including tremors, slow movement, muscle stiffness and impaired speech).

PD is diagnosed when Parkinsonian motor symptoms precedes cognitive impairment (and vice versa in DLB) and the primary manifestations include bradykinesia (slowness of movement) plus at least one of the following motor symptoms: tremor, rigidity or postural instability (World Health Organization, 2020). In addition, non-motor manifestations may present in individuals with PD including neuropsychiatric features and autonomic dysfunction (World Health Organization, 2020).

FTD encompasses a group of neurodegenerative disorders which affect the frontal and temporal lobes (World Health Organization, 2020). Onset of FTD is usually characterised by behavioural changes such as executive dysfunction, declining social cognition, apathy and repetitive behaviours or by linguistic deficits including issues with comprehension and speaking. However, memory function, motor function and visual perception/spatial abilities usually are usually not affected in FTD, especially in the early stages (World Health Organization, 2020).

Table 1.1: Characteristics of the most prevalent neurodegenerative diseases which are characterised by dementia. *Table adapted from MacBean at al. (2020).*

Phenotype	**Prevalence (% of dementia cases)**	**Neuropathological hallmarks**	**Pathological progression**	**Clinical characteristics**
Alzheimer's Disease (AD)	60-80 (Alzheimer's Association, 2019)	Amyloid-beta (Aβ) Neurofibrillary tangles of Tau (NFT)	Aβ deposits are found in the neocortex and then build up in the hippocampus, then striatum and finally the cerebellum and other brain regions. NFT starts to build up in the trans-entorhinal cortex spreading to the hippocampus, and then to the neocortex.	Memory impairment, apathy, depression, impaired communication, disorientation, confusion and behavioural changes. At the later stages individuals may present with psychotic symptoms, aggression, changes to gait and mobility issues.
Dementia with Lewy Bodies (DLB)	4.6 (Kane *at al.*, 2018)	Alpha-synuclein (α-synuclein)	Build-up of α-synuclein deposition starts in the brainstem and progresses to the substantia nigra of the midbrain, then the transentorhinal region, the hippocampus and eventually to the neocortex.	Memory impairment, impaired judgement and ability to decide or plan, difficulties with motor functioning (e.g. poor balance).
Parkinson's Disease (PD)	2 (Kane *at al.*, 2018; Nussbaum & Ellis, 2003)	Alpha-synuclein (α-synuclein)	Pathology is similar to DLB: build-up of α-synuclein deposition starts in the brainstem and progresses to the limbic regions as pathology increases.	Impaired motor functioning and movement including tremors, poor balance and changes in gait.
Vascular Dementia (VaD)	8-15 (Goodman *at al.*, 2017)	To date, there are no accepted neuropathological criteria for diagnosing VaD	Damage to blood vessels in the affected area of the brain causes infarction (tissue death).	Memory impairment, apathy, depression, impaired communication, disorientation, confusion, behavioural changes, sleep disturbances, visual hallucinations, poor balance and slowness.

Frontotemporal Dementia (FTD)	1.1-5 (Hogan *at al.*, 2016)	Neurofibrillary tangles of Tau (NFT) TAR DNA-binding protein 43 (TDP-43) Fused sarcoma protein (FUS)	Deposits of NFT, TDP-43, or FUS in the frontotemporal region cause lobar degeneration. TDP-43 used to distinguish between three subtypes of FTD (subtypes FTD-TDP type 1-3) depending on where it accumulates in the brain: type I) TDP-43 is found in dystrophic neurites (DN), neuronal cytoplasmic inclusions (NCI), intra-neuronal inclusions (NII) in the frontal and temporal cortex; type 2) TDP-43 is found in the DN in the lower layers of cerebral cortex; and type 3) TDP-43 is found in the NCI in the cerebral cortex and hippocampus.	Changes to personality and behaviour, difficulties with speech and understanding language. FTD-TDP type 1 clinically presents as either the behavioural variant (where behavioural changes occur) of frontotemporal dementia or progressive non-fluent aphasia (where a person's ability to use language is affected). FTD-TDP type 2 clinically presents as Semantic dementia (characterised by loss of semantic memory in both the verbal and non-verbal domains). FTD-TDP type 3 clinically presents as motor neuron disease.

1.1.2 Methods used to diagnose dementia

Several methods are used assess dementia symptoms but a *definitive* diagnosis of dementia is only possible via neuropathological assessment post-mortem (Love, 2005)(see **Section 1.1.4**). To clinically define dementia the first step in the diagnostic process after presentation to services is generally informant questionnaire based which helps to identify individuals with mild cognitive impairment (MCI; the stage between cognitive decline of normal aging and dementia) and dementia (Galvin *at al.*, 2005; Jorm & Jacomb, 1989; Sabbagh *at al.*, 2010). Informant questionnaires have shown good accuracy in detecting dementia and correlate with cognitive screening tests used in the diagnosis of AD. Cognitive testing is used to assess decline in general cognition beyond that of normal aging. The Mini-Mental State Examination (MMSE) (Folstein *at al.*, 1983) is frequently used at an initial assessment and involves individuals answering a 30-point questionnaire which tests numerous cognitive domains including memory, attention and language. The strength of this test is the increase in sensitivity with repeat assessments, therefore it can be used to track the progression of cognitive decline (Borson, Scanlan, Watanabe, Tu, & Lessig, 2005). Other cognitive tests used in the diagnosis of AD include the Mini-Cog (Borson *at al.*, 2005) a short three item recall test; the Montreal Cognitive Assessment (MoCa) (Nasreddine *at al.*, 2005), a short test which assesses attention, numeracy, executive functions, conceptual thinking, memory, language, visual perception/ construction and orientation; and the Clinical Dementia Rating (CDR) (Morris, 1993), a 5-point scale test used to characterize six domains of cognitive and functional performance which are indicative of dementia (memory, judgment / problem solving, orientation, community affairs, hobbies and life at home, and lastly personal hygiene/ care).

In addition to informant questionnaires and cognitive assessments ‘aggregated risk’ is considered when diagnosing an individual with dementia as epidemiological studies suggest many factors increase the risk for developing dementia including certain health conditions and lifestyle factors (Silva *at al.*, 2019). Aggregated risk analysis involves identifying if an individual has any medical conditions which increases their risk of developing dementia (e.g. hypertension (Skoog *at al.*, 1996; Staessen, Richart, & Birkenhäger, 2007), diabetes (Li, Song, & Leng, 2015), obesity (Fitzpatrick *at al.*, 2009), high cholesterol (Popp *at al.*, 2013), head trauma (Mortimer *at al.*, 1991) and cardiovascular disease (Fitzpatrick *at al.*, 2009)) in addition to considering their demographic information (e.g. age (Guerreiro & Bras, 2015) and family history (Honea, Vidoni, Swerdlow, Burns, & Alzheimer's Disease Neuroimaging Initiative, 2012)).

Neuroimaging and physical examinations can be used to detect focal neurologic deficits affecting the CNS and gait disturbances (Chen, Wang, Liou, & Shaw, 2013; Weingarten, Sundman, Hickey, & Chen, 2015). These exams help identify cerebrovascular disease, Parkinsonism or if hydrocephalus is present (the build-up of fluid in the ventricles within the brain). Additionally, there are laboratory tests which can be used to aid the diagnosis of dementia. For example, to detect deficiencies which are known to be associated with developing dementia such as B_{12}, vitamin D and thyroid stimulating hormone (Chai *at al.*, 2019; Choi *at al.*, 2020; Wang *at al.*, 2001). Nevertheless, these deficiencies are non-specific and can be identified in all dementia sub-types, limiting the usefulness of these tests.

1.1.2.1 Pre-mortem biomarkers of Alzheimer's disease

Due to the inaccessibility of the brain, biomarkers are the best method to investigate the biology of Alzheimer's pathology pre-mortem (for a discussion of AD pathology see section **1.1.4.1**). In addition to their clinical phenotype criterion, the International Working Group (IWG) (Dubois *at al.*, 2014) and the National Institute on Aging and Alzheimer's Association (NIA-AA) (Montine *at al.*, 2012) have both incorporated biomarkers into their diagnostic processes for detecting AD. Biomarkers for AD include: 1) structural magnetic resonance imaging (MRI) and fluorodeoxyglucose (FDG) positron emission tomography (PET) to detect neurodegeneration; 2) molecular neuroimaging with PET to detect amyloid and tau; and 3) CSF analysis of amyloid-beta (Aβ) or neurofibrillary tangles of tau (NFT). The incorporation of biomarkers in the diagnostic criteria makes it possible for AD to be diagnosed in the prodromal stage and can confirm the type of dementia and how far disease has progressed. Despite the increased use of biomarkers in AD trials - which can help with confidence in diagnosis - there is still risk of over-/misdiagnosis using these methods. To date, the only reliable way to validate these techniques and confirm AD is through post-mortem neuropathology examinations (Love, 2005).

1.1.3 Familial versus sporadic AD

There are three forms of Alzheimer's disease (Bekris, Yu, Bird, & Tsuang, 2010). There are two early forms which affect individuals under the age of 65: sporadic early onset AD (sEOAD) and familial EOAD (fEOAD), and late onset / sporadic AD (LOAD), affecting individuals over the age of 65 (Koedam *at al.*, 2010). EOAD is less common than LOAD, accounting for ~5% of AD cases (Koedam *at al.*, 2010), with 90% of these cases being attributed to autosomal recessive mutations (Wingo, Lah, Levey, & Cutler, 2012). fEOAD only accounts for 1% of AD cases (Bekris *at al.*, 2010) and it has a clear inheritance pattern; it is caused by autosomal dominant mutations in one of three genes: the amyloid precursor protein gene (*APP*; located on chromosome 21), the presenilin 1 gene (*PSEN1*; located on chromosome 14) and the presenilin 2 gene (*PSEN2*; located on chromosome 1) (Bekris *at al.*, 2010). LOAD accounts for 95% of AD cases and is a complex disease; risk is mediated by a combination of genetic, lifestyle and environmental factors (Koedam *at al.*, 2010).

Although there are clear differences in the age of onset between EOAD and LOAD, both are characterised by deficits in episodic, short-term and working memory, and the neuropathological hallmarks are analogous (Joubert *at al.*, 2016). However, studies have shown that individuals with EOAD have larger deficits in executive functioning, language and visual perception abilities, whereas individuals with LOAD have a more dispersed pattern of cognitive impairment including greater semantic (long-term) memory deficits (Aziz *at al.*, 2017; Joubert *at al.*, 2016). Additionally, in EOAD atrophy is widespread and includes frontotemporoparietal areas and in LOAD it is limited to temporal regions (Aziz *at al.*, 2017).

1.1.4 Neuropathology of neurodegenerative diseases

Most neurodegenerative diseases are proteinopathies; their pathogenesis is characterised by the aggregation of specific proteins in intracellular inclusions or extracellular aggregates within the brain (Ross & Poirier, 2004). Although the proteins and the areas involved in the aetiology of neurodegenerative diseases differ (see **Table 1.1**), the progressive accumulation of these deposits ultimately leads to neuronal cell death and brain atrophy (Ross & Poirier, 2004). Of note, AD and other dementias are rarely found without other neurodegenerative co-pathologies. A major challenge in understanding development of these disorders relates to clinical and neuropathological heterogeneity (Dickerson, Brickhouse, McGinnis, & Wolk, 2017); the concordance of clinically defined AD and neuropathologically defined AD ranges from 70-92% (Gauthreaux *at al.*, 2020; Lim *at al.*, 1999; Nagy *at al.*, 1998; Petrovitch *at al.*, 2001; Tasaki, Gaiteri, Mostafavi, De Jager, & Bennett, 2018).

1.1.4.1 Neuropathology of Alzheimer's disease

Pathology is thought to initiate up to two decades before clinical symptoms manifest, and by the time an individual is symptomatic there is usually significant neurodegeneration (Rajan, Wilson, Weuve, Barnes, & Evans, 2015). The only way to accurately determine the severity of AD is through post-mortem examination (Love, 2005)

AD is a progressive neurodegenerative disorder which causes brain atrophy in the limbic regions and neocortex (DeTure & Dickson, 2019). AD is characterised by two histopathological hallmarks: the accumulation of extracellular Aβ plaques and accretion of intracellular NFTs (Braak, Alafuzoff, Arzberger, Kretzschmar, & Del Tredici, 2006; Thal, Rüb, Orantes, & Braak, 2002). There are different measures used to quantify neuropathology in AD including Braak NFT staging (Braak *at al.*, 2002), a measure of NFT pathology; Thal Phasing (Thal *at al.*, 2006), a measure of amyloid deposits (both diffuse and dense-core); and the 'Consortium to Establish a Registry for AD' (CERAD) score (Mirra *at al.*, 1991),which describes the density of neuritic amyloid plaques (dense-core) .

The Braak NFT scale of AD neuropathology spans seven stages (0-VI) and measures both the quantity and regional locality of NFTs throughout the brain (Braak, Alafuzoff, Arzberger, Kretzschmar, & Del Tredici, 2006) (see **Figure 1.1**). Of note, the cerebellum is mostly unaffected by NFTs in AD, even at the latest stages of the disease.

- Braak NFT stages I-II - the entorhinal stage: tau tangles are confined to the trans-entorhinal layer and the earliest site of AD NFT pathology is the Entorhinal Cortex (EC).
- Braak NFT stages III/IV - the limbic stage: AD is clinically emerging and NFT starts to spread to the hippocampus, the area of the brain involved in memory processing.

- Braak NFT stages V/VI - the neocortical stage: represents fully developed AD, with NFT accumulation through all areas of the cortex.

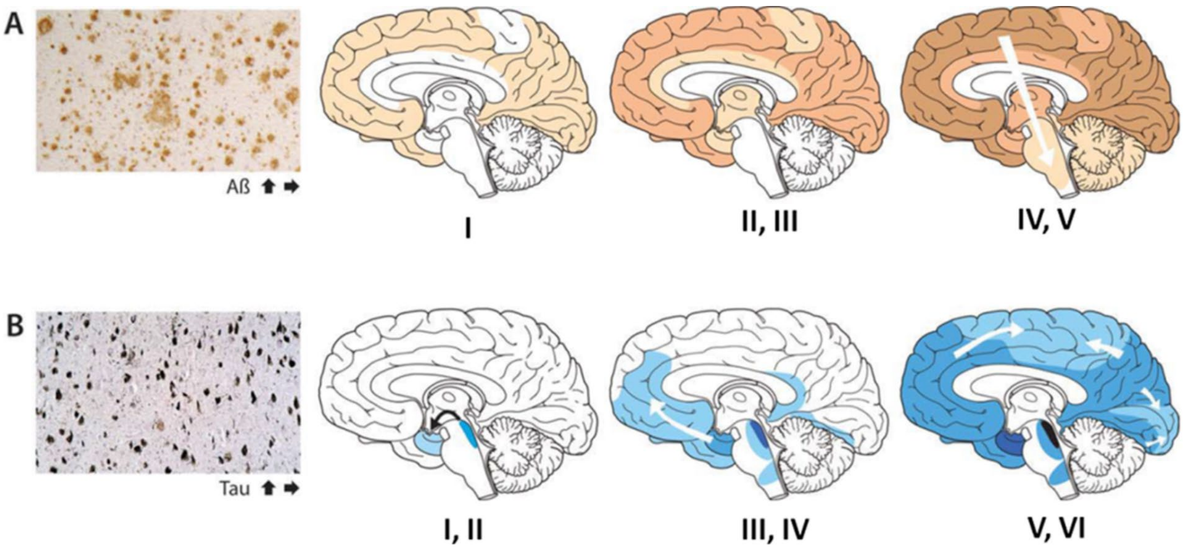


Figure 1.1: The progression of the pathological hallmarks of AD. (A) *Thal amyloid phases – amyloid deposit progression* ***(B)*** *Braak NFT Tangle stages - NFT progression. Figure adapted from (Jouanne, Rault, & Voisin-Chiret, 2017).*

Similarly to NFT accumulation, Aβ deposits spread throughout different regions of the brain in a progressive but predictable manner. Thal phasing measures the spatial-temporal distribution of amyloid deposits in the encephalon and is divided into six phases (0-5) (Thal, Rüb, Orantes, & Braak, 2002)(see **Figure 1.1**):

1. Thal phase 1 - Aβ deposits are found exclusively in the neocortex.
2. Thal phase 2 - Aβ spreads throughout the allocortical regions (e.g. the hippocampus)
3. Thal Phase 3 - Aβ starts to accumulate in the striatum.
4. Thal phase 4 - several brainstem nuclei become involved.
5. Thal phase 5 - the presence of Aβ deposits in the cerebellum and other brain areas.

CERAD neuropathologists use a semi-quantitative approach to assess the frequency of senile plaques (neuritic). CERAD density refers to the abundance of dense-core neuritic plaques in three areas of the isocortex (frontal, temporal and parietal) and is measured on a four tier scale (Mirra *at al.*, 1991):

1. No pathology

2. Sparse pathology
3. Moderate pathology
4. Frequent or high pathology

The NIA-AA developed an ABC scoring system which incorporates all three measures to quantify the extent of AD neuropathology where Thal amyloid phasing = A; Braak NFT staging = B, and CERAD neuritic amyloid plaque density = C (see **Table 1.2**) (Montine *at al.*, 2012). Of note, the presence of NFTs is essential for an AD diagnosis. Two amyloid measures are included in the scores as it is unknown which amyloid marker is more informative in regards to either clinical or pathological measures. According to the ABC scoring system, for an individual to be classified with either intermediate or high AD neuropathology, they can have a low Thal phase, but must have a Braak stage greater than III and a CERAD score in the moderate to frequent range (see **Table 1.2**) (Cummings, 2019; Hyman *at al.*, 2012).

Table 1.2: The NIA-AA ABC scoring system. *AD neuropathological progression is assessed using an "ABC" score from three scales:* ***(A)****, Aβ plaques by the method of Thal phase;* ***(B)*** *NFT stage by the method of Braak; and* ***(C)*** *neuritic plaque score by the method of CERAD. The combination of A, B, and C scores receive a descriptor of "Not", "Low", Intermediate" or "High" AD neuropathological change. "Intermediate" or "High" AD neuropathological change is considered sufficient explanation for dementia. Figure and legend taken directly from (Hyman at al., 2012).*

A: Aβ/amyloid plaque score (Thal phases)[2]	**C: Neuritic plaque score (CERAD)**	**B: NFT score (Braak stage)**		
		B0 or B1 (None or I/II)	**B2 (III/IV)**	**B3 (V/VI)**
A0 (0)	**C0** (none)	Not	Not	Not
A1 (1/2)	**C0 or C1** (none to sparse)	Low	Low	Low
	C2 or C3 (mod. to freq.)	Low	Intermediate	Intermediate
A2 (3)	**Any C**	Low	Intermediate	Intermediate
A3 (4/5)	**C0 or C1** (none to sparse)	Low	Intermediate	Intermediate
	C2 or C3 (mod. to freq.)	Low	Intermediate	High

The presence of NFTs continues to rise in patients with mild cognitive decline and starts to plateau when they reach dementia status, whereas Aβ plaque levels plateau before an individual is symptomatic, suggesting NFTs are correlated more strongly with disease symptoms and neurodegeneration than Aβ (Jack *at al.*, 2010).

1.1.5 The aetiology of Alzheimer’s disease

The neuropathology of AD is well characterised but the mechanisms of disease onset and progression are not fully understood. There are two main hypotheses which have been posited to explain the aetiology of AD disease pathology: 1) the amyloid hypothesis, which involved the accumulation of Aβ (Hardy & Allsop, 1991); and 2) the tau hypothesis, which involves the accumulation of NFTs (Kosik, Joachim, & Selkoe, 1986). Although both amyloid and tau pathology are involved in the aetiology of AD, the amyloid hypothesis has been the mainstream concept driving AD research for the past 20 years. However, due to failures in clinical trials targeting Aβ, recent studies propose that tau is the main factor underling the progression of AD (Kametani & Hasegawa, 2018).

1.1.5.1 Amyloid cascade hypothesis

The amyloid cascade hypothesis focuses on the aggregation of Aβ (see **Figure 1.2**), which is largely driven by known familial risk mutations such as those in *APP*. Aβ aggregation can also be triggered in sporadic AD as consequence of dysfunctional Aβ clearing mechanisms as a result of *APOE*-ε4 inheritance or faulty Aβ degradation (see **Figure 1.3**) (Hardy & Selkoe, 2002; Selkoe & Hardy, 2016). Aβ is a ~40-base peptide which is formed by the cleavage of *APP* (a transmembrane protein involved in neuronal development and axonal transport) by the enzymes β-secretase (*BACE1*) and γ-secretase. This process produces Aβ40 and Aβ42 segments in a 9:1 ratio (Hardy & Selkoe, 2002). Although the Aβ42 isoform is the least abundant of the Aβ peptides, it is the more pathogenic form in AD as it makes up the majority of Aβ plaques in the brain. On the other hand, Aβ40 is generally found in cerebrovascular plaques, where amyloid build up on the walls of the arteries in the brain. In unaffected individuals Aβ is eliminated from *APP* by *BACE1* and subsequently γ-secretase is released outside of the cell where it will degrade. In older individuals or where there are pathological conditions, the ability to degrade Aβ decreases, leading to the accumulation of Aβ peptides. The increase in Aβ peptides then induces Aβ amyloid fibril formations which aggregate and develop into senile plaques (see **Figure 1.2**). The senile plaques create a neurotoxic environment in the brain and induce tau neuropathology, ultimately causing neuronal cell death (Hardy & Allsop, 1991).

Although overproduction of Aβ is primarily responsible for fEOAD, studies suggest that there is a 30% impairment in the clearance of both Aβ42 and Aβ40 in sporadic AD, highlighting that Aβ is likely important in the development of both familial and sporadic forms of the disease (Mawuenyega *at al.*, 2010). In sporadic AD, Aβ clearance can be altered as a result of *APOE* since the rate of clearance depends on the different isoforms; the rate is fastest for *APOE*-ε2 and slowest for *APOE*-ε4 (Deane *at al.*, 2008). In addition, mechanistic studies have linked several LOAD-associated GWAS risk genes (including *SORL1*, *BIN1* and *PICALM*) to aspects of Aβ homeostasis, providing further support for the role of Aβ in LOAD pathogenesis (Selkoe & Hardy, 2016).

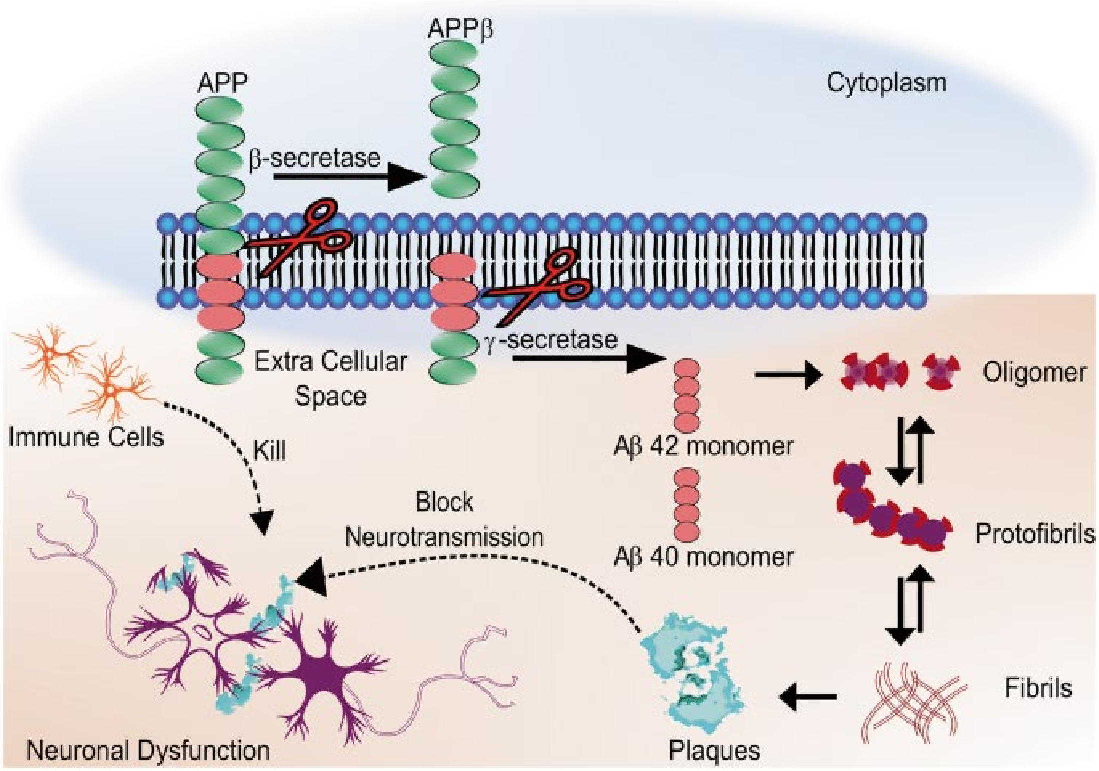


Figure 1.2: The amyloid cascade hypothesis. *This figure shows how amyloid-beta plaques form during the progression of Alzheimer's disease, leading to neurodegeneration. Figure taken directly from (Pandey & Ramakrishnan, 2020).*

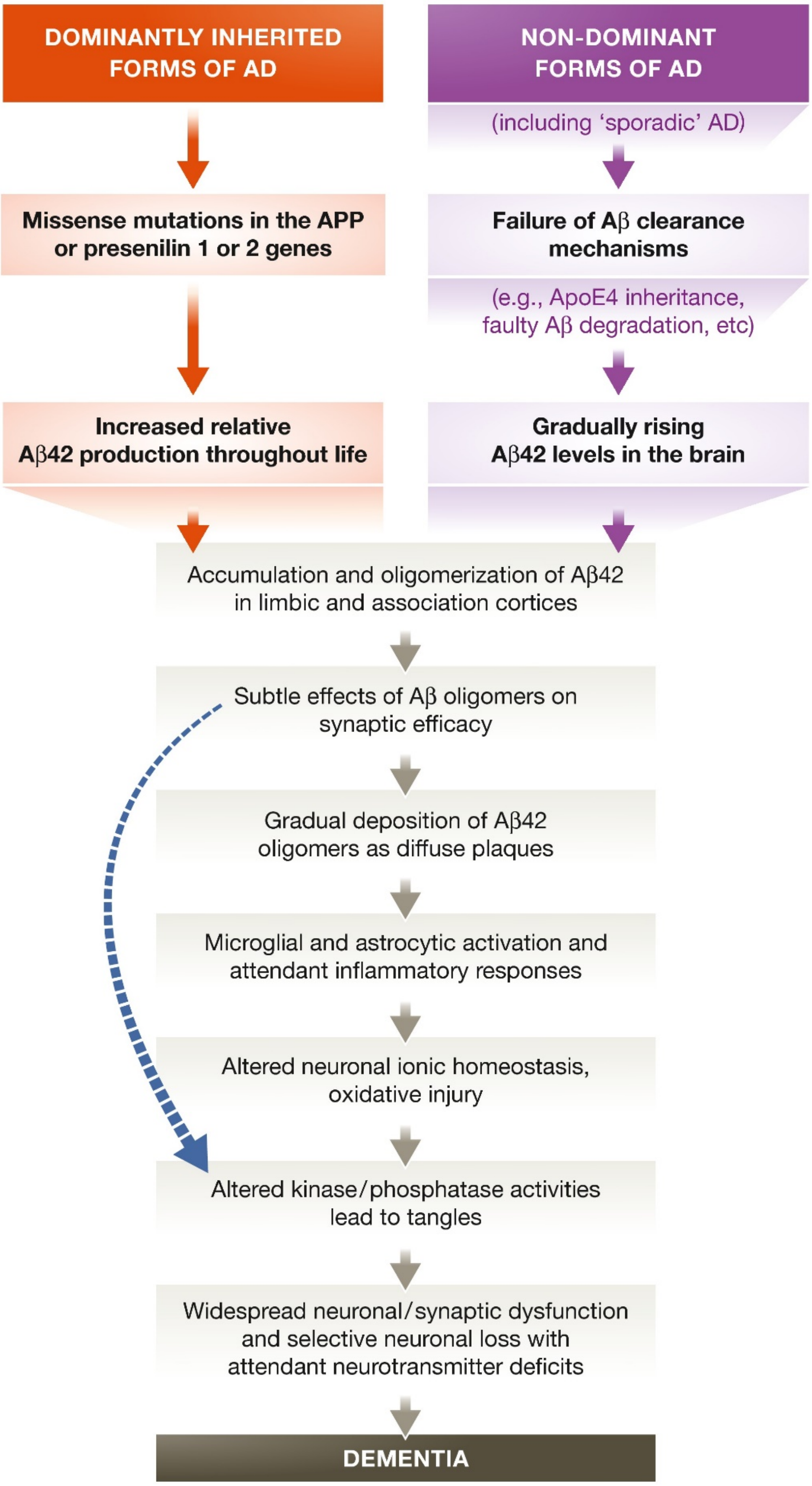

DOMINANTLY INHERITED FORMS OF AD
NON-DOMINANT FORMS OF AD
(including 'sporadic' AD)
Missense mutations in the APP or presenilin 1 or 2 genes
Failure of Aβ clearance mechanisms
(e.g., ApoE4 inheritance, faulty Aβ degradation, etc)
Increased relative Aβ42 production throughout life
Gradually rising Aβ42 levels in the brain
Accumulation and oligomerization of Aβ42 in limbic and association cortices
Subtle effects of Aβ oligomers on synaptic efficacy
Gradual deposition of Aβ42 oligomers as diffuse plaques
Microglial and astrocytic activation and attendant inflammatory responses
Altered neuronal ionic homeostasis, oxidative injury
Altered kinase/phosphatase activities lead to tangles
Widespread neuronal/synaptic dysfunction and selective neuronal loss with attendant neurotransmitter deficits
DEMENTIA

Figure 1.3: The sequence of major pathogenic events leading to AD proposed by the amyloid cascade hypothesis. *The curved blue arrow indicates that Aβ oligomers may directly damage the synapses and neurites of brain neurons, in addition to activating microglia and astrocytes. Figure and legend taken directly from (Selkoe & Hardy, 2016).*

1.1.5.2 Tau hypothesis

Tau is a microtubule-associated protein which is involved in regulating stability of tubulin assembly and the maintenance of neuronal structures (Strang, Golde, & Giasson, 2019). The tau gene (*MAPT*) is located on chromosome 17 with two tau isoforms (3R and 4R) being expressed in the adult human brain. The tau hypothesis posits that the primary causal substance leading to AD is tau (Kosik, Joachim, & Selkoe, 1986). In individuals unaffected by pathology, the tau-protein is actively phosphorylated in order to regulate neuronal axon length. Within AD brains 3R and 4R tau aggregates are found in a hyper-phosphorylated state as pathological inclusions (see **Figure 1.4**). The pathological inclusions are referred to as NFTs when found in the neuronal cell bodies or threads if found in dendrites and axons. It is hypothesised that these pathological inclusions cause the degeneration of neurons (Kosik, Joachim, & Selkoe, 1986). Tau pathology is also present in other neurodegenerative disorders including forms of FTD and PD dementias which are linked to chromosome 17 (e.g. FTD-17) (Strang, Golde, & Giasson, 2019).

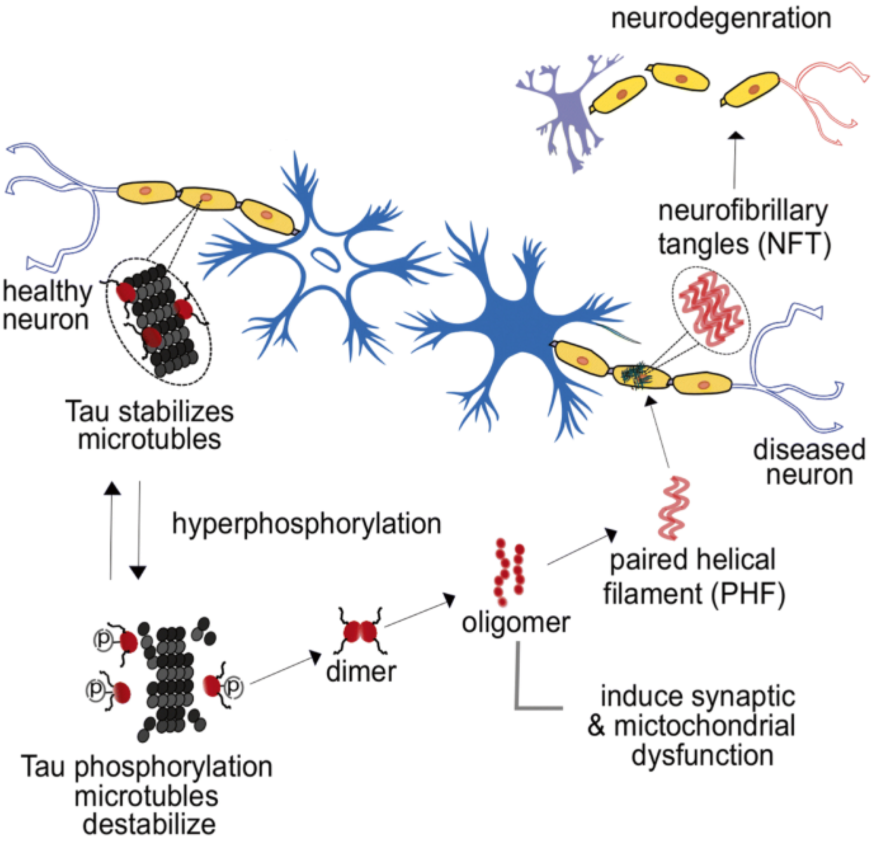

neurodegenration
neurofibrillary tangles (NFT)
healthy neuron
Tau stabilizes microtubles
diseased neuron
hyperphosphorylation
paired helical filament (PHF)
oligomer
dimer
P
induce synaptic & mictochondrial dysfunction
Tau phosphorylation microtubles destabilize

Figure 1.4: The Tau Hypothesis. *This figure shows how neurofibrillary tangles of tau form during the progression of Alzheimer's disease, leading to neurodegeneration. Figure taken directly from (Pandey & Ramakrishnan, 2020).*

1.1.6 Neuropathology of other neurodegenerative diseases

1.1.6.1 Lewy-body related Neuropathology

Lewy-body (LB) pathology is involved in the pathogenesis of DLB and PD and involves the accumulation and of α-synuclein in neuronal cell bodies as Lewy bodies (LBs) and in neuronal cell processes (e.g. axons) as Lewy neurites. LB pathology is quantified by Braak LB staging, which provides a measure for α-synuclein throughout the brain. Similarly to Braak NFT staging, there are seven Braak LB stages (0-6):

1. Braak LB stage 1: α-synuclein starts to accumulate in the motor nucleus of the medulla oblongata.
2. Braak LB stage 2: α-synuclein spreads to the locus coeruleus.
3. Braak LB stage 3: it progresses to the substantia nigra of the midbrain.
4. Braak LB stage 4: it spreads to the trans-entorhinal region and CA2 of the hippocampus.
5. Braak stage 5: the neocortex is affected.
6. Braak LB stage 6: α-synuclein has spread through all of the neocortex and is detected in the premotor and motor regions.

The exact mechanism(s) which cause α-synuclein to misfold and form pathogenic inclusions are not fully understood, however once they are formed there is usually spread of pathology between cells, leading to neurotoxicity and cell death. Of note, LB pathology is not limited to LB dementias and is frequently found in AD cases (Hamilton, 2006).

1.1.6.2 Neuropathology of TDP-43 proteinopathies

FTD and amyotrophic lateral sclerosis (ALS) are primary TAR DNA-binding protein 43 (TDP-43) proteinopathies, where TDP-43 is the main driver in disease pathogenesis (Chou *at al.*, 2018). When TPD-43 is not the primary driver of pathogenesis and occurs in association with other distinct pathological processes, these diseases are referred to as secondary TDP-43 proteinopathies, for example AD cases are often found with some level of TDP-43 (Josephs *at al.*, 2014). TDP-43 is a specific mark used in the characterisation of subtypes of neurological disorders, particularly FTD. TDP-43 can be used to distinguish between three subtypes of FTD depending on where it accumulates in the brain (Mackenzie, Rademakers, & Neumann, 2010). In FTD-TDP type 1 (which clinically presents as either the behavioural variant of frontotemporal dementia or progressive non-fluent aphasia), TDP-43 is found in dystrophic neurites (DN), neuronal cytoplasmic inclusions (NCI), intra-neuronal inclusions (NII) in the frontal and temporal cortex. In FTD-TDP type 2 (which clinically presents as semantic dementia), TDP-43 is found in the DN in the lower layers of cerebral cortex. In FTD-TDP type 3 (which clinically presents as motor neuron disease), TDP-43 is found in the NCI in the cerebral cortex and hippocampus. ~40% of AD cases accumulate TDP-43 in the DN and NCI in the hippocampus and amygdala (Mackenzie, Rademakers, & Neumann, 2010).

1.2 Genetics of Alzheimer's disease

The identification of the familial AD genes *APP* and the presinilins (*PSEN1* and *PSEN2*) was a pivotal point for understanding fEOAD. Unlike fEOAD, which has a heritability estimate of 92-100% (Wingo, Lah, Levey, & Cutler, 2012), LOAD does not have a clear Mendelian inheritance pattern. Using twin studies (an experimental design used to quantifying genetic and environmental influences based on comparing the concordance of a particular phenotype between monozygotic and dizygotic twins) the heritability of LOAD is estimated to be between 56-79% (Gatz *at al.*, 2006), suggesting there is a large genetic component contributing to disease aetiology.

The first LOAD risk gene discovered was Apolipoprotein E (*APOE*), which is found on chromosome 19 and remains the strongest risk gene for LOAD (van der Lee *at al.*, 2018). The three main alleles of *APOE* (ε2, ε3 and ε4) are generated by variants in two single nucleotide polymorphisms (SNPs): rs429358 and rs7412 (Farrer, 1997). Individuals with one ε4 allele (ε4 heterozygotes) are four times as likely to develop AD compared to the average risk, and individuals with two ε4 alleles (ε4 homozygotes) are 15 times more likely to develop AD (Farrer, 1997). In contrast, the rarer ε2 allele has a protective effect (Farrer, 1997). Although the mechanisms by which alleles of the *APOE* gene alter risk for AD are not fully understood, it is hypothesised that conformational changes to the shape of *APOE* decreases the protein's ability to bind ligands (e.g. Aβ and TREM) (Holtzman, Herz, & Bu, 2012). In addition, the ε4 allele is less efficient in mediating the clearance of Aβ, leading to its aggregation and in turn, neurodegeneration (Holtzman, Herz, & Bu, 2012).

APOE genotype explains ~13% of variation in LOAD (Ridge *at al.*, 2016), meaning much of the inherited component of LOAD must be attributable to other variants. Over the past decade, there has been a focus on uncovering the additional genetic variants that contribute to LOAD using genome-wide association studies (GWAS). GWAS use an experimental design which identifies common genetic variation at the nucleotide level (Wang, Barratt, Clayton, & Todd, 2005). GWAS uses a hypothesis free approach to compare the frequency of SNPs at specific loci with polygenic traits (Wang, Barratt, Clayton, & Todd, 2005).

SNPs associated with traits identified by GWAS are either the causal variant or are in linkage disequilibrium (LD) with the causal variant. LD describes the extent to which an allele of one SNP is inherited or correlated with an allele of another SNP within a given population (Bush & Moore, 2012). SNPs in LD do not always reach genome-wide significance (p<5e-08) but can be utilised to identify a causal SNP; they can associate with a trait at a higher p-value. GWAS have identified thousands of SNPs associated with complex brain disorders including autism, schizophrenia and dementia (Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium, 2017; Ferrari *at al.*, 2014; Guerreiro *at al.*, 2018; Jansen *at al.*, 2019; Kunkle *at al.*, 2019; Nalls *at al.*, 2019; Sullivan, Daly, & O'Donovan, 2012; van Rheenen *at al.*, 2016; Visscher *at al.*, 2017). GWAS have shown that LOAD and other non-familial neurodegenerative diseases are polygenic disorders caused by multiple variants of small effect (Ferrari *at al.*, 2014; Guerreiro *at al.*, 2018; Jansen *at al.*, 2019; Kunkle *at al.*, 2019; Nalls *at al.*, 2019), although the precise mechanisms leading to disease are unclear. Each trait-associated SNP generally has a very small effect size, and given the heterogeneity of brain disorders, sample size is a limiting factor in GWAS; in order to increase statistical power, studies have focussed on increasing sample sizes by conducting meta-analyses. A meta-analysis involves merging the summary statistics of individual studies.

The first GWAS of clinically diagnosed LOAD were conducted in 2009 (Harold *at al.*, 2009; J.-C. Lambert *at al.*, 2009) with samples sizes totalling ~15,000 (~30% cases) across both stages of their analyses (see **Table 1.3**). In addition to *APOE* these GWAS identified variants annotated to *CLU* (a multifunctional glycoprotein involved in lipid transport and immune modulation which is thought to be involved in altering Aβ aggregation and clearance), *PICALM* (an endocytic-related protein which is required for the formation and maturation of autophagic precursors and has been found to influence the processing of *APP*) and *CR1* (a membrane glycoprotein that functions in the innate immune system and promotes phagocytosis of immune complexes and cellular debris, as well as Aβ) (Harold *at al.*, 2009; J.-C. Lambert *at al.*, 2009). Since 2009, additional LOAD GWAS have been conducted (Bellenguez *at al.*, 2020; de Rojas *at al.*, 2020; Hollingworth *at al.*, 2011; Jansen *at al.*, 2019; Kunkle *at al.*, 2019; Lambert *at al.*, 2013; Marioni *at al.*, 2018; Naj *at al.*, 2011; Sims *at al.*, 2017), collectively identifying >75 LOAD loci; these successes can be attributed to international collaborative efforts which have focused on increasing sample size and power by meta-analysing the summary statistics from multiple cohorts (see **Table 1.3**; **Figure 1.5**). One method which has been adopted across several AD GWAS studies to increase power has been to use a 'proxy' measure of AD based on family history. This method was first adopted by Marioni and Colleagues (Marioni *at al.*, 2018). They used the self-report questions "Has/did your father ever suffer from Alzheimer's disease/dementia?" and "Has/did your mother ever suffer from Alzheimer's disease/dementia?" to derive a proxy of AD. They excluded participants with parents under the age of 60 or parents who died before the age of 60. They identified 26 AD risk loci and seven of these were novel. Genetic correlation analysis showed the proxy measure of AD was accurate for clinical diagnosis for both maternal (r=0.91) and paternal (r=0.67) AD when correlated with a clinical AD case-control GWAS. Due to the high genetic correlations between proxy AD and clinical AD, several other studies have utilised the AD-by proxy methodology (Bellenguez *at al.*, 2020; de Rojas *at al.*, 2020; Jansen *at al.*, 2019; Wightman *at al.*, 2020).

Table 1.3: GWAS of LOAD over the past decade

Reference	Approach	Sample size		Total number of loci
		Total	**Number of Cases**	
(Lambert *at al.*, 2009)	Diagnosed case-control GWAS	Stage 1: 7,360 Stage 2:7,275	Stage 1: 2,032 Stage 2: 3,978	3
(Harold *at al.*, 2009)	Diagnosed case-control GWAS	Stage 1: 11,789 Stage 2:4,363	Stage 1: 3,941 Stage 2: 2,023	3
(Hollingworth *at al.*, 2011)	Diagnosed case-control GWAS	Stage 1: 20,373 Stage 2: 9,799 Stage 3: 29,544	Stage 1: 6,688 Stage 2: 4,896 Stage 3:8,286	6
(Naj *at al.*, 2011)	Diagnosed case-control GWAS	Stage 1: 15,675 Stage 2:7,096	Stage 1: 8,309 Stage 2: 3,531	9
(Lambert *at al.*, 2013)	Diagnosed case-control GWAS	Stage 1:54,162	Stage 1: 17,008	20
		Stage 2: 74,046	Stage 2: 8,572	
(Sims *at al.*, 2017)	Diagnosed case-control GWAS of rare variants using exome-wide arrays	85,133	37022	3
(Marioni *at al.*, 2018)	Diagnosed and proxy case-control GWAS	388,324 (314,278 [81%] proxy)	67,314 (42,034 [62%] proxy)	26
(Jansen *at al.*, 2019)	Diagnosed and proxy case-control GWAS	455,258 (383,378 [84%] proxy)	71,880 (46,613 [65%] proxy)	29
(Kunkle *at al.*, 2019)	Diagnosed case-control GWAS	94,437	33,814	25
(de Rojas *at al.*, 2020)	Diagnosed and proxy case-control GWAS	409,435	50,737 (14,338 [28%] proxy)	40
(Wightman *at al.*, 2020)	Diagnosed and proxy case-control GWAS	1,126,563	90,338 (46,613 [52%] proxy)	38
(Bellenguez *at al.*, 2020)	Diagnosed and proxy case-control GWAS	487,511	85,934 (46,828 [54%] proxy)	75
(Schwartzentruber *at al.*, 2021)	Diagnosed and proxy case-control GWAS	408,942 + Stage 1 Kunkle *at al.* (n = 63,926)	53,042 (52,791 [99%] proxy) + Stage 1 Kunkle *at al.* (n = 21,982)	37

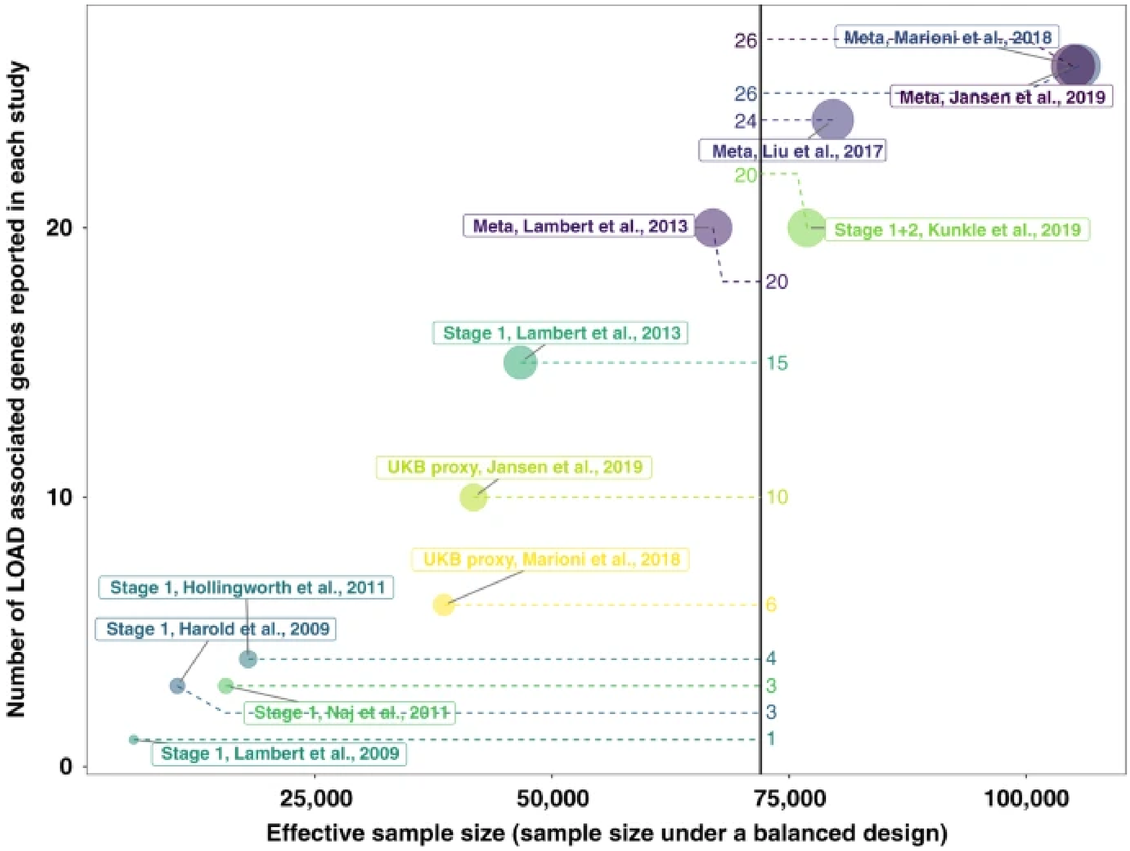


Figure 1.5: The relationship between sample size and the number of identified genes. *Sample size is calculated as the number of cases and controls under a 50:50 ratio. Genes associated with LOAD were collected from various GWAS. The genes are the closest to the SNPs (minor allele frequency > 0.01 and p<5e-08). Stage 1 = stage 1 from the study. Meta = meta-analysis from the study. Figure taken directly from (Zhang at al., 2020).*

Two of the most recent AD GWAS with publicly-available summary statistics were conducted by Kunkle and colleagues (2019) and Jansen and colleagues (2019), and the results from these studies have been used throughout my thesis. The Kunkle *at al.* (2019) GWAS included clinically and autopsy-documented LOAD cases (35,274 cases and 59,163 controls). Kunkle and colleagues conducted a GWAS meta-analysis of non-Hispanic whites from the International Genomics of Alzheimer's Project (IGAP) (which is composed of four consortia: Alzheimer Disease Genetics Consortium (ADGC), Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (CHARGE), The European Alzheimer's Disease Initiative (EADI), and Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease Consortium (GERAD/PERADES). 25 risk loci were identified (see **Figure 1.6**) The Jansen *at al.* (2019) GWAS was based both on clinically diagnosed AD and AD-by proxy (based on family history) cases (71,880 cases, 383,378 controls). Jansen and colleagues conducted a GWAS meta-analysis using three independent AD-case-control consortia (Alzheimer's disease working group of the Psychiatric Genomics Consortium (PGC-ALZ), the International Genomics of Alzheimer's Project (IGAP), and the Alzheimer's disease Sequencing Project (ADSP)). These cohorts were meta-analysed with the UK Biobank which used parental data as a proxy for LOAD status. The number of parents with AD was weighted by the probability of being a case or control based on parental age. 29 risk loci were identified (see **Figure 1.6**). There is a high correlation between the two studies (r=0.89; see **Figure 5.6**)

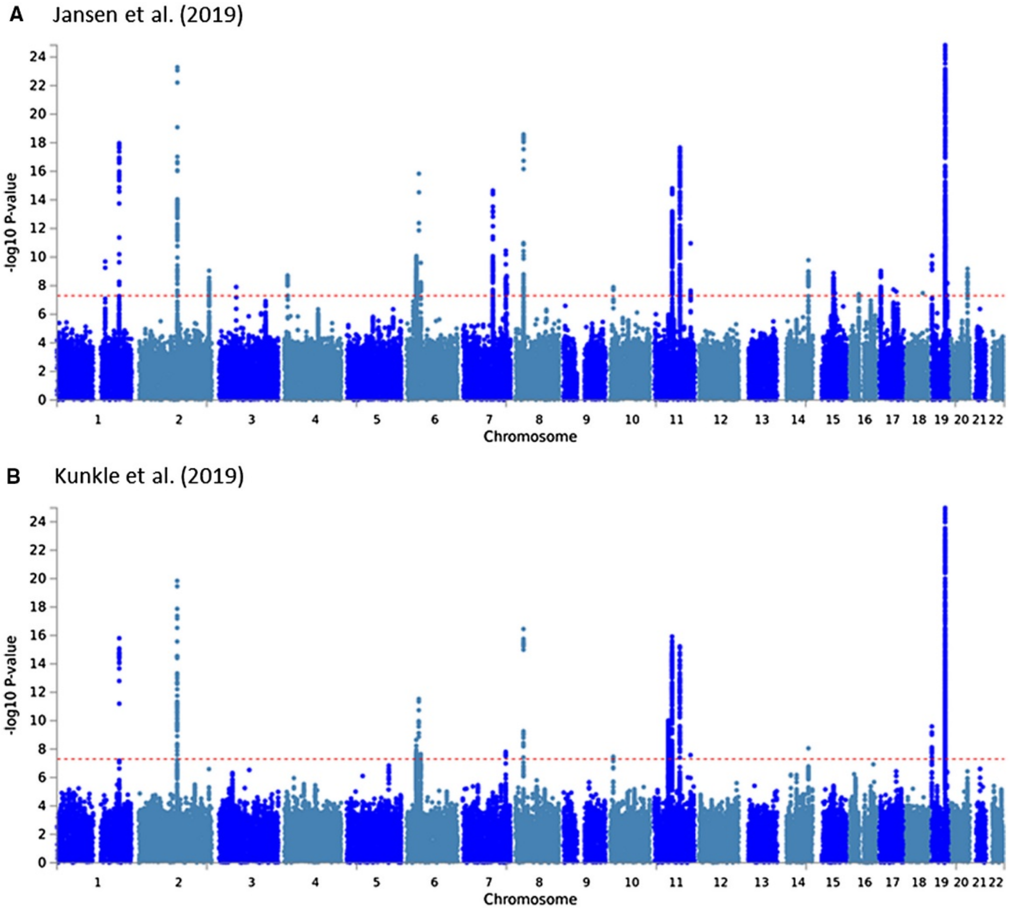


Figure 1.6: Manhattan plots of the two latest publicly available LOAD GWAS. ***(A)*** *GWAS was conducted by Jansen and colleagues (2019); and* ***(B)*** *GWAS was conducted by Kunkle and colleagues (2019). The x axis is the genomic position, segregated by chromosome. Shown on the y-axis is the –log10 p-value from each GWAS. Each point represents a SNP. The red horizontal line represents the genome-wide significance (p<5e-08). P-values are truncated at 1e-25. Figure taken directly from (Bertram & Tanzi, 2020).*

Table 1.4: Genetic variants identified in the Jansen et al (2019) and Kunkle et al (2019) GWAS. *Chr, Base position, A1 and A2 are taken from the Jansen at al. nearest gene determined from UCSC genome browser (GRCh37; hg19). AD pathway as determined in either or both GWAS: immune = immune system response; lipid = lipid metabolism, APP = APP metabolism, tau = tau protein binding.* Stage 1 result from Kunkle at al.** Stage 2 result from Kunkle at al. *** Stage 3 result from Kunkle at al. MAF = minor allele frequency from European controls as provided on GnomAD [v.2.1.1.; https://gnomad.broadinstitute.org/]. Table and legend adapted from (Bertram & Tanzi, 2020).*

Chr	Base position	Lead SNP	A1	A2	MAF	P-value Jansen et al	P-value Kunkle et al	AD effect	Nearest gene	AD pathway	Potential link to AD pathogenesis
1	161155392	rs4575098	A	G	0.24	2.05E-10	2.34E-02*	Risk	*ADAMTS4*	None	Neuroprotection: Extracellular Matrix Protease
1	207786828	rs2093760	A	G	0.225	1.10E-18	1.66E-15*	Risk	*CR1*	Immune	Innate Immunity; Neuroinflammation
2	127891427	rs4663105	A	C	0.412	3.38E-44	2.16E-26*	Risk	*BIN1*	Lipid	Cellular Protein Trafficking
2	233981912	rs10933431	G	C	0.24	8.92E-10	3.42E-09**	Protection	*INPPD5*	None	Autophagy; Viral Infection
4	11026028	rs6448453	A	G	0.228	1.93E-09	4.90E-05*	Risk	*CLNK*	None	Innate Immunity; Neuroinflammation
6	32583357	rs9469112	T	A	0.153	8.41E-11	2.32E-07**	Protection	*HLA-DRB1*	Immune	Adaptive Immunity
6	47432637	rs9381563	C	T	0.344	2.52E-10	3.57E-10**	Risk	*CD2AP*	None	Blood Brain Barrier; Aβ Transcytosis
7	99971834	rs4727449/ rs1859788	A	G	0.323	2.22E-15	1.22E-09**	Protection	*ZCWPW1/ NYAP1*	None	Innate Immunity; Neuroinflammation
7	143108158	rs7810606	T	C	0.425	3.59E-11	1.13E-06**	Protection	*EPHA1*	None	Signal Transduction
8	27464929	rs28834970/rs4236673	A	G	0.39	2.61E-19	5.60E-23**	Protection	*CLU/PTK2B*	Immune; lipid; tau	Aβ clearance/Signal Transduction
10	11717397	rs11257238	C	T	0.382	1.26E-08	2.61E-07**	Risk	*ECHDC3*	None	Lipid Metabolism

11	47380340	rs3740688	G	T	0.458	4.50E-05	5.46E-13**	Protection	*SPI1/CELF1*	Immune	Innate immunity; cellular communication; myeloid cell development
11	59958380	rs2081545	A	C	0.342	1.55E-15	5.35E-17**	Protection	*MS4A6A*	Immune	Innate Immunity; Neuroinflammation
11	85776544	rs867611	G	A	0.342	2.19E-18	3.41E-19**	Protection	*PICALM*	*APP*	Blood Brain Barrier; Aβ Transcytosis
11	121435587	rs11218343	C	T	0.035	1.09E-11	2.88E-12**	Protection	*SORL1*	Lipid; *APP*	Cellular Protein Trafficking
14	53391680	rs17125924	G	A	0.099	5.26E-06	1.42E-09**	Protection	*FERMT2*	*APP*	*APP* processing
14	92938855	rs12590654	A	G	0.347	1.65E-10	8.73E-09*	Protection	*SLC24A4*	None	Calcium Homeostasis
15	59022615	rs442495	C	T	0.334	1.31E-09	2.51E-7**	Protection	*ADAM10*	Immune	Sheddase; *APP* Processing
15	63569902	rs117618017	T	C	0.132	3.35E-08	2.38E-04*	Risk	*APH1B*	None	γ-secretase; *APP* Processing
16	19808163	rs7185636	C	T	0.156	1.40E-01	2.4E-08 ***	Protection	*IQCK*	Unknown	Unknown
16	31133100	rs59735493	A	G	0.324	3.98E-08	7.42-03*	Protection	*KAT8*	None	Transcriptional Regulation
16	79355857	rs62039712	G	A	0.094	7.66E-01	3.70E08 *	Risk	*WWOX*	Lipid; Tau; *APP*	Cholesterol metabolism; regulates tau hyper-phosphorylation; neurofibrillary formation; Aβ aggregation
17	5138980	rs113260531	A	G	0.118	9.16E-10	3.70E-04**	Risk	*SCIMP*	None	Innate Immunity; Neuroinflammation
17	47450775	rs28394864	A	G	0.471	1.87E-08	4.85E-03*	Risk	*ABI3*	None	Innate Immunity; Neuroinflammation

17	61538148	rs138190086	A	G	0.017	2.65E-04	5.30E-09***	Risk	*ACE*	Immune	Aβ degradation; blood pressure regulation
18	56189459	rs76726049	C	T	0.011	3.30E-08	1.76E-01*	Risk	*ALPK2*	None	Signal Transduction
19	1039323	rs111278892	G	C	0.165	7.93E-11	1.10E-07*	Risk	*ABCA7*	Lipid; *APP*	Lipid Metabolism; Innate Immunity
19	45411941	rs429358	C	T	0.155	<1E-900	1.17E-881*	Risk	*APOE*	Lipid; *APP*; tau	Aβ clearance/Lipid Metabolism
19	46241841	rs76320948	T	C	0.059	4.64E-08	1.22E-04*	Risk	*AC074212.3*	None	unknown
19	51727962	rs3865444	A	C	0.336	6.34E-09	5.27E-06**	Protection	*CD33*	None	Innate Immunity; Neuroinflammation
20	54998544	rs6014724	G	A	0.089	6.56E-10	3.65E-07*	Protection	*CASS4*	None	Signal Transduction
21	28156856	rs2830500	A	C	0.336	1.65E-02	2.60E-08 ***	Protection	*ADAMTS1*	*APP*	Not fully explored but thought to be involved in amyloidosis
Rare variants											
3	57226150	rs184384746	T	C	0.002	1.24E-08	n.a.	Risk	*HESX1*	None	Homoebox Gene; Development
6	41129252	rs75932628	T	C	0.002	2.95E-15	2.95E-12*	Risk	*TREM2*	Immune system response	Innate Immunity; neuroinflammation
7	145950029	rs114360492	T	C	0.0003	2.10E-09	n.a.	Risk	*CNTNAP2*	None	Neuronal Development
16	81942028	rs72824905	G	C	0.01	2.11E-03	7.92E-03*	Protection	*PLCG2*	None	Microglial activation; neuroinflammation
17	47297297	rs616338	T	C	0.01	7.81E-07	n.a.	Risk	*ABI3*	None	Microglial activation; neuroinflammation

In the past year several preprints documenting additional LOAD GWAS analyses have been published (see **Table 1.3**) (Bellenguez *at al.*, 2020; de Rojas *at al.*, 2020; Schwartzentruber *at al.*, 2021; Wightman *at al.*, 2020), although the summary statistics from these studies were not publicly available for use in my thesis. The first GWAS to combine the datasets used by Jansen *at al.* (2019) and Kunkle *at al.* (2019) was conducted by De Rojas and colleagues (de Rojas *at al.*, 2020). They combined three sets of summary statistics of European ancestry: a Spanish case-control study (GR@*ACE*/DEGESCO study; n = 12,386), the case-control study of International Genomics of Alzheimer project (IGAP; n = 82,771) and the UK Biobank (UKBB) AD-by-proxy (based on family history) case-control study (n=314,278). In addition to the *APOE* locus, de Rojas *at al.* identified 39 LOAD -associated SNPs which surpassed genome-wide significance (p<5e-08) (de Rojas *at al.*, 2020). This GWAS built on previous findings (see **Figure 1.7**) and the increase in power from meta-analysing past LOAD summary statistics in addition to the inclusion of new previously unanalysed cohorts, led to the identification of six novel SNPs, annotated to the genes CHRNE, *APP*, PRKD3/NDUFAF7, PLCG2 and SHARPIN. Perhaps the most interesting finding from the de Rojas *at al.* GWAS was the identification of the common variant rs2154481 which resides in the *APP* locus which is also a familial EOAD gene and suggests common pathways between the familial and sporadic forms of AD.

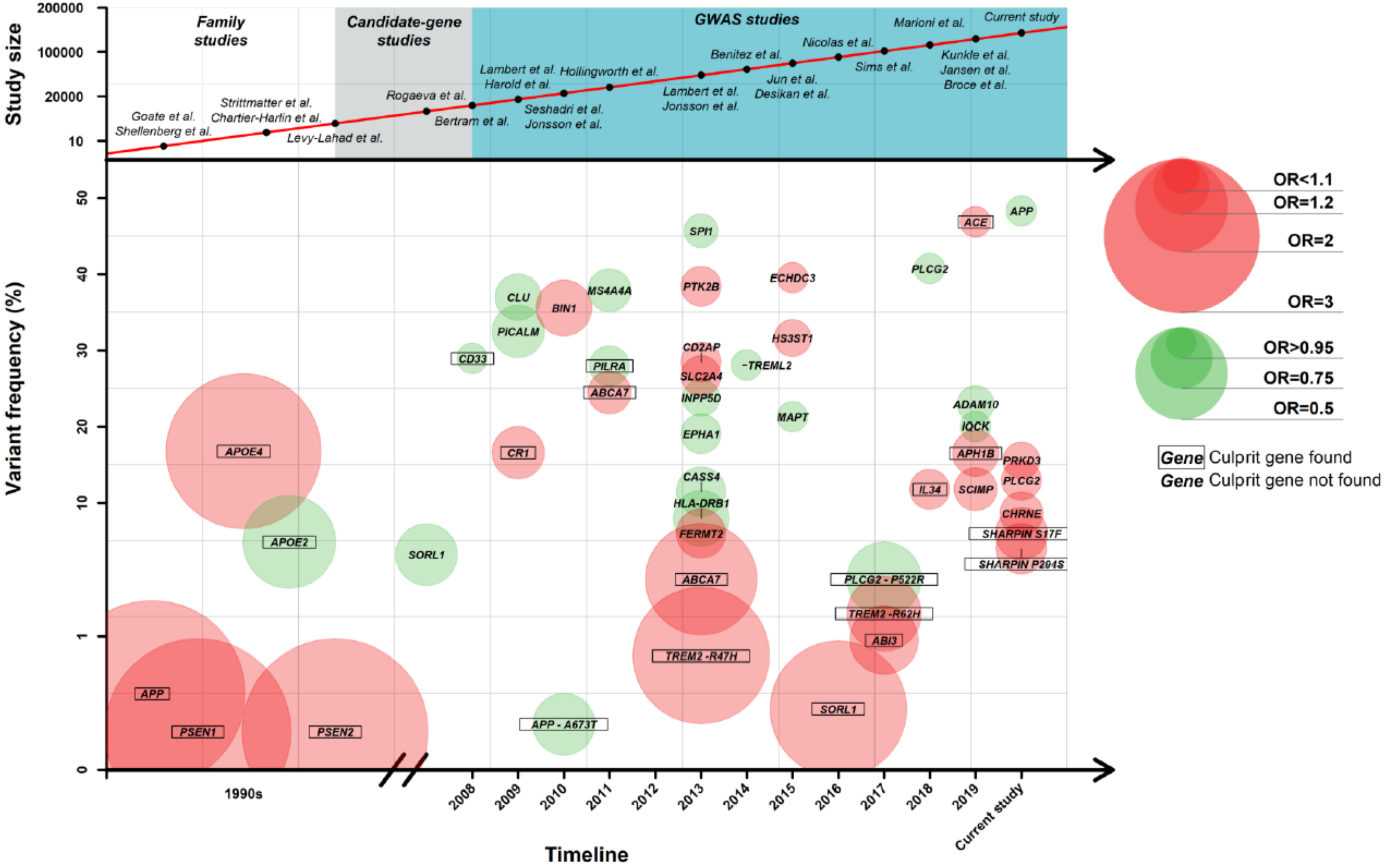


Figure 1.7: Landscape for Alzheimer's disease. *Shown are the genes implicated in late onset Alzheimer's disease identified by GWAS. OR = odds ratio. Green represents a protective effect. Red represents a risk effect. Figure taken directly from (de Rojas et al., 2020).*

A more recent preprint by Bellenguez and colleagues (Bellenguez *at al.*, 2020) identified 75 loci (42 novel) with a sample size of 487,511 (85,934 cases and of these 46,828 were AD by proxy). Pathway analyses of genes annotated to these variants support the involvement of amyloid and tau pathways and further highlight the role of microglia in AD (**Figure 1.8**). Of note, several of the novel loci are also associated with FTD (e.g. MAPT, GRN, and TMEM106B), potentially reflecting misclassification in the diagnosis of clinical AD and the proxy-AD diagnoses. Alternatively, as mentioned previously, AD and other dementias are rarely found without other neurodegenerative co-pathologies (Wingo, Lah, Levey, & Cutler, 2012), therefore it is likely there are similarities in the underlying architecture for the dementia subtypes. This hypothesis is further supported by research indicating there is genetic pleiotropy (where SNPs have an effect on more than one trait) between LOAD, FTD and PD (Ferrari *at al.*, 2017).

A recent meta-analysis conducted by Schwartzentruber and colleagues (2021) identified 37 loci in total and four novel variants (annotated to CCDC6, TSPAN14, NCK2 and SPRED2). Of note, Stage 1 of this study included only 898 clinically confirmed AD cases with the remaining 52,791 cases (99% of cases) being AD-by proxy. These results were then meta-analysed with the Stage 1 Kunkle *at al.* (2019) GWAS results. They identified that 21 of the 37 associated variants had over 50% probability of being causally involved in AD risk via the use of fine-mapping methods (Schwartzentruber *at al.*, 2021). Notably, although genetic correlation analysis suggests the proxy measure of AD is accurate for clinical diagnosis, the accuracy varies, particularly when looking at the paternal correlations estimated by Marioni and colleagues (r=0.67) (Marioni *at al.*, 2018). This variation needs to be taken into consideration when interpreting LOAD GWAS including by-proxy measures. However, the increase in predictive power when including by-proxy measures is likely to compensate for the fact that some individuals will be misclassified using this method.

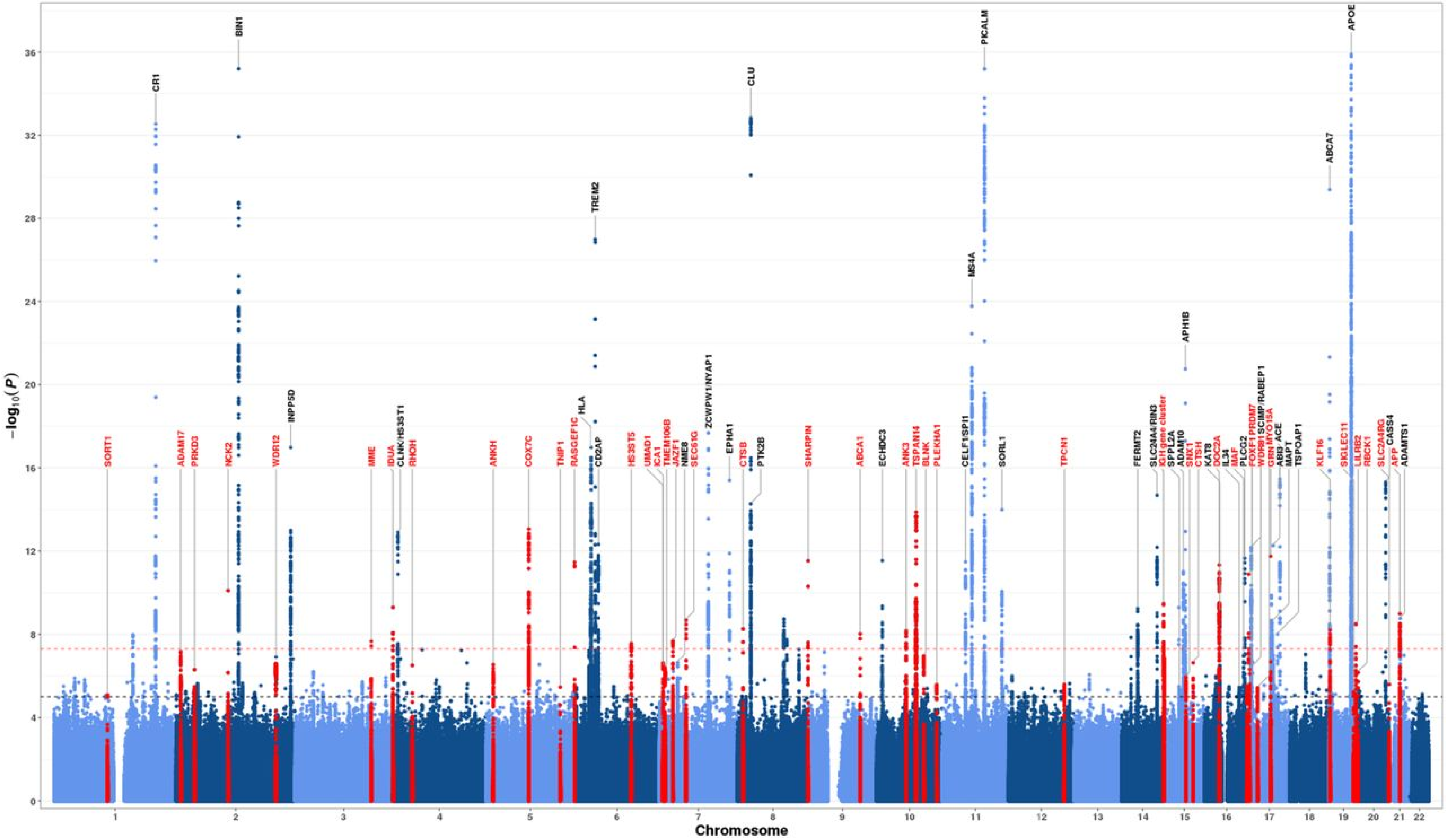


Figure 1.8: Manhattan plot of the latest LOAD GWAS conducted by Bellenguez and colleagues (2020). *The x axis is the genomic position, segregated by chromosome. Shown on the y-axis is the –log10 p-value from each GWAS. Each point represents a SNP. The red horizontal line represents the genome-wide significance (p<5e-08). P-values are truncated at 1e-36. Figure taken directly from (Bellenguez et al., 2020)*

In addition to the common genetic variants of small effect identified by GWAS, rare variants (with a population frequency <1%) associated with a moderate to high effect on an individual's risk have been identified and validated using next generation sequencing technologies (NGS) including whole-genome and exome sequencing. Rare variants associated with LOAD have been annotated to genes including *TREM2* (Bellenguez *at al.*, 2017) (a transmembrane receptor expressed in cells of the myeloid lineage and is thought to be a driver within the immune and inflammatory pathways in the cause of the AD); *SORL1* (Bellenguez *at al.*, 2020)(which encodes SorLA – a protein involved in the processing of *APP* and the secretion of Aβ) and *ABCA7* (Bellenguez *at al.*, 2020) (a transmembrane protein which is thought to influence AD pathogenesis through various mechanisms, including the regulation of *APP* processing and clearance of Aβ via phagocytosis). Of note, a number of rare LOAD variants are found in genes where common GWAS variants have been identified, which suggests common pathways to disease susceptibility and multiple ways in which the same gene can be disrupted (Sims, Hill, & Williams, 2020).

To better understand the biology of GWAS variants and the genes implicated, pathway analysis methods have been developed that test if there is an excess of association signal (i.e. an enrichment) in sets of genes based on independent annotations. Within this thesis pathway analysis has been run in **Chapters 4** and **5** (see Chapter 2 section **2.3.4** for more details). Pathway analyses have been used to identify disease-relevant processes and the application of these methods suggest a role for multiple biological pathways in LOAD including immune regulation, synaptic pathways, cholesterol transport / lipid metabolism, endocytosis, ubiquitination, protein folding, Aβ clearance and tau biology (Kunkle *at al.*, 2019).

1.2.1 Genetics of other neurodegenerative diseases

Although genetic studies of other neurodegenerative diseases have been considerable smaller than for AD, GWAS have identified variants associated with other dementias (see **Table 1.5**) including PD (Chang *at al.*, 2017; Nalls *at al.*, 2019), where a recent study identified 90 PD-associated genome-wide significant SNPs including multi-signal loci residing in the genes *GBA, NUCKS1* and *RAB29, GAK* and *TMEM175, SNCA*, and *LRRK2;* LBD (Guerreiro *at al.*, 2018), where four SNPs have been identified residing in the genes *APOE*, *SNCA, GBA* and *CNTN1*; and FTD (Ferrari *at al.*, 2014), where 3 SNPS have been identified which are all located within the HLA locus. However, these studies may still be statistically underpowered to detect all associated variants with neurodegenerative diseases (Escott-Price *at al.*, 2015; Escott-Price, Shoai, Pither, Williams, & Hardy, 2017) and much of the genetic signal associated with these diseases is yet to be robustly identified (Ridge, Mukherjee, Crane, Kauwe, & Alzheimer's Disease Genetics Consortium, 2013).

Table 1.5: GWAS conducted in other neurodegenerative diseases.

Dementia Phenotype	Reference	Approach	Sample size		Total number of loci
			Total	Number of Cases	
Parkinson's Disease	Chang *at al.* (2017)	Diagnosed case-control GWAS	416,518	19,476	41
Parkinson's Disease	Nalls *at al.* (2019)	Diagnosed and proxy case-control GWAS	1,474,097 (436,419 [30%] proxy)	56,306 (18,618 [33%] proxy)	90
Lewy Body Dementia	Guerreiro *at al.* (2018)	Diagnosed case - control GWAS	5007	1216	4
Frontotemporal Dementia	Ferrari *at al.* (2014)	Diagnosed case - control GWAS	12928	3526	3

1.2.2 Polygenic Risk scores

To further investigate the genetic architecture of polygenic traits such as LOAD, methods have been developed to combine the information provided by independent SNPs. Polygenic risk scores (PRS) quantify genetic burden as an accumulative genetic score for each individual in a sample and are calculated as a sum of trait-associated variants, weighted by effect sizes estimated from the discovery GWAS (see **Figure 1.9**). Essentially, PRS summarise GWAS data into a single score for individuals in the target sample. The standard method of generating a PRS is based on an additive method. This involves 'clumping' the GWAS summary statistics such that the most significant variant in each LD block is retained and the PRS is calculated for each individual in a dataset as the number of reference alleles multiplied by the effect size (e.g odds ratio or beta), and then summed across all retained clumped variants with a GWAS p-value meeting a certain threshold (see **Figure 1.9**). SNPs which do not independently meet genome-wide significance are often included in PRS analysis as evidence has shown that the increase in predictive power when aggregating the SNPs compensates the increase in false positives when using higher significance thresholds (Euesden, Lewis, & O'Reilly, 2015).

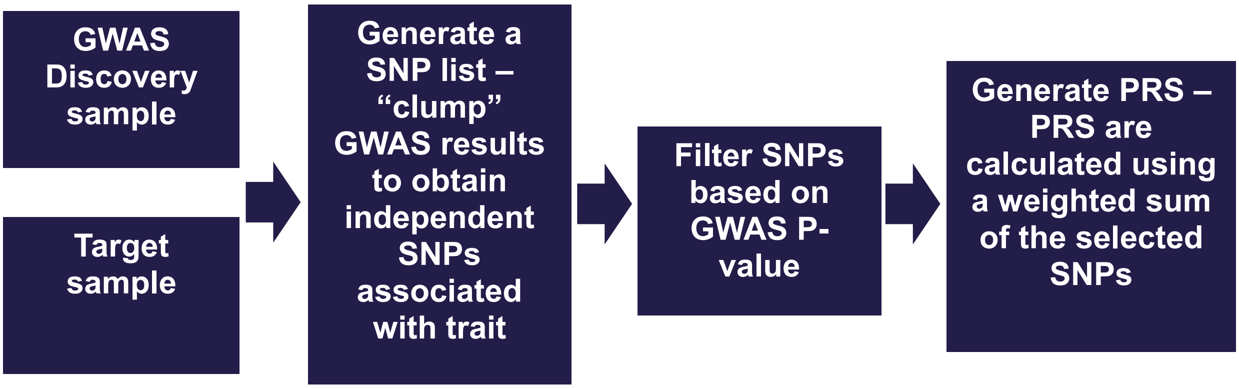

GWAS Discovery sample
Target sample
Generate a SNP list – "clump" GWAS results to obtain independent SNPs associated with trait
Filter SNPs based on GWAS P-value
Generate PRS – PRS are calculated using a weighted sum of the selected SNPs

Figure 1.9 The process of generating PRS using the additive method.

PRS have primarily been used on the phenotype they were trained on to test prediction of case-control status. A measure of association is generally used to evaluate the PRS. For example, the proportion of variation in liability for developing a phenotype explained by the PRS is often used (R^2 for continuous traits and Nagelkerke's pseudo-R^2 for binary traits) or the area under the curve (AUC), which provides a measure of how well a parameter (in this case the PRS) can distinguish between two diagnostic groups. It has been statistically hypothesised that cases for a disease will have a higher PRS for the disorder than controls. Provided the sample size is large enough, this is the general trend seen across neurological and psychiatric disorders, however the differences are small, and many individuals, including cases, have scores close to the population average (see **Figure 1.10**). PRS are not currently clinically useful for risk prediction due to these small differences between case-control groups. However, PRS might represent good biomarkers for mechanistic studies, for example to explore the regulatory genomic mechanisms leading to disease; this is the crux of **Chapter 6** of my thesis.

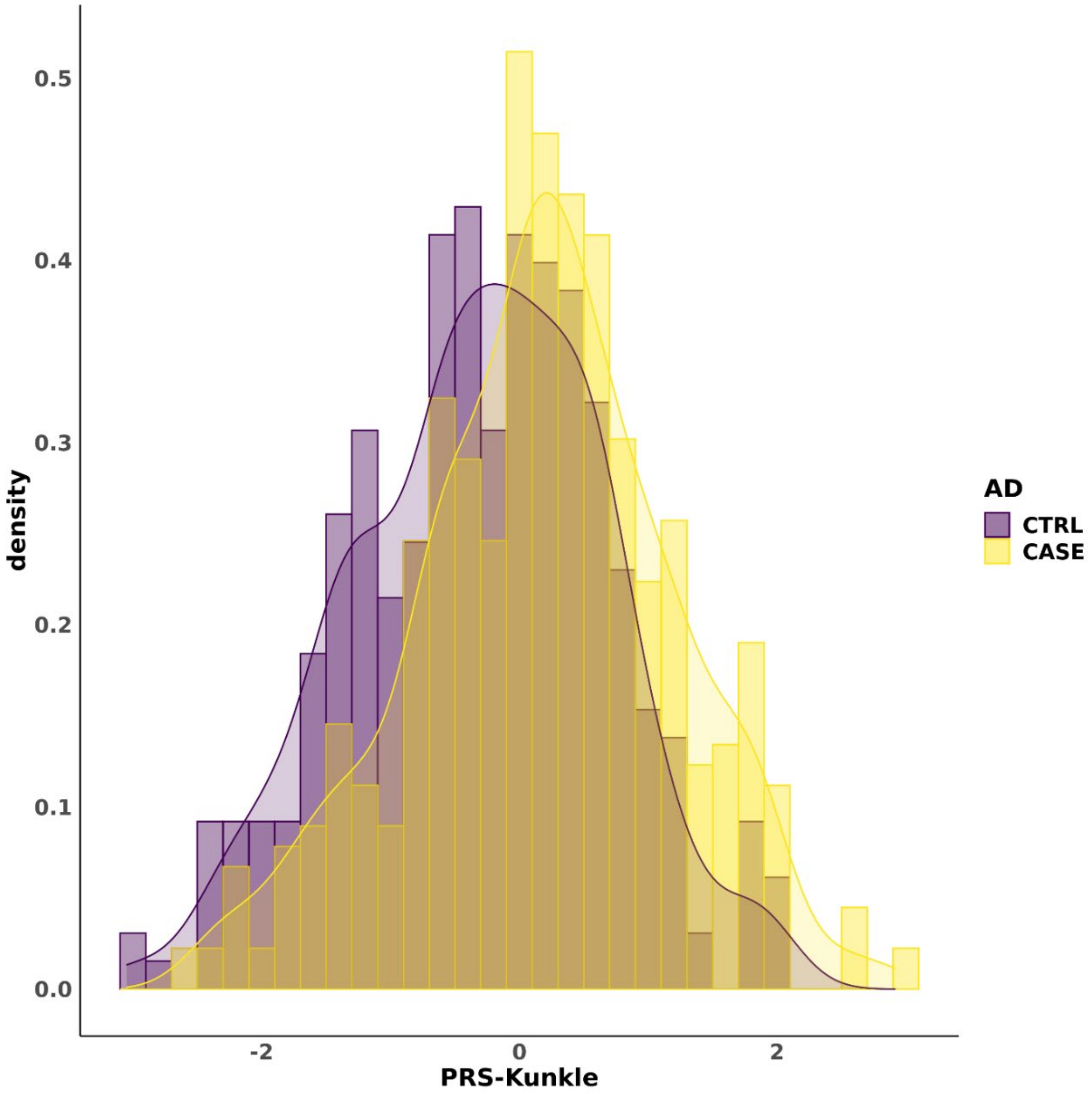


Figure 1.10: PRS in cases is close to population mean in brain disorders. *Shown is a density histogram of PRS generated using the Brains for Dementia research Cohort, split by AD status. PRS were calculated using the Kunkle at al. (2019) GWAS summary statistics excluding the APOE region and were standardised to have a mean of 0 and standard deviation of 1.*

PRS capture common genetic variation which contributes to an individual's susceptibility to disease. Although the additive method to quantify genetic loading is simplistic, it is currently one of the best methods developed to measure this variation; there is little evidence suggesting there is any interaction between the variants included in a PRS (Lewis & Green, 2021). In addition, this method works because the effects are small and therefore all very similar. However, within the additive method gene-gene and gene-environment interactions cannot be modelled within the PRS. Conversely, a large meta-analysis of heritability estimated from twin studies supports the additive PRS model for the majority of traits and in particular neurological diseases, where 99.5% of twin studies tested in their study were consistent with a model where all resemblance was due to additive genetic variance (Polderman *at al.*, 2015).

Other polygenic risk score methods model the correlation structure between SNPs without identifying the minimal subset of variants for predictions (i.e. not filtering on p-value). Methods which do this include Bayesian LD approaches which have focussed on increasing the power of PRS. For example, *SBayesR* (Lloyd-Jones *at al.*, 2019) is a novel method which uses a Bayesian multiple regression model and has been shown to improve the variance explained by the PRS in comparison to the standard clumping and p-value thresholding method. Furthermore, using tagging SNPs (the most associated SNP in a sample) in the additive method as opposed to the causal variant limits the precision of PRS. New methodologies such as *PleioPred (Hu at al., 2017),* which uses GWAS summary statistics as its input, models multiple genetically correlated diseases as well as a variety of external information such as LD and functional annotations, which has been shown to increase the accuracy of risk prediction.

There are four key aspects to consider when generating a PRS to better understand what the score is capturing (Lewis & Green, 2021): 1) Known information – which shows where an individual lies on a risk scale along the normal distribution in relation to other individuals in that population; 2) The unknown information from environmental exposures not included in the model and incomplete genetics; 3) The possibility for individuals to be assigned to the incorrect phenotypic group/ the accuracy of the association statistics in the GWAS and 4) What the PRS will be applied to i.e. it is unlikely to be sufficient to justify therapeutic interventions based on PRS alone. The first two considerations can be statistically inferred using a measure of variance explained by the PRS, however it is harder to statistically evaluate the last two but they are still important limitations to reflect upon when applying PRS.

1.2.3 Polygenic risk scores in LOAD

There has been success in applying PRS in LOAD to significantly predict disease status. Using the IGAP dataset (>17,000 cases and >37,000 controls), Escott-Price and colleagues developed the Cardiff PRS (Escott-Price *at al.*, 2015). They produced an algorithm using >87,000 variants (p-value threshold < 0.5) in addition to age and sex, and the predictor yielded an AUC of 78% (i.e. cases/controls are accurately predicted 78% of the time) (Escott-Price *at al.*, 2015). The Cardiff PRS has been validated on other cohorts, and performs best when applied to pathologically confirmed LOAD cases (AUC=84%)(Escott-Price, Myers, Huentelman, & Hardy, 2017). It must be considered that the variance explained by the LOAD PRS is generally low (~1-5%) (Escott-Price *at al.*, 2015) and only adds a modest amount of information to a prediction model, reducing its clinical application. On the other hand, PRS are currently one of the best predictors for LOAD and have aided understanding some of the contribution of variation in the disease. To separate the effects of *APOE* from the other AD variants the *APOE* region is often excluded from analysis. Studies have shown that the AUC for the PRS excluding *APOE* is 75% and for *APOE* alone is 75% (Leonenko *at al.*, 2019). The combined inclusion of both the PRS and *APOE* increases the AUC (to ~82%) (Leonenko *at al.*, 2019). In this thesis (**Chapter 6**) I generate PRS both including and excluding the *APOE* region to explore epigenomic changes associated with AD risk.

Beyond the prediction of disease status, PRS can be used as a tool to investigate how genetic risk mediates the development of specific symptoms, and has the potential to separate the primary causal features from the secondary consequences of the disease. In addition to the robust associations between PRS and disease status, LOAD PRS have been shown to correlate with MCI and cognitive decline (Felsky *at al.*, 2018; Ge *at al.*, 2018; Kauppi, Rönnlund, Nordin Adolfsson, Pudas, & Adolfsson, 2020; Marioni *at al.*, 2017; Mormino *at al.*, 2016); memory impairment (Marioni *at al.*, 2017; Mormino *at al.*, 2016); brain measurements such as hippocampal volume (Axelrud *at al.*, 2018) and cortical thickness (Sabuncu *at al.*, 2012); cerebrospinal biomarkers (Martiskainen *at al.*, 2015); inflammatory biomarkers (Morgan *at al.*, 2017); and neuropathological measures (Desikan *at al.*, 2017; Felsky *at al.*, 2018; Hannon *at al.*, 2020). The extent of the associations with LOAD PRS highlights the complexity of understanding the genetic pathways involved in LOAD pathogenesis.

1.3 Regulatory Genomic Variation in Alzheimer’s disease

Even with the successes of GWAS, relatively little is known about the functional mechanisms by which risk variants mediate disease susceptibility. As the majority of GWAS variants associated with LOAD do not index coding variants affecting protein structure they are hypothesized to influence gene regulation. For example, in the 2019

Kunkle *at al.* GWAS, only ~2% of LOAD-associated variants were located in exons, with the majority (58%) being located in introns (Kunkle *at al.*, 2019). This notion is further supported by research suggesting LOAD GWAS variants are enriched in regulatory domains including enhancers and regions of open chromatin (Kikuchi *at al.*, 2019; Marzi *at al.*, 2018).

1.3.1 Introduction to epigenetics

Advances have been made in understanding the role of functional genomic processes in complex disease phenotypes. Of particular interest is the epigenome, which encompasses numerous chemical modifications to DNA and histone proteins which directly and dynamically influence the regulation of gene expression (Schübeler, 2015). The term epigenetics has many definitions; in this thesis I use it to refer to the study of mitotically heritable, but reversible, changes in gene expression which arise independently of genetic variation (Henikoff & Matzke, 1997). Unlike the DNA sequence, the epigenome is dynamic and can influence a cell without altering the genetic sequence.

Epigenetic mechanims play a key role in cell differentiation and are involved in the regulation of gene expresion, genomic imprinting (Ferguson-Smith, 2011) (a mammalian inheritence process controlled by epigenetic markers) and X-chromosome inactivation (Heard, Clerc, & Avner, 1997) (a dosage compensation mechanism which ensures that XX females and XY males have equal levels of gene expression from the X-chromosome). The 'Epigenetic Landscape' first described by Waddington (Waddington, 1957)(see **Figure 1.11**) provides a metaphorical framework for how gene regulation is involved in cell differentiation during development. In Waddington's model marbles roll from the top of a landscape but due to the uneven terrain, the trajectory will be different for each marble running down the valley, and thus they will have a different end point. The marbles in Waddington's metaphor represent all nucleated cells and the fact they contain the same genetic information, but due to the complex interplay of epigenetic mechanisms regulating genes at the cellular level (the uneven terrain in Waddington's model), the cells end up physiologically different at the end.

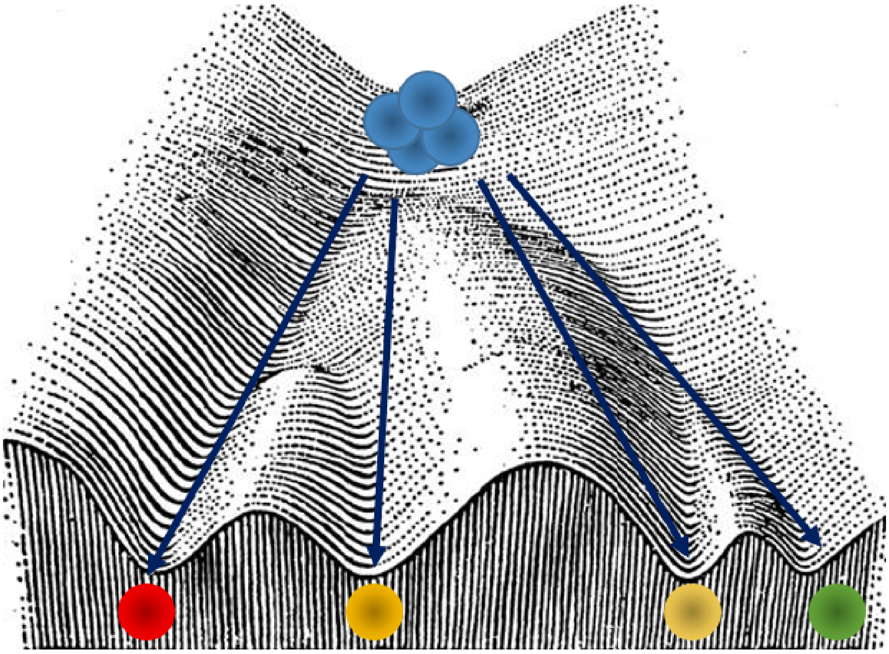


Figure 1.11: Waddington's Epigenetic Landscape*. This figure represents a metaphorical framework for how gene regulation modulates cell differentiation during development. The marbles represents cells, which follow a different pathway through the 'valleys' and the different trajectories determine the cell fates. Figure adapted from Waddington (1957).*

Epigenetic mechanisms are cell-type-specific and are dynamic during development and aging. Epigenetic marks defining cellular phenotypes are mitotically inherited although they can be influenced by environmental (e.g. pollution (Breton & Marutani, 2014)) and lifestyle (e.g. tobacco smoking (Elliott *at al.*, 2014) and medication (Viuff *at al.*, 2016)) factors and stochastic changes. As epigenetic changes are dynamic, there is the potential for disease associated epigenetic modifications to be targeted for future therapeutic development due to this "reversible" nature.

Waddington's epigenetic landscape can also be used to illustrate the hypothesis of how the landscape is ultimately under genetic control. This hypothesis states that genes indirectly control development via a network of interacting biochemical products. Shown in **Figure 1.12**, is a different perspective of the landscape, where the contours on the landscape are controlled by genes. The ropes represent the gene products altering the landscape and the connections represent the biochemical interactions between those gene products (Waddington, 1957).

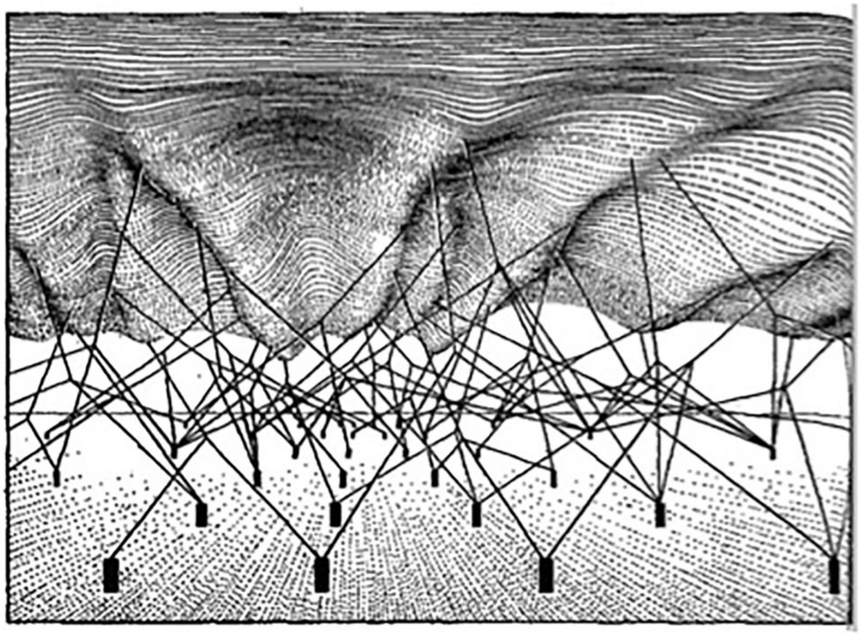


Figure 1.12: Waddington's epigenetic landscape show that its topology is underpinned by the influence of genes. *Figure taken directly from Waddington (1957).*

1.3.2 DNA methylation and other cytosine modifications

DNA methylation (DNAm) is the best characterised and most stable epigenetic mechanism involving the addition of a methyl group to the 5' carbon in a cytosine ring which leads to the formation of 5-methylcytosine (5mC; see **Figure 1.13**) (Horvath, 2013; Schübeler, 2015). DNAm almost exclusively occurs at CpG (cytosine-guanine) sites in mammals (Schübeler, 2015). Around 70% of CpG dinucleotides in the genome are methylated, and are prone to deamination, which causes a cytosine to thymine transition, leading to the frequency of guanine and cytosine to be depleted in the genome (Bird, 1986; Strichman-Almashanu *at al.*, 2002) except at large, dense regions of CpG sites called CpG Islands (CGIs), which are estimated to account for 1-2% of the genome (Antequera & Bird, 1993). CGIs have been a target of DNAm research, with many CGIs (~40%) being located in promoters of housekeeping control genes (Larsen, Gundersen, Lopez, & Prydz, 1992; Strichman-Almashanu *at al.*, 2002). DNAm can influence the functional state of regulatory regions without changing the DNA sequence itself and has predominantly been associated with transcriptional repression (Ng & Adrian, 1999) (see **Figure 1.13**) but its properties are not fully established. The primary hypothesis of gene silencing via DNAm occurs proposes the methylation of the cytosine in a CpG dinucleotide influences transcriptional machinery by blocking transcription factors from binding and additionally, signalling methyl-binding proteins, which in turn causes chromatin to compress and silences genes. However, recent research suggests that the relationship between DNAm and transcription is more complicated; where gene-body DNAm and non-CpG methylation has been associated with gene expression and alternative splicing as opposed to repression. Recent literature supports a role for other cytosine modifications in gene regulation such as 5-hydroxymethylcytosine (5hmC) which is a product of the de-methylation and oxidation of 5mC by the ten-eleven translocation (TET) group of enzymes. In the past 5hmC was considered to have no effect on gene regulation but recent literature suggests it is enriched in the brain (more specifically within synaptic genes) and is thought to have important roles during neurodevelopment (Spiers, Hannon, Schalkwyk, Bray, & Mill, 2017).

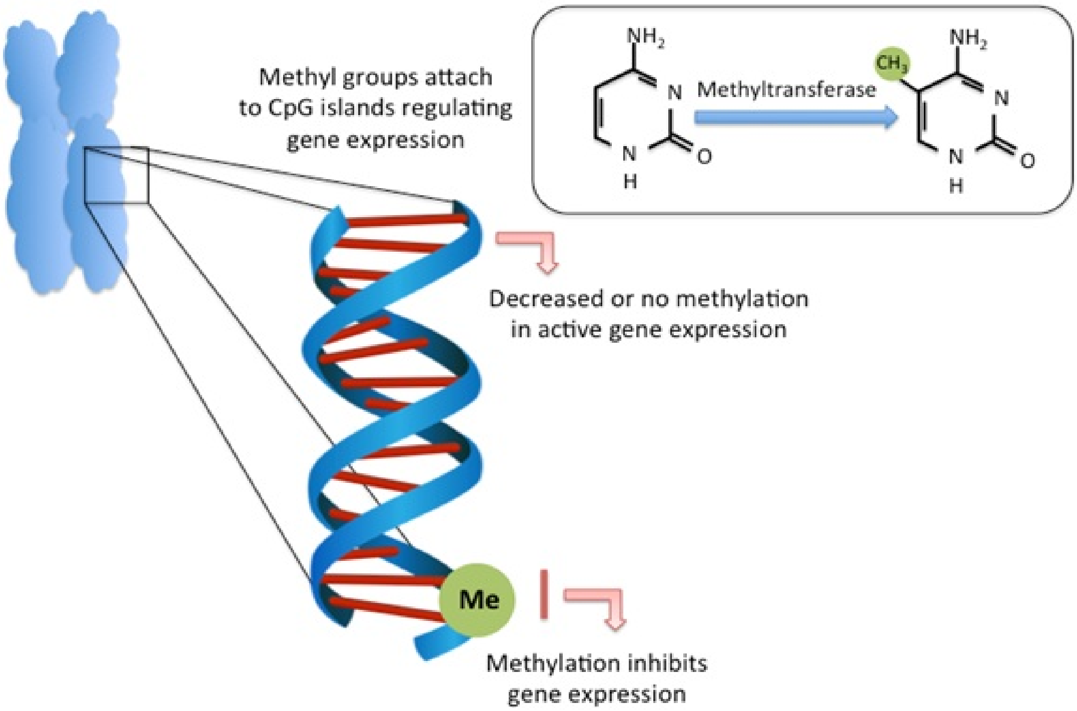

Methyl groups attach to CpG islands regulating gene expression
NH_2
N
N
H
O
Methyltransferase
CH_3
NH_2
N
N
H
O
Decreased or no methylation in active gene expression
Me
Methylation inhibits gene expression

Figure 1.13: DNA methylation involves the addition of a methyl group to the 5' carbon in a cytosine ring. *Top right: on the left is cytosine and on the right is 5-methylcytosine. Figure taken directly from (Nevin & Carroll, 2015)*

1.3.3 Epigenetic Variation associated with AD

Multiple studies suggest epigenetic variation plays a role in the development of complex diseases of the brain including LOAD (see **Table 1.6**) (Roubroeks *at al.*, 2020; A. R. Smith *at al.*, 2016; A. R. Smith, Smith, Pishva, *at al.*, 2019; R. G. Smith *at al.*, 2018; R. Smith *at al.,* 2021; Vasanthakumar *at al.*, 2020). Similarly to GWAS, genome-wide variation in DNAm has been explored using epigenome-wide association studies (EWAS), where DNAm at each CpG site is independently tested for association with the trait of interest.

1.3.3.1 Epigenetic studies of LOAD using post-mortem brain tissue

Unlike germline genetic variation, epigenetic signatures are tissue specific and therefore the tissue type used in an EWAS is important to consider; most EWAS of neurodegenerative diseases have been conducted in regions of the cortex. The first EWAS of AD in brain tissue was conducted by Bakulski *at al.* (2012), and DNAm was quantified in the frontal cortex using the Illumina Infinium Human Methylation 27K BeadChip array (27K array) - which measures methylation at >27,000 DNAm sites - in 12 LOAD cases and 12 cognitively normal donors. Despite the small sample size, they identified 948 differentially methylated positions (DMPs) which were annotated to 918 unique genes, with the most significant located in the *TMEM59* loci which has been implicated in *APP* processing. However, they did not control for multiple testing, which would lead to an increase in false positives and consequently the majority of the DMPs have not been validated in subsequent LOAD EWAS.

The development of the Illumina Infinium 450K Beadarray (450K array) (see Chapter 2 section **2.1.1**) allowed for DNAm to be quantified at > 450,000 CpG sites. Two EWAS studies were conducted in parallel in 2014 which utilized the newer array technology. De Jager *at al.*, using a large cohort of 708 donors (prefrontal cortex samples - 460 AD, 263 controls), identified 71 DMPs which were associated with AD, including sites residing in LOAD GWAS loci such as *ABCA7* and *BIN1*. Other genes identified include *ANK1, CDH23, DIP2A, RHBDF2, RPL13, SERPINF1* and *SERPINF2*. Lunnon *at al.* conducted an association study in 122 individuals (three cortical regions [entorhinal cortex, superior temporal gyrus, and prefrontal cortex] and the cerebellum) with Braak NFT stage and identified significant hypermethylation in the *ANK1* gene and additionally, 11 of De Jager *at al.*'s 71 DMPs were validated within this sample. A number of other brain studies investigating LOAD have been conducted using the 450K array (see **Table 1.6**) (Altuna *at al.*, 2019; Lardenoije *at al.*, 2019; A. R. Smith *at al.*, 2016; A. R. Smith, Smith, Pishva, *at al.*, 2019; R. G. Smith *at al.*, 2018; R. Smith *at al.*, 2021; Watson *at al.*, 2016), in which multiple DMPs and differentially methylation regions (DMRs; where variable DNAm across a region consisting of multiple CpG sites is associated with a trait) have been identified. Of particular interest is the *HOXA* region, which has repeatedly been recognised to have an influence on LOAD (Altuna *at al.*, 2019; R. G. Smith *at al.*, 2018; R. Smith *at al.*, 2021). Recently, Smith, Pishva and colleagues conducted a meta-analysis of AD EWAS studies (R. Smith *at al.*, 2021), combining data from six 450K array analyses (N=1,453 unique individuals) to identify differential methylation associated with Braak NFT across multiple cortical regions. This is the largest AD cortex meta-analysis to date. In their cross-cortex meta-analysis (N=1,408 donors) they identified 220 DMPs associated with Braak NFT stage, annotated to 121 genes. These results further support a role for differential DNAm across many genes in AD.

The majority of AD EWAS analyses have been conducted in bulk tissue but these studies cannot capture cell-type specific changes, which may play a vital role in the pathogenesis of LOAD. Therefore, recent work by Gasparoni and colleagues investigated this hypothesis, whereby they profiled both bulk (52 controls and 76 AD cases from frontal cortex and temporal cortex samples) and sorted neuronal and non-neuronal nuclei (isolated from 31 occipital cortex samples) to investigate if there are cell-type specific DNAm patterns. They identified that DNAm differences in the *HOXA*3 regions are predominantly driven by modifications in neuronal cells and variable DNAm in *ANK1* was driven by modifications in glial cells, highlighting the limitations of bulk only studies and the importance of considering single-cell populations.

1.3.3.2 Epigenetic studies of LOAD using *peripheral* tissues

Although understanding epigenetic dysregulation in the brain is crucial to aid our understanding of dementia, investigating epigenetic changes in peripheral tissues of LOAD individuals may be useful in the context of identifying disease biomarkers. Biomarkers are defined as biological measures that have the potential to aid diagnosis, determine patient-specific aetiology and monitor disease progression. Epigenetic biomarkers with a clinical application have been developed for diseases where regulatory mechanisms undergo large, system-wide changes, such as in cancer (Das & Singal, 2004). Many clinical and epidemiological studies examine DNAm in easily accessible tissues and recently there has been an interest in profiling epigenetic variation in whole blood. In LOAD, blood biomarkers have the potential to aid clinical diagnosis, or be utilized as a tool to measure declining cognition over time. However, research investigating correlations between blood and brain deduced that whole blood DNAm measures provide limited information for disorders of the brain, though there are a proportion of sites where inter-individual variation is correlated between the two tissues (Hannon, Lunnon, Schalkwyk, & Mill, 2015).

Although the majority of AD EWAS have been conducted in brain, as this is the tissue directly affected by neuropathology, several LOAD EWAS have been conducted in peripheral tissues, identifying multiple DMPs and DMRs associated with disease (Lardenoije *at al.*, 2019; Roubroeks *at al.*, 2020; Vasanthakumar *at al.*, 2020)(see **Table 1.6**). Notably, the recent study by Roubroeks and colleagues (2020) which used the 450K array to quantify DNA methylation in 207 individuals (86 AD , 89 controls, and 109 MCI) and identified 9 DMRs associated with conversion from MCI to AD and 4 DMRs associated with baseline diagnosis, including one which was annotated to HOXB6, a region previously identified in brain EWAS. Another recent whole blood EWAS study, conducted by Vasanthakumar *at al.* (2020) used the Illumina Infinium EPIC Beadarray (EPIC array), which incorporates >850,000 DNAm sites (see Chapter 2 section **2.1.2** for more details), to quantify DNAm in 653 unique individuals (94 AD, 336 MCI, 223 controls) and identified 42 DMPs to be associated with AD vs control, 13 with AD vs MCI and 25 with MCI vs control. The DMPs were enriched in brain-specific genes such as *CLIP4*, *BIN1*, *BDNF* and *APOC1*.

Table 1.6: DNA methylation studies of Alzheimer's disease. *Table adapted from MacBean at al. (2020).*

Tissue type	Methods	Results	Reference
Prefrontal Cortex Entorhinal Cortex Middle Temporal gyrus Cerebellum	1,453 unique individuals of varying Braak NFT stage from six sample cohorts were used for discovery (Braak 0-II: 332, Braak III-IV: 627, Braak V-VI:494) Replicated in 661 samples. 450K array	220 Braak NFT stage-associated DMPs were identified cross-cortex 236 Braak NFT stage-associated DMPs were identified in the prefrontal cortex, 95 DMPs in the middle temporal gyrus and 10 DMPs in the entorhinal cortex Several DMPs identified were annotated to previously identified AD genes such as *ANK1*, *HOXA*, *PPT2/PRRT1* and *RHBDF2* Many novel DMPs identified	(R. Smith *at al.*, 2021)
Superior temporal gyrus (STG) Inferior frontal gyrus (IFG)	127 samples (67 AD and 60 cognitively normal control) for STG 117 samples (60 AD and 57 cognitively normal control) for IFG EWAS against Braak NFT Stage EPIC Array	5 and 14 DMPs associated with neuropathology in the STG and IFG, respectively including DNAm sites annotated to *ABCA7* and the *HOXA* gene cluster 21 and 173 DMRs associated with neuropathology in the STG and IFG, respectively Previously reported Braak NFT Stage -associated DMPs annotated to *RMGA, GNG7, HOXA3, GPR56, SPG7, PCNT, RP11-961A15.1, MCF2L, RHBDF2, ANK1, PCNT, TPRG1*, and *RASGEF1C* were replicated ($p < 0.0001$)	(Li, Sun, & Wang, 2020)

Whole blood	284 individuals (86 AD, 89 controls,109 MCI -38 MCI converted to AD within 1 year) 450K array	*MOV10L1* was associated with differences between all 3 groups 4 DMRs associated with baseline diagnosis, including those annotated to *HOXB6* and *CSNK1E* 9 DMRs associated with conversion from MCI to AD	(Roubroeks *at al*., 2020)
Whole blood	653 unique individuals (94 AD, 336 MCI, 223 controls) longitudinal with multiple data points EPIC array	42 DMPs associated with AD vs control 13 DMPs associated with AD vs MCI 25 DMPs associated with MCI vs control DMPs were enriched in brain related genes such as *CLIP4*, *BIN1*, *BDNF* and *APOC1*	(Vasanthakumar *at al*., 2020)
Hippocampus	38 individuals (26 AD, 12 controls) 450K array	118 DMPs associated with AD status were identified in the hippocampus AD-related DMPs were annotated to neurodevelopmental and neurogenesis-related genes and candidate “hotspots” such as *HAND2, HOXA3, HIST1H3E, NXN, PAX3, RBMS1*, and *RHOB*	(Altuna *at al*., 2019)
Middle temporal gyrus (MTG)	80 individuals (45 AD, 35 controls) for MTG 450K array with bisulfite (BS) and oxidative BS -converted DNA	1 DMR, 1 differentially hydroxymethylated region, and 11 differentially unmodified regions that were associated with AD annotated to genes including *OXT*	*(Lardenoije at al., 2019)*

Whole blood (WB)	99 individuals (42 converters, 44 controls, 13 non-converters but later diagnosis) for WB 450K array – BS only	Regional analysis identified 15 and 21 DMRs associated with conversion to AD at baseline and follow-up including *OXT* (blood-brain similarities)	
Whole blood	23 AD discordant twin pairs RRBS - Illumina HiSeq2500/3000	11 DMPs with consistent methylation differences in each zygosity group	(Konki *at al.*, 2019)
Anterior hippocampus	12 individuals (6 AD and 6 controls) RRBS - Illumina HiSeq2500/3000	Multiple sites identified including those annotated to genes previously linked to AD (e.g. *ADARB2;* located in an exon) identified in both blood and cortex	
Entorhinal cortex (EC) Superior temporal gyrus Cerebellum Striatum Substantia nigra	369 individuals (60 AD, 119 DLB, 27 VaD, 22 HD, 36 PD, and 105 controls) 96 (discovery), 104 (first replication), and 96 (second replication) brain samples covering the Braak NFT stage spectrum 450K array (on BS and oxidative BS -treated DNA)	Hypermethylation and hyperhydroxymethylation associated with elevated AD neuropathology in *ANK1* Identified significant DNA methylation changes in the EC in multiple diseases, including AD, HD, and PD, with significant DNA hypermethylation across the amplicon in AD and HD Results suggest previous methylation values may be cofounded by 5-hmC	(A. R. Smith, Smith, Burrage, *at al.*, 2019)
Frontal Cortex Temporal cortex Occipital cortex	128 individuals covering the Braak NFT stage spectrum (76 AD, 52 controls) Isolated neuronal and glial nuclei in 31 occipital cortex samples covering the Braak NFT stage spectrum 450K array	Differential methylation identified in known AD genes (e.g. *APP*, *HOXA*3 and ADAM17) Novel AD loci identified including *LRRC8B* and *MCF2L*	(Gasparoni *at al.*, 2018)

Whole blood	84 individuals (45 AD and 39 controls) EPIC array	477 DMPs 106 DMPs were annotated to genes (21 hypermethylated; 97 hypomethylated) A proportion of DMP-associated genes and their products previously implicated in LOAD pathogenesis including *B3GALT4, FLOT1, OXT,* and *DLG2*	(Madrid *at al.*, 2018)
Superior temporal gyrus Prefrontal cortex	147 covering the Braak NFT stage spectrum 450K array	Differentially methylation across the *HOXA* gene cluster (covering 48kb) Hypermethylation associated with elevated Braak NFT stage	(R. G. Smith *at al.*, 2018)
Hippocampus	5 individuals (3 AD and 2 controls) Reduced representation hydroxymethylation profiling (RRHP)	Pathway analysis implicated genes that play a role in the pathophysiology of LOAD including *CR1*, *BIN1*, and *CLU* (AD GWAS genes)	(Ellison, Bradley-Whitman, & Lovell, 2017)
Whole blood	12 individuals (4 AD, 4 MCI, and 4 controls) 450K array	11 DMPs identified which differentiates AD, MCI and control. Replicated and validated hypomethylation in *NCAPH2/LMF2*	(Kobayashi *at al.*, 2016)

Superior temporal gyrus	68 individuals (34 AD and 34 controls) 450K array	479 DMRs hypermethylated with AD pathology Hypermethylated DMRs preferentially overlapped promoter regions DMRs overlapped with genes involved in neuron function, development, and cellular metabolism and included genes previously reported in Alzheimer’s disease genome-wide and epigenome-wide association studies	(Watson *at al.*, 2016)
Dorsolateral Prefrontal cortex	708 brains (460 AD and 263 controls as discovery) 450K array Data from Lunnon *at al.* used as replication (Lunnon *at al.*, 2014)	Differential methylation at 71 CpGs associated with elevated AD pathology including CpGs in the *ABCA7* and *BIN1* region 11 DMPs were validated in the independent replication cohort Genes implicated include: *ANK1*, *CDH23, DIP2A, RHBDF2, RPL13, SERPINF1* and *SERPINF2*	(De Jager *at al.*, 2014)
Entorhinal cortex Superior temporal gyrus Prefrontal cortex Cerebellum	122 individuals (multiple tissue samples per individual) in discovery cohort with varying Braak NFT stage 450K array	Hypermethylation was associated with elevated Braak NFT stage in the *ANK1* region in AD cortex	(Lunnon *at al.*, 2014)
Prefrontal cortex	12 AD cases 12 controls 27K array	948 DNAm sites annotated to 918 genes associated with AD *THEM59* hypomethylated with AD - a gene involved in Aβ processing	(Bakulski *at al.*, 2012)

1.3.4 Epigenetic Variation associated with other types of dementia

The focus of most EWAS analyses of neurodegenerative disease has been on some aspect of AD, however several studies have explored associations between variable DNAm and PD (Chuang *at al.*, 2017; Masliah, Dumaop, Galasko, & Desplats, 2013), DLB (Sanchez-Mut *at al.*, 2016; Fernandez *at al.*, 2012) and FTD (Li *at al.*, 2014) in both brain and peripheral tissues (see **Table 1.7**). Although many of these studies are not well powered due to low sample numbers, particularly in disease groups, several DNAm sites have been associated with other neurodegenerative diseases. For example, a study conducted by Sanchez-Mut *at al.* (2016) focused on identifying common pathways involved in PD, DLB, AD and down-syndrome cases and identified multiple DMRs associated with disease. In their subsequent pathway analysis they reported significant over-representation in pathways related to brain function and immune response. They also identified that *ANK1* was differentially expressed in DLB, which corresponds with results from previous AD EWAS indicating there is hypermethylation of *ANK1* (De Jager *at al.*, 2014; Gasparoni *at al.*, 2018; Lunnon *at al.*, 2014; Smith *at al.*, 2019), although recent research reported that the hypermethylation of *ANK1* in DLB was likely confounded by AD pathology. A recent EWAS in peripheral blood of patients with FTD and progressive supranuclear palsy (PSP) compared to controls found a specific methylation signature associated with tauopathy, suggesting this signature as a risk factor for neurodegeneration and is not a specific pathway in AD. Due to the nature of DNAm, which is influenced by a variety of factors, it can be difficult to distinguish if associations are driven specifically by the disease itself (e.g. disease specific neuropathology), by some other aspect of neurodegeneration or due to other factors.

Table 1.7: Epigenetic studies in dementias other than LOAD. *Table adapted from MacBean at al. (2020).*

Phenotype	Tissue type	Methods	Results	References
Parkinson's disease	Whole blood	2,131 individuals (1,132 PD cases, 999 controls) 450K array	2 DMPs. Hypermethylation in PD is associated with down-regulation of the *SLC7A11* gene. Consistent with an environmental exposure (unlikely consequence of medications or genetic effects on DNA methylation)	(Vallerga *at al.*, 2020)
Parkinson's disease	Whole blood	232 PD samples followed up from baseline (197 European samples were the focus of analysis) Longitudinal study 450K array	Numerous significant DMPs in blood associated with declining cognition in PD including sites annotated to *KCNB1, DLEU2*, and *SATB1*	(Chuang *at al.*, 2017)
Dementia with Lewy bodies	Dorsolateral prefrontal cortex	107 individuals (AD, 5 DS-AD, 23 DLB, 15 PD, and 32 controls) 450K array	Identified differentially methylation in *ANK1* and *RHBDF2* in AD and *ANK1* in DLB Neurodegenerative disorders might have similar pathogenic mechanisms	(Sanchez-Mut *at al.*, 2016)
Frontotemporal Dementia	Whole blood	351 individuals (128 FTD, 43 PSP, and 185 controls) 450K array	Multiple DMPs were identified in both diseases. Three DMPs were located in 17q21.31 region – a reported PSP risk gene	(Li *at al.*, 2014)
Progressive supranuclear palsy (PSP)				
Parkinson's disease	Frontal cortex Whole blood	11 individuals (5 PD and 6 controls) 450K array	Differential methylation identified in both blood and cortex Four genes strongly associated with PD: *HLA-DQA1*, *GFPT2*, *MAPT*, and *MIR886*	(Masliah, Dumaop, Galasko, & Desplats, 2013)
Dementia with Lewy bodies	Cerebral cortex lesions	437 individuals (424 control tissues, 13 DLB) GoldenGate DNA methylation BeadArray	Differential methylation identified across the groups; methylation patterns were able to make a distinction between DLB samples from normal and cancer tissues	(Fernandez *at al.*, 2012)

1.4 Regulatory genomic variation and ageing biomarkers

Advancing age is associated with declining physical and cognitive function, and as previously mentioned (see section **1.1.3**) is a major risk factor for many human brain disorders including dementia and other neurodegenerative diseases (Harper, 2014; Sierra, 2019). Understanding the biological mechanisms involved in ageing will be a critical step towards preventing, slowing or reversing age-associated phenotypes. Due to the substantial inter-individual variation in age-associated phenotypes, there is considerable interest in identifying robust biomarkers of 'biological' age, a quantitative phenotype that is thought to better capture an individuals' risk of age-related outcomes than actual chronological age (Jylhävä, Jiang, Foebel, Pedersen, & Hägg, 2019). Several data modalities have been used to generate estimates of biological age; these include measures of physical fitness (e.g. muscle strength) (Sosnoff & Newell, 2006), cellular phenotypes (e.g. cellular senescence due to the deterioration and functional characteristics of cells which is most commonly caused by DNA double strand breaks) (Baker *at al.*, 2011), genomic changes (e.g. telomere length) (Jylhävä, Pedersen, & Hägg, 2017; Sanders & Newman, 2013) and epigenetic mechanisms (e.g. DNA methylation) (Horvath, 2013).

1.4.1 Epigenetic clocks

There has been recent interest in the dynamic changes in epigenetic processes over the life course, and a number of 'epigenetic clocks' based primarily on DNAm have been developed that appear to be highly predictive of chronological age (Hannum *at al.*, 2013; Horvath, 2013). The landmark DNAm clock was developed by Horvath (Horvath, 2013), who applied elastic net regression to Illumina DNAm array data from a large number of samples derived from a range of tissues (n = ~ 8,000 across 51 tissue and cell types), and generated a predictor based on DNAm at 353 CpG sites that is accurate at predicting chronological age (Horvath, 2013). Given that changes in DNAm are known to index exposure to certain environmental risk factors (for example, tobacco smoking) (Elliott *at al.*, 2014; Sugden *at al.*, 2019) that are associated with diseases of old age, and variable DNAm is robustly associated with a number of age-associated disorders (Chouliaras *at al.*, 2018; Chuang *at al.*, 2017; A. R. Smith *at al.*, 2016), there has been interest in the hypothesis that DNAm clocks might robustly quantify variation in biological age. Horvath's DNAm age clock, for example, has been widely applied to identify accelerated epigenetic ageing - where DNAm age predictions deviate from chronological age such that individuals appear older than they really are - in the context of numerous health and disease outcomes (Horvath & Ritz, 2015; Levine, Lu, Bennett, & Horvath, 2015; Marioni *at al.*, 2015; McCartney *at al.*, 2018). Although the original DNAm clocks were primarily developed to predict chronological age and are not robustly predictive of clinical health measures (e.g. blood pressure) (Quach *at al.*, 2017), more recent DNAm clocks such as Levine's 'pheno age' clock (Levine *at al.*, 2018) incorporate surrogate measures of biological age and are more directly aimed at predicting mortality and health-span.

1.4.2 Biases in existing epigenetic clocks

A strength of many existing epigenetic clocks is that they work relatively well across different types of sample; the Horvath multi-tissue clock, for example, can accurately predict age in multiple tissues across the life-course. Importantly, as with any predictor, the composition of the training data used to develop the clock influences the generality of the model. To date, there has been limited research comparing the prediction accuracy and potential bias of existing clock algorithms across different tissues and ages. Recent analyses have highlighted potential biases when using Horvath's clock in older samples (>~60 years) and in samples derived from certain tissues, especially the central nervous system (El Khoury *at al.*, 2019). This is important for the interpretation of studies of possible relationships between accelerated epigenetic age and age-related diseases affecting the human brain (e.g. neurodegenerative phenotypes); reported associations between accelerated DNAm age and disease may actually be a consequence of fitting a suboptimal predictor to available datasets. Potential confounders include differential changes in DNAm with age across tissues and the age distribution of the samples used to train existing classifiers. Resolution of these biases requires the construction of specific DNAm clocks developed using data generated on the relevant tissue-type and including broad representation of the age spectrum they will be used to interrogate. Recently, a number of tissue-specific DNA methylation clocks have been described, including clocks designed for whole blood (Hannum *at al.*, 2013; Zhang *at al.*, 2019), muscle (Voisin *at al.*, 2020), bone (Gopalan, Gaige, & Henn, 2019) and paediatric buccal cells (McEwen *at al.*, 2019). Importantly, although these DNAm age estimators have increased predictive accuracy within the specific tissues in which they were built, they lose this precision when applied to other tissues (Zhang *at al.*, 2019). A key aim of **Chapter 3** in my thesis was the development of a novel epigenetic clock specifically calibrated for use on human cortex tissue.

1.4.3 Age acceleration in neurodegenerative phenotypes

Since age is a major risk factor for dementia and other neurodegenerative brain disorders, there is particular interest in the application of epigenetic clock algorithms to these phenotypes, especially as differential DNAm in the cortex has been robustly associated with diseases including AD and PD (see **Table 1.6** and **Table 1.7**)(Lunnon *at al.*, 2014; A. R. Smith *at al.*, 2016; Yu *at al.*, 2015). Recent studies have reported an association between accelerated DNAm age and specific markers of AD neuropathology in the cortex (e.g. neuritic plaques, diffuse plaques and Aβ load) (Levine *at al.*, 2018, 2015). Furthermore, among individuals with AD, DNAm age acceleration is associated with declining global cognitive functioning and deficits in episodic and working memory (Levine *at al.*, 2018, 2015). These studies infer DNAm age is important to consider when conducting DNAm studies in neurodegenerative diseases. A key aim of this thesis is the development of a novel DNAm clock specifically calibrated for the human cortex and its application to measures of neuropathology and AD (see **Chapter 3**).

1.5 Genetic influences on DNA methylation

Although lifestyle factors and environmental influences such as disease status, nutrition, ageing, stress, and chemical exposures can alter epigenetic marks, the epigenome can also be directly influenced by genetic factors. This is further supported by evidence demonstrating that DNA methylation is heritable at a large proportion of sites, with the average heritability at each DNAm site ranging from 16-20% (Hannon *at al.*, 2018; McRae *at al.*, 2014, 2018).

There are several hypotheses for how genetic-epigenetic mediation occurs. For example, alterations in the genetic sequence have the potential to influence the formation of epigenetic marks, such as a SNP which is located within a CpG site could create or remove that specific site; SNPs could cause direct changes to epigenetic machinery; and more recently there have been studies which suggest global genetic risk influences molecular markers (Hannon *at al.*, 2016, 2018; Viana *at al.*, 2017). Recent studies have sought to better understand the molecular mechanisms underlying disease phenotypes by using integrated omic methods and literature suggests that the genetic mediation of the methylome provides a link between genetic variation and complex phenotypes (Wagner *at al.*, 2014).

1.5.1 Genetic effects on DNAm at specific sites in the genome

There has been recent interest investigating the relationship between genetic and epigenetic variation at the individual level, there have been studies which have examined the effect individual SNPs have on molecular processes (Hannon *at al.*, 2018, 2019; Hannon, Weedon, Bray, O'Donovan, & Mill, 2017; Liu, Wang, Jing, Meng, & Yang, 2021; Zhao, Hu, Zang, & Wang, 2019). Genetic variants that are associated with DNAm at CpG sites are defined as methylation quantitative trait loci (mQTLs; see **Figure 1.14**). Most current mQTL databases have been generated using 450K data (Hannon, Weedon, Bray, O'Donovan, & Mill, 2017; McRae *at al.*, 2018). In this thesis, I generate two novel mQTL databases utilising EPIC array data: for whole blood and for the cortex (see **Chapter 5**). Utilising the EPIC array substantially increases the number of sites across the genome in which we can identify mQTLs. The cortex mQTL dataset will represent the largest EPIC mQTL database currently generated for that tissue. This is a key aim of **Chapter 5**.

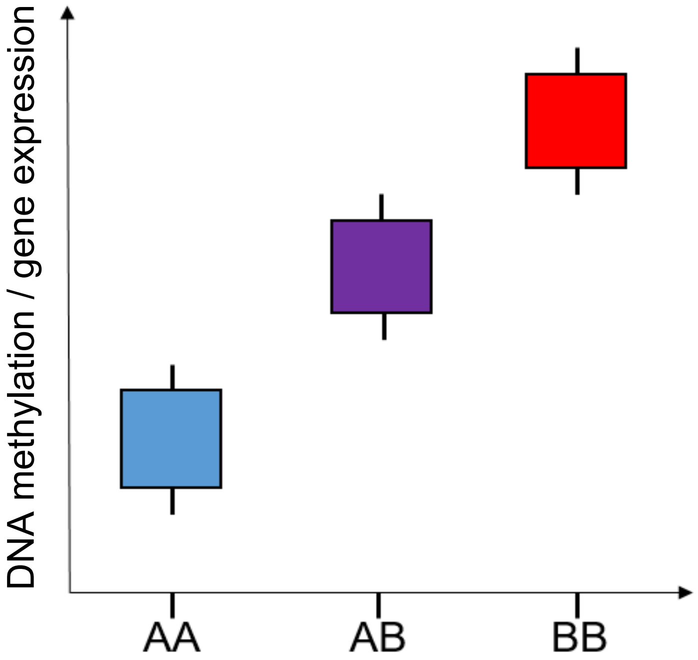


Figure 1.14: mQTLs and eQTL share the underlying principle that SNPs have an effect on either DNA methylation (mQTL) or gene expression (eQTL). *This is a hypothetical QTL example and it shows that being homozygous for the B allele results in increased DNA methylation/ gene expression in comparison to the homozygous A allele.*

1.5.1.1 Leveraging mQTLs to fine-map genomic regions associated with complex disease

As many GWAS variants associated with complex traits reside in non-coding regions, it has been challenging to identify which genes are relevant for disease aetiology (Amlie-Wolf *at al.*, 2018; Giambartolomei *at al.*, 2018). A potential approach to explore the mechanisms by which non-coding risk variants regulate gene expression is through integration of datasets that measure the association of molecular phenotypes including mQTLs and gene expression quantitative trait loci (eQTLs; where a SNP is associated with gene expression). If the same genetic variant is driving the association signal in the GWAS but is also driving expression at close-by DNAm and gene expression sites, then this could be indicative of a putative disease mechanism. eQTL studies have been conducted looking at the association between eQTLs and disease, highlighting a causal gene and the tissue in which the effect it mediated (He *at al.*, 2013; Nica *at al.*, 2010). Identifying overlap between complex disease-associated variation and eQTL variants has provided evidence of shared molecular mechanisms.

However these earlier studies, such as the one conducted by Nica and colleagues (2010), did not formally test the null hypothesis for co-localisation and the methodology was instead based on the residual association of the most significant GWAS SNP. A formal test of colocalisation was since developed based on a regression framework which consists of testing the null hypothesis of proportionality of regression coefficients for two traits across a set of SNPs – this assumption will be met if these two traits share causal variants. This method makes the assumption that if there are multiple causal variants then they are all shared. However, since this method relies on specifying the subset of SNPs to be included in analysis, biases can arise due to over estimation of the effect sizes of the tested SNPs, a concept known as ‘the winners curse’ (Wallace, 2013); this notion states that the most associated SNP identified in a GWAS may in fact not be the causal SNP, which can lead to the statistical rejection of co-localisation in situations where the causal SNP is shared.

Giambartolomei and colleagues (2014) developed a Bayesian test for colocalisation between pairs of genetic association studies which utilises summary statistics. Their model is related to a method first developed by Flutre *at al.* (Flutre, Wen, Pritchard, & Stephens, 2013) who focussed on maximising power to discover eQTLs in expression datasets across multiple different tissues. Giambartolomei’s method provides posterior probabilities and only requires the input of single SNP p-values and their minor allele frequencies. Their framework tests five different hypotheses: 1) There is no association with either trait; 2) There is an association with trait one but not with trait two; 3) There is an association with trait two, but not with trait one; 4) There is an association with trait one and trait two but these are two independent SNPs; and 5) There is an association with trait one and trait two, and there is a single causal SNP.

Recent research by Hannon and colleagues (2018) integrated the external information from mQTLs and eQTLs to identify which genetic signals are explained by regulatory effects and applied the Bayesian framework for co-localisation developed by Giambartolomei and colleagues to their data. In this study they characterised mQTLs in a collection of 166 human foetal brain samples and identified >160,000 mQTLs. They found that the foetal brain mQTLs were enriched in risk loci identified in a Schizophrenia (SCZ; a psychiatric disorder with a neurodevelopmental component) GWAS. They utilised the mQTLS to refine GWAS signals and identified discrete sites of regulatory variation associated with the SCZ variants.

More recently, another method called Summary-data based Mendelian Randomisation (SMR) (Zhu *at al.*, 2016) has been developed to identify whether there is pleiotropy between genetic variants and molecular markers (DNAm and gene expression). This method is based on the premise of Mendelian randomisation (MR): an approach which uses genetic variation as a natural experiment to investigate the ‘causal’ relationships between phenotypes in observational data (Burgess, Dudbridge, & Thompson, 2016). In other words, if the expression or DNAm of a gene is influenced by a SNP (an eQTL/ mQTL), then there will be various levels of gene expression/ DNAm among individuals who are homozygous or heterozygous for specific genetic variants (e.g. AA – low expression/ DNAm, AB – median expression/ DNAm, BB – high expression/ DNA). If the expression/ DNAm levels have a significant effect on a trait then these differences will be observable in the three distinct genetic groups (see **Figure 1.14**). In MR analysis a genetic instrument is used (e.g. a SNP) to test for the causative effect of an exposure (e.g. gene expression) on an outcome (a phenotype). However, given that GWAS have identified thousands of SNPs associated with complex traits, a single common SNP will have little effect on a trait. Subsequently, Zhu and colleagues (Zhu *at al.*, 2016) developed SMR to integrate summary-level data (e.g. effect sizes) from GWAS as well as eQTL and mQTL studies. The aim of this method is to identify genes whose expression levels are associated with a trait due to pleiotropy (see **Figure 1.15**). SMR has been applied to multi-omics datasets. For example, Wu *at al.* (2018), performed an integrative analysis using summary level SNP data from multi-omic studies to identify if DNAm sites are associated with gene expression and a phenotype through shared causal effects. They identified pleiotropic associations between >7800 DNAm sites and >2700 genes and found that the DNAm sites were enriched in promoters and enhancers. Their next analysis linked both the transcriptome and the methylome to twelve traits and identified 149 DNAm sites and 66 genes. These results indicate there is a mechanism where a SNP has an effect of a trait mediated though genetic regulation of DNAm and gene expression. Similarly, Hannon and colleagues (2018) utilised the SMR tool to characterise the relationship between genetic, epigenetic and transcriptomic variation in > 60 traits and identified ~1700 pleiotropic associations between 36 complex traits and >1200 DNAm sites. They also identified ~6,800 pleiotropic associations between >5,400 DNAm sites and the transcription of >1700 genes.

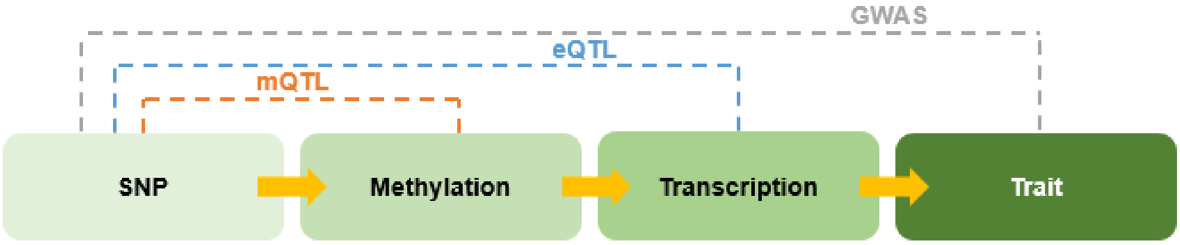


Figure 1.15: Schematic of integrative analysis utilizing multi-omics data. *A hypothetical model of a mediation mechanism tested in SMR analysis: an SNP exerts an effect on the phenotype by altering the DNAm level, which regulates the expression levels of a functional gene. Figure and legend adapted from Wu at al. (2018).*

These studies have characterised the relationship between genetic, epigenetic and transcriptomic variation using co-localisation analyses and Mendelian randomisation based methods, fine-mapping regulatory variation involved in disease pathology, increasing understanding of the mechanisms leading to disease phenotypes (Giambartolomei *at al.*, 2018; Hannon *at al.*, 2016). I will extend these analyses and apply them to cortex and whole blood datasets (see **Chapter 5**).

1.5.1.2 Genetic-epigenetic variation at a global level

In addition to identifying associations between genetic and epigenetic variation at the individual level, there have been studies which have examined the effect global genetic risk has on molecular processes. Genetic variation mediating epigenetic variation has been observed in complex diseases of the brain including schizophrenia (SCZ) and autism spectrum disorder (ASD) (Hannon *at al.*, 2016, 2018; Viana *at al.*, 2017). Recent studies have quantified global genetic risk into PRS and associated this with genome-wide DNAm to further explore the molecular genomic mechanisms involved in disease pathogenesis. Recent research by Hannon *at al.*, explored this hypothesis; they calculated PRS for 639 individuals within in a case/control SCZ cohort (Hannon *at al.*, 2016). PRS were significantly higher in cases compared to controls and significantly predicted disease status. Subsequently, they conducted an epigenome-wide association study (EWAS) of SCZ PRS against genome-wide DNAm identifying multiple DMPs associated with disease status (Hannon *at al.*, 2016). PRS EWAS relationships in brain disorders has also been investigated in peripheral tissues. Hannon and colleagues conducted an autism spectrum disorder (ASD) – PRS study whereby they quantified neonatal methylomic variation from archived blood spots in 1263 infants (50% later developed ASD) and identified multiple sites where the ASD PRS was associated with hypermethylation. PRS have proven a useful tool for associating the genetic contribution of a disease - essentially as a biomarker - with epigenetic variation, enabling the exploration of molecular genomic mechanisms driving disease pathogenesis. The dynamic nature of epigenetic processes including DNAm means that unlike in genetic epidemiology a range of potentially confounding factors (e.g. medication, other environmental toxins, and reverse-causality) need to be taken into consideration. PRS-associated epigenetic variation is potentially less affected by these factors which are associated with the disease itself.

1.6 General aims of my thesis

There is a growing body of evidence highlighting the functional complexity of the genome and how genetic and epigenetic mechanisms play a role in the aetiology of AD and other forms of dementia. There is limited literature investigating the relationship between genetic and molecular mechanisms (e.g. DNAm and gene expression) in these diseases and therefore, within my thesis I aim to expand our knowledge in this area. There are three overarching aims of my thesis:

1) To identify if variable DNAm is associated with measures of neuropathology that are hallmarks of different types of dementia, with a particular focus on LOAD.

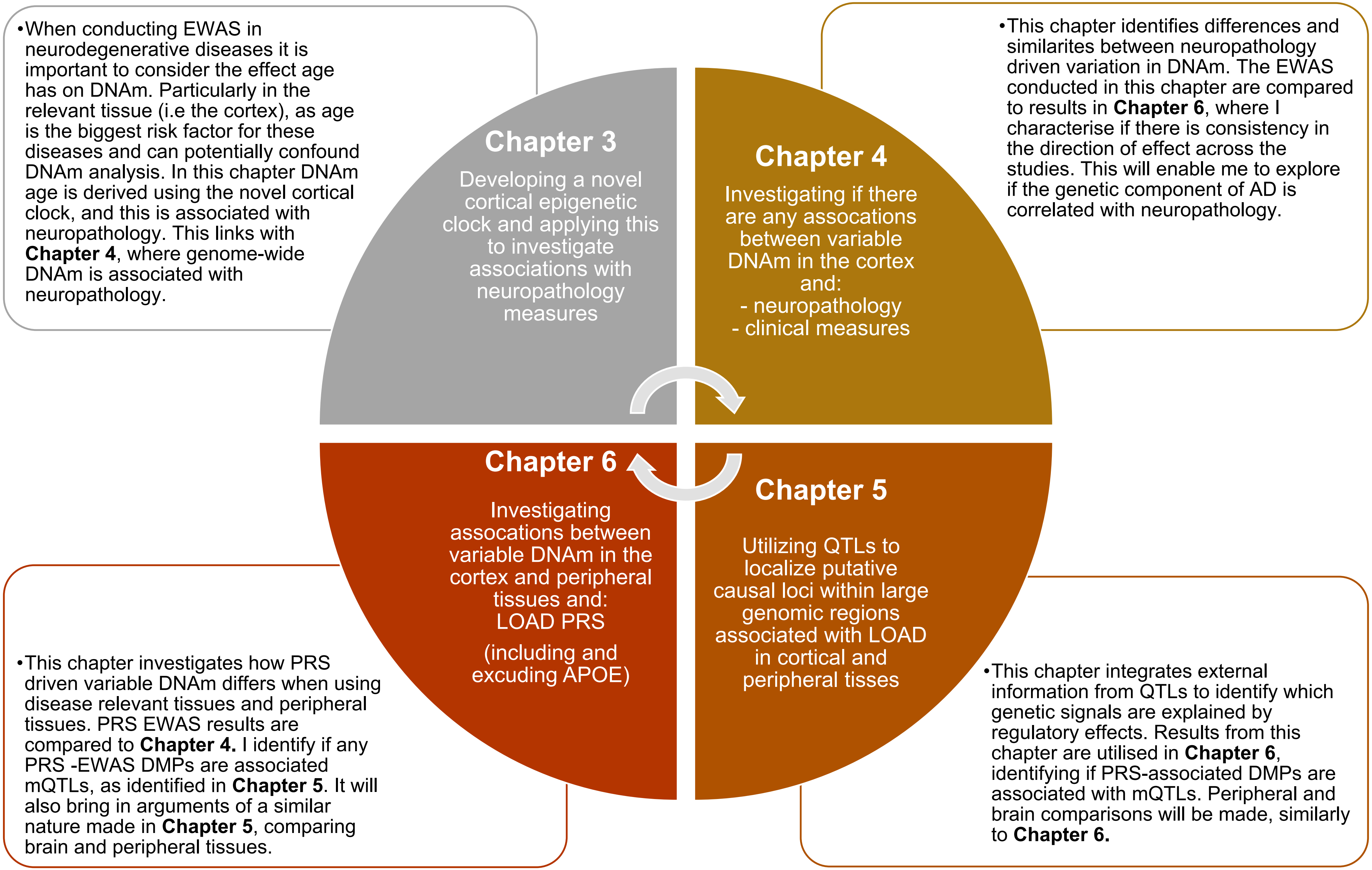
2) To investigate how polygenic burden for LOAD influences genomic regulatory processes (i.e. epigenomic and transcriptomic variation) across multiple tissues and how these results compare with disease driven DNAm variation.

3) To functionally annotate genomic regions associated with LOAD to fine-map regulatory variation involved in neuropathology, to help refine the genes involved in disease pathogenesis.

These aims are addressed across four empirical chapters:

- In my first empirical chapter (**Chapter 3**), I develop a novel DNAm clock that is specifically designed for application in DNA samples isolated from the human cortex and is accurate across the lifespan including in tissue from older donors. I build an epigenetic clock which minimises the potential for spurious associations with ageing phenotypes (e.g. neurodegenerative diseases) relevant to the brain.
- In my second empirical chapter (**Chapter 4**) I investigate if variable DNAm in the cortex is associated with neuropathology and clinical measures of dementia.
- In my third empirical chapter (**Chapter 5**) I utilise methylation and expression quantitative trait loci to localize putative causal loci within large genomic regions associated with LOAD in both brain and peripheral tissues.
- In my fourth empirical chapter (**Chapter 6**) I first investigate if PRS and *APOE* genotype are associated with neuropathology. Second, I investigate if variable DNAm in cortex and peripheral tissues is associated with polygenic risk for LOAD. I compare advantages and disadvantages of using brain (i.e. the most relevant for disease) and peripheral tissues for PRS analysis in EWAS.

Figure 1.16 describes the integration of my empirical chapters of my thesis.



- When conducting EWAS in neurodegenerative diseases it is important to consider the effect age has on DNAm. Particularly in the relevant tissue (i.e the cortex), as age is the biggest risk factor for these diseases and can potentially confound DNAm analysis. In this chapter DNAm age is derived using the novel cortical clock, and this is associated with neuropathology. This links with **Chapter 4**, where genome-wide DNAm is associated with neuropathology.

Chapter 3

Developing a novel cortical epigenetic clock and applying this to investigate associations with neuropathology measures

Chapter 4

Investigating if there are any assocations between variable DNAm in the cortex and:
- neuropathology
- clinical measures

- This chapter identifies differences and similarites between neuropathology driven variation in DNAm. The EWAS conducted in this chapter are compared to results in **Chapter 6**, where I characterise if there is consistency in the direction of effect across the studies. This will enable me to explore if the genetic component of AD is correlated with neuropathology.

Chapter 6

Investigating assocations between variable DNAm in the cortex and peripheral tissues and:
LOAD PRS

(including and excuding APOE)

Chapter 5

Utilizing QTLs to localize putative causal loci within large genomic regions associated with LOAD in cortical and peripheral tisses

- This chapter investigates how PRS driven variable DNAm differs when using disease relevant tissues and peripheral tissues. PRS EWAS results are compared to **Chapter 4.** I identify if any PRS -EWAS DMPs are associated mQTLs, as identified in **Chapter 5**. It will also bring in arguments of a similar nature made in **Chapter 5**, comparing brain and peripheral tissues.

- This chapter integrates external information from QTLs to identify which genetic signals are explained by regulatory effects. Results from this chapter are utilised in **Chapter 6**, identifying if PRS-associated DMPs are associated with mQTLs. Peripheral and brain comparisons will be made, similarly to **Chapter 6.**

Figure 1.16: Integration of Chapters 3 to 6 of this thesis. *EWAS = epigenome-wide association study; DNAm = DNA methylation; DMP = differentially methylated position; QTL = quantitative trait loci; mQTL = DNA methylation QTL; PRS = polygenic risk score.*

2 General Methods

In this Chapter I describe general methods which are used across multiple chapters in my thesis. This includes in-depth descriptions of quality control pipelines, array chemistry and general statistical methods. Additional methods specific to each individual study are provided in the relevant chapter(s).

2.1 DNA Methylation Profiling

The analysis of sodium bisulfite ($NaHSO_3$) treated DNA is currently the gold-standard method for quantifying DNA methylation (DNAm); it is efficient and provides quantification at the base-pair resolution (Y. Li & Tollefsbol, 2011). This process involves deamination of the non-methylated cytosines to uracil, which is then replaced by a thymine molecule in downstream procedures including the polymerase chain reaction (PCR). Cytosines which are methylated are protected from deamination, and therefore remain as cytosine (Y. Li & Tollefsbol, 2011). This enables the level of methylated and unmethylated cytosines at individual sites in the genome to be quantified directly. The step-by-step process of sodium bisulfite conversion is shown in **Figure 2.1**(Biolabs, 2016).

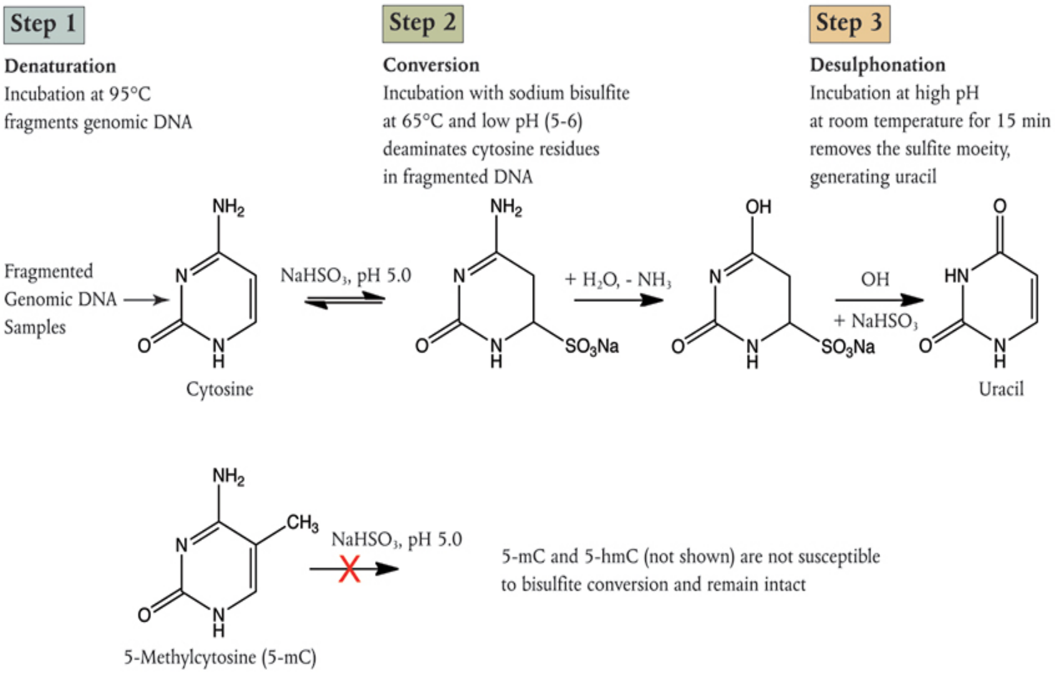

Step 1
Denaturation
Incubation at 95°C fragments genomic DNA
Step 2
Conversion
Incubation with sodium bisulfite at 65°C and low pH (5-6) deaminates cytosine residues in fragmented DNA
Step 3
Desulphonation
Incubation at high pH at room temperature for 15 min removes the sulfite moeity, generating uracil
Fragmented Genomic DNA Samples
NaHSO₃, pH 5.0
+ H₂O, - NH₃
OH
+ NaHSO₃
Cytosine
Uracil
NaHSO₃, pH 5.0
5-mC and 5-hmC (not shown) are not susceptible to bisulfite conversion and remain intact
5-Methylcytosine (5-mC)

Figure 2.1: DNA sodium bisulphite treatment. *Taken from New England Biolabs (Ipswich, MA, USA) webpage (Biolabs, 2016).*

The samples in the DNAm datasets I worked with in my thesis underwent sodium bisulfite conversion using the EZ DNAMethylation-Gold Kit (Zymo Research, Irvine, CA, USA), according to manufacturer's instructions. After sodium bisulfite conversion the human DNA samples were profiled using Illumina BeadChip DNA methylation arrays.

2.1.1 Infinium HumanMethylation450 BeadChip

Several of the cohorts in this thesis (used in **Chapter 3** and **Chapter 6**) were profiled using the Illumina Infinium HumanMethylation450 BeadChip (450K array), which were scanned on an iScan Microarray Scanner (Illumina, SanDiego, CA, USA) using the manufacturer's instructions. The Illumina 450K array quantifies DNAm at 484,577 DNAm sites and coves 99% of RefSeq genes (https://www.ncbi.nlm.nih.gov/refseq/ (O'Leary *at al.*, 2016)) as well as regulatory regions including CpG islands (96% covered), 5′ and 3′ UTRs, island shores, island shelves, promoters and gene bodies (Bibikova *at al.*, 2011). The array combines two technically distinct assays: 1) the Infinium I assay (type I probes) and 2) the Infinium II assay (type II probes) (M. E. Price *at al.*, 2013). Type I probes employ two probes per CpG locus – one specific to methylated DNA and one specific to unmethylated DNA (see **Figure 2.2**). The 3' terminus of each of these probes matches either the protected cytosine (methylated) or the thymine based which arose from sodium bisulfite conversion (unmethylated). The design of type I probes is based on the assumption that methylation is regionally correlated with a 50bp span (i.e. CpGs within this region are correlated with the query CpG) (Shoemaker, Deng, Wang, & Zhang, 2010). Type II probes are characterised by one probe per locus (see **Figure 2.2**), with the DNAm state determined at the single base extension after hybridisation of two dyes (red = unmethylated, green=methylated). The requirement for a single bead type enables more DNAm sites to have their methylation state quantified.

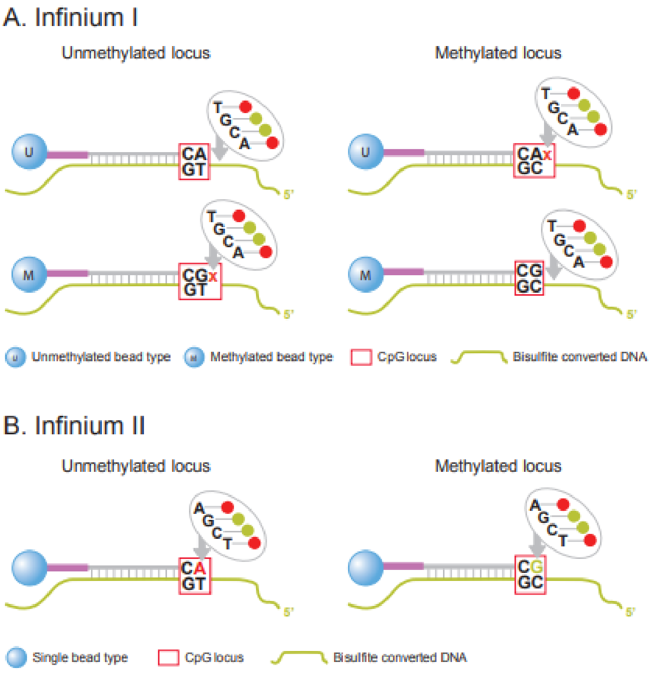


Figure 2.2: Type I and type II probes. *Figure and legend downloaded the Illumina (San Diego, CA, USA) webpage (Illumina, 2015). BeadChips employ both Infinium I and Infinium II assays, enhancing their breadth of coverage. A) Infinium I assay design employs two bead types per CpG locus, one each for the methylated and unmethylated states. B) The Infinium II design uses a single bead type, with the methylated state determined at the single base extension step after hybridization.*

2.1.2 Infinium HumanMethylationEPIC BeadChip

The other cohorts used in this thesis (used in **Chapter's 3-6**) were profiled using the latest Illumina BeadChip technology, the Illumina Infinium HumanMethylationEPIC BeadChip (EPIC array) and were scanned on an iScan Microarray Scanner (Illumina, SanDiego, CA, USA) using the manufacturer's instructions. The EPIC array quantifies DNAm at 866,836 DNAm sites, including 90% of the 450K array probes and an additional 413,743 DNAm sites (95% of these are type II probes and mainly target gene body, intergenic and non-CpG island regions). Overall, the array has good coverage across important regulatory regions such including CpG islands, 5′ and 3′ UTRs, island shores, island shelves, promoters and gene bodies.

2.1.3 Quantifying DNA methylation

In order to quantify DNAm at each site, a ratio of the fluorescence intensity for methylated (M) and unmethylated (U) signal is used to produce a 'β value' for each site, which ranges from 0 (all DNA alleles at that DNAm site are unmethylated) to 1 (all DNA alleles at that DNAm site are methylated). The equation for the β value is shown below, where 'M' = intensity methylated, 'U' = intensity unmethylated, and 'a' = 100 (this is added to M + U to stabilise the beta values if both M and U are small) (Weinhold, Wahl, Pechlivanis, Hoffmann, & Schmid, 2016).

$$\beta = \frac{\mathrm{M}}{M + U + a}$$

2.1.4 Quality control of DNA methylation data

I conducted quality control (QC) for multiple large DNAm datasets throughout my PhD including the Brains for Dementia research (BDR) cohort (DNAm data used in **Chapter 3**-**6**) and the Exeter Ten Thousand (EXTEND) cohort (DNAm data used in **Chapter 5-6**). Our group previously developed the *wateRmelon* (Pidsley *at al.*, 2013) and *BigMelon* (Gorrie-Stone *at al.*, 2019) packages for conducting array QC. There are several steps we take to ensure we have high-quality data which I explain in more detail below, using examples from the EXTEND and BDR QC pipeline.

a) Check Signal Intensities

The intensity check is the biggest indicator of sample and raw data quality. The median methylated signal intensity and unmethylated signal intensity for each sample is calculated (see **Figure 2.3**). Although an arbitrary threshold of 2000 is recommend by Illumina, the threshold can be adjusted for each dataset and visually ascertained when plotting M against U (see **Figure 2.4**.), removing samples which clearly deviate from the main cluster.

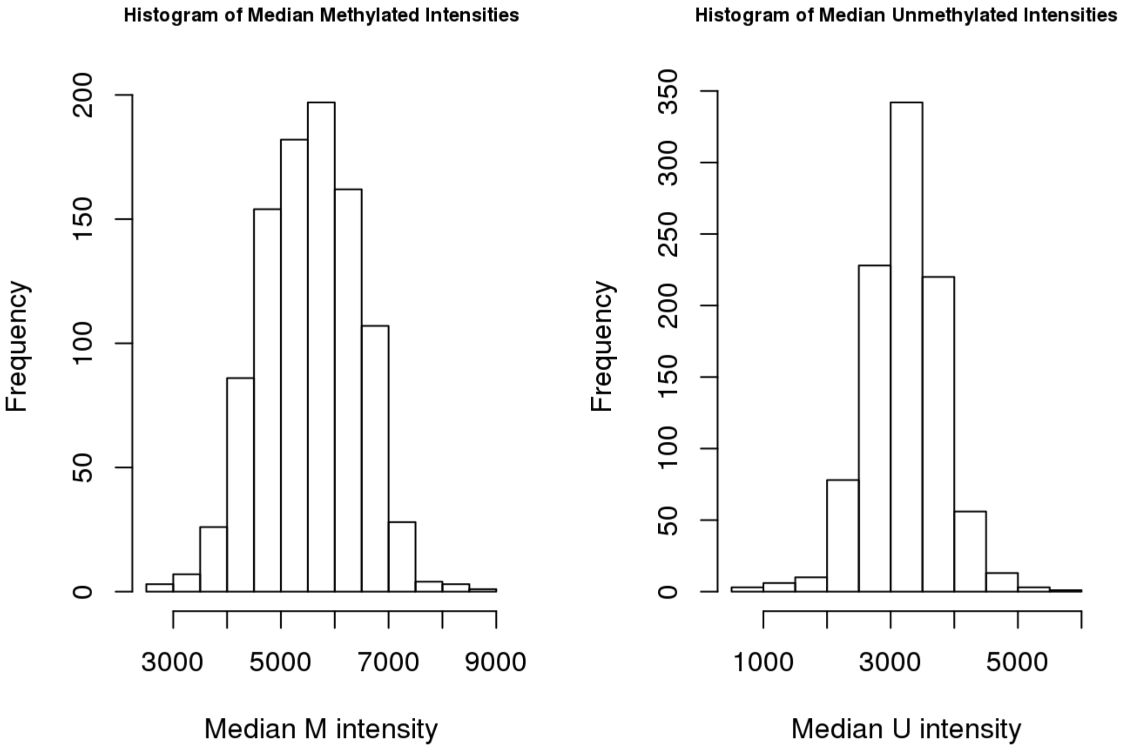


__Figure 2.3: Signal intensities of the EXTEND cohort.__ Histograms of median methylated and unmethylated signal intensities.

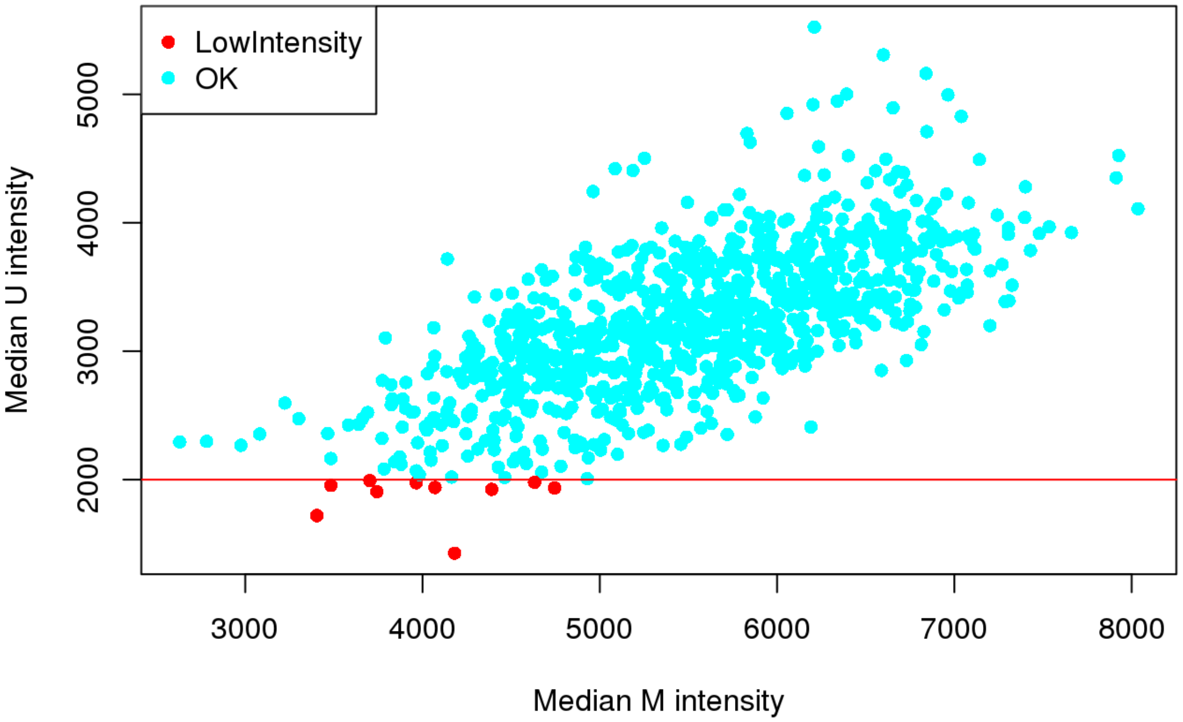


Figure 2.4: EXTEND samples with a signal intensity < 2000 for either M or U were removed based on the Illumina threshold.

b) Bisulfite conversion statistic

A bisulfite conversion statistic for each sample is calculated using the *bscon* function from *wateRmelon*, and a histogram of the results are plotted (see **Figure 2.5**). Samples with a conversion rate <80% fail this QC step and are excluded.

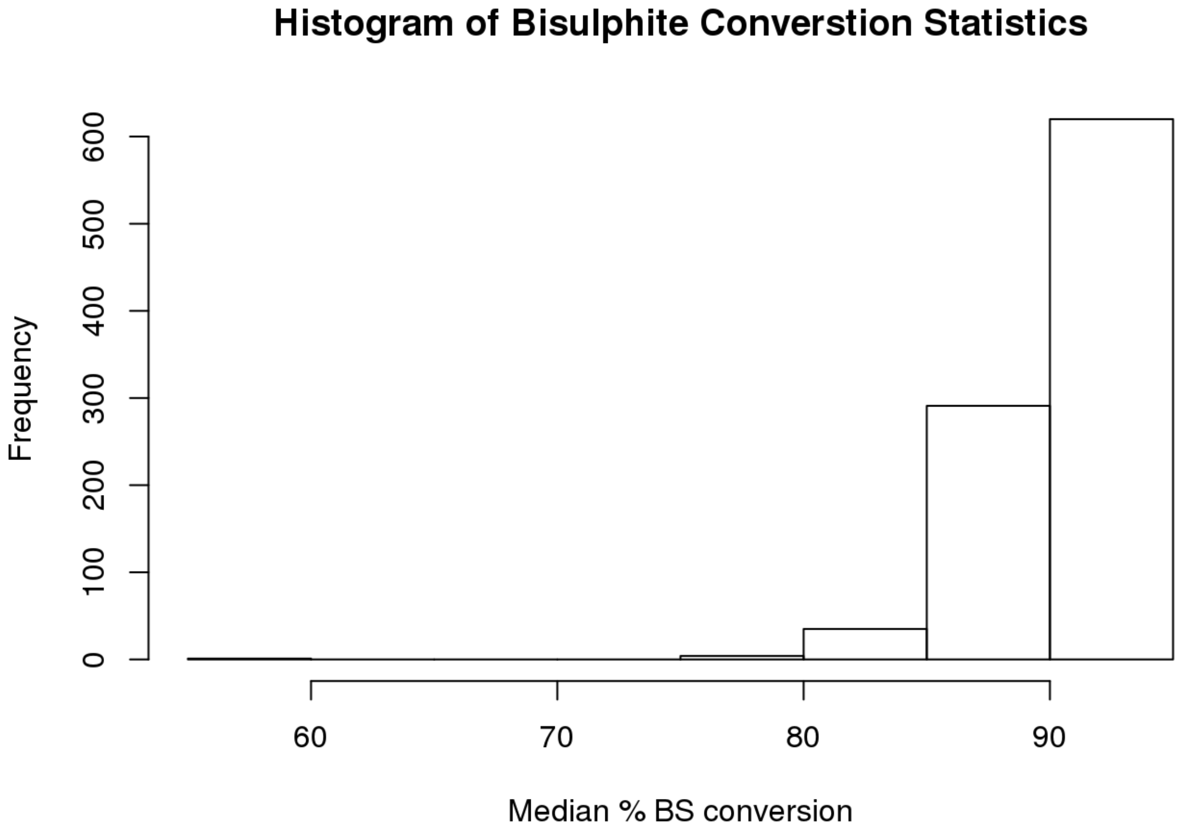


Figure 2.5: Histogram of Bisulphite conversion statistics for EXTEND.

c) Check Sex

A principal component analysis (PCA) of the DNAm data is used to confirm the sex of the samples. The principal components are calculated, and the two which correlate overall most with sex are found. These can be used to generate a scatter plot where the sexes are clearly separated. Samples with mismatched sex (i.e. observed sex is different from the expected/reported sex) are removed (see **Figure 2.6**).

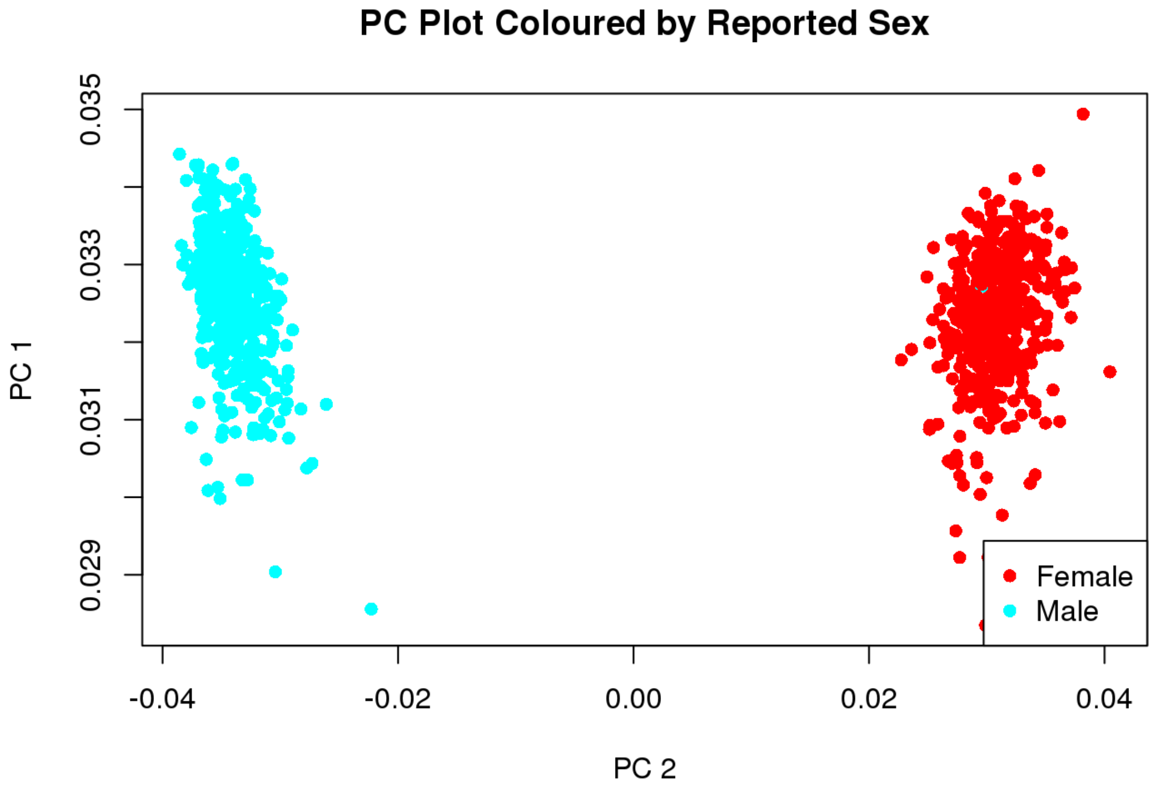


Figure 2.6: PC1 plotted against PC2 coloured by reported sex.

d) Check samples match their genotype data

On the EPIC array there are 59 SNP probes and on the 450K array there are 65 SNP probes – these are selected from high-quality SNP probes included on Illumina genotyping arrays and are useful for sample ID tracking. If we have genotype data for these samples, we can compare the methylation on these SNP probes to the known genotypes to confirm they are from the expected individual. Correlations are calculated, and samples with a correlation <0.9 are excluded (see **Figure 2.7**).

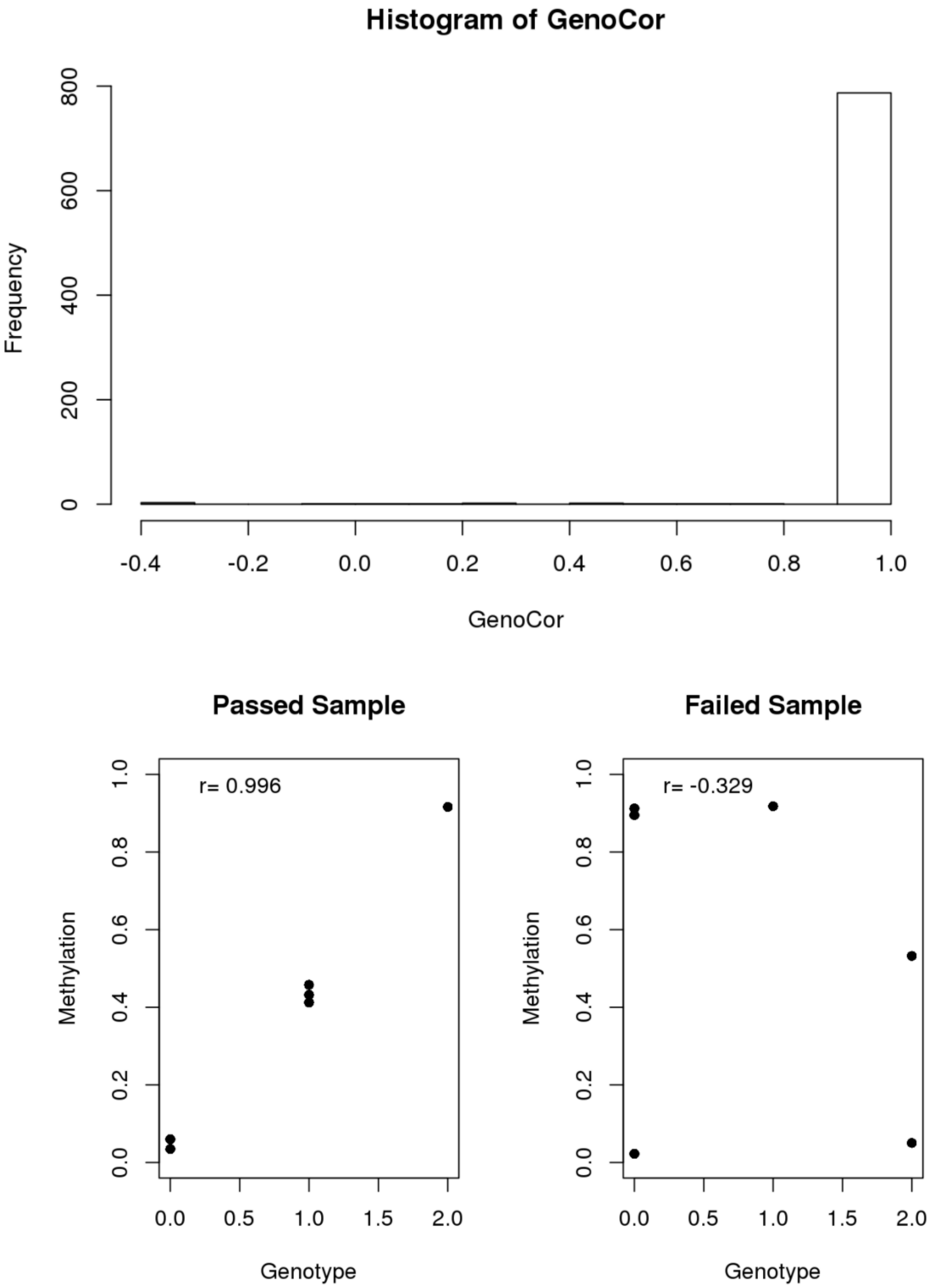


Figure 2.7: Check if EXTEND samples match their genotype data. *Top: correlations between genotype and DNAm data based on 59 SNPs on the EPIC array. Bottom: an example of a passed and failed sample.*

e) Check samples match both tissues

In some datasets there will be more than one sample per individual (e.g. in BDR we profiled both occipital cortex [OCC] and prefrontal cortex [PFC] samples from each donor) and the SNP probes can be used to confirm samples match up. A histogram of results is plotted (see **Figure 2.8**) and matched samples which have a correlation < 0.9 fail the QC and are excluded.

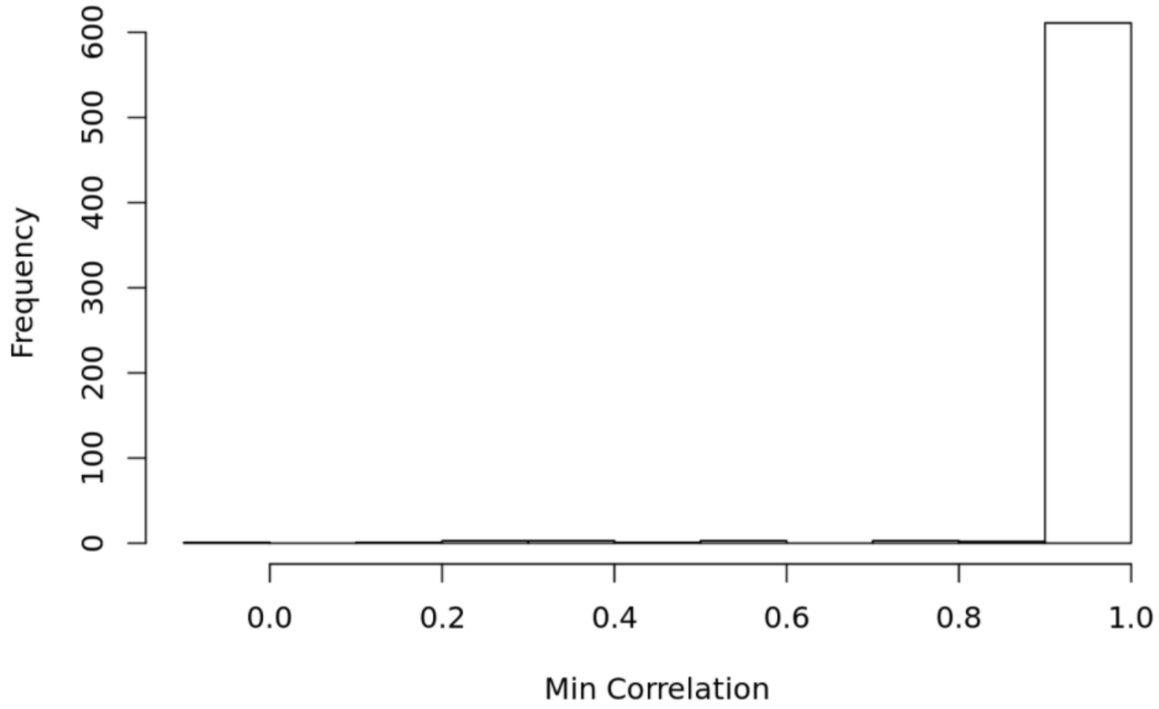


Figure 2.8: Correlation between samples from the same individual in the BDR cohort.

f) Check for duplicate samples

The SNP probes on the DNA methylation arrays can also be used to estimate genetic correlations between samples. This small number of probes means that only identical samples (such as samples from the same individual, monozygotic twins or samples duplicated by error) can be identified and no lower proportion of genetic relatedness (e.g. siblings) can be inferred. Since we expect all samples to be unrelated in the cohorts included in my thesis, all genetic correlations between samples should be low (approximately < 0.8). For each sample we can find the maximum correlation with any other sample and plot a histogram of the results (see **Figure 2.9**). Samples with a correlation > 0.8 are excluded.

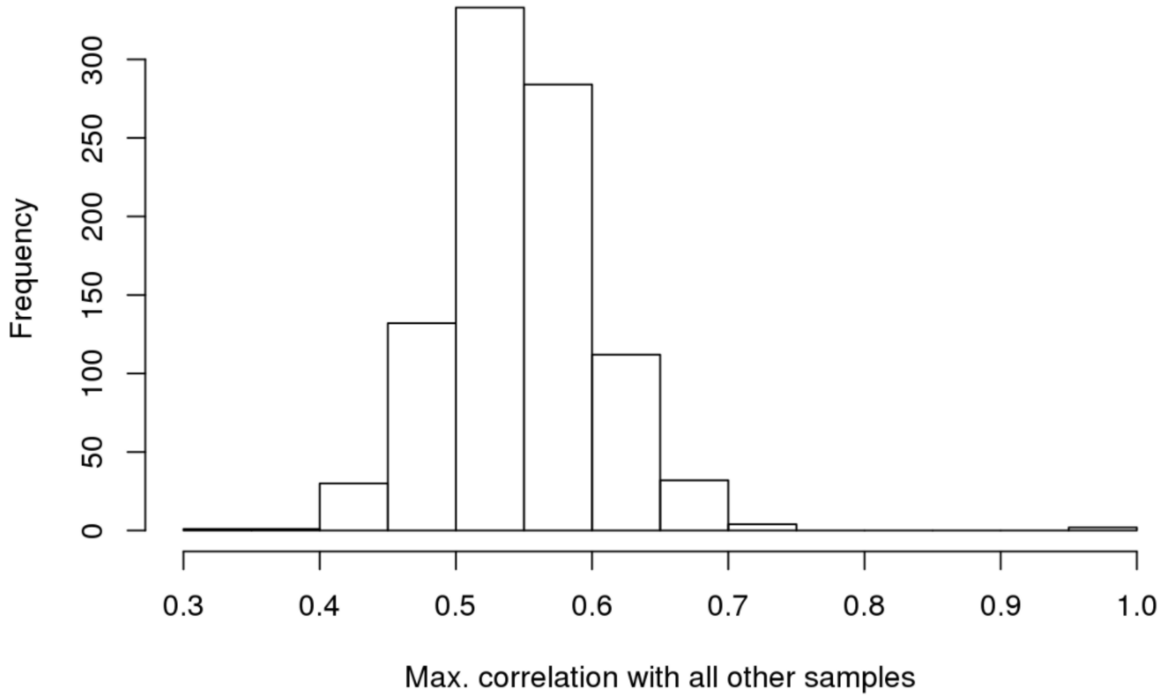
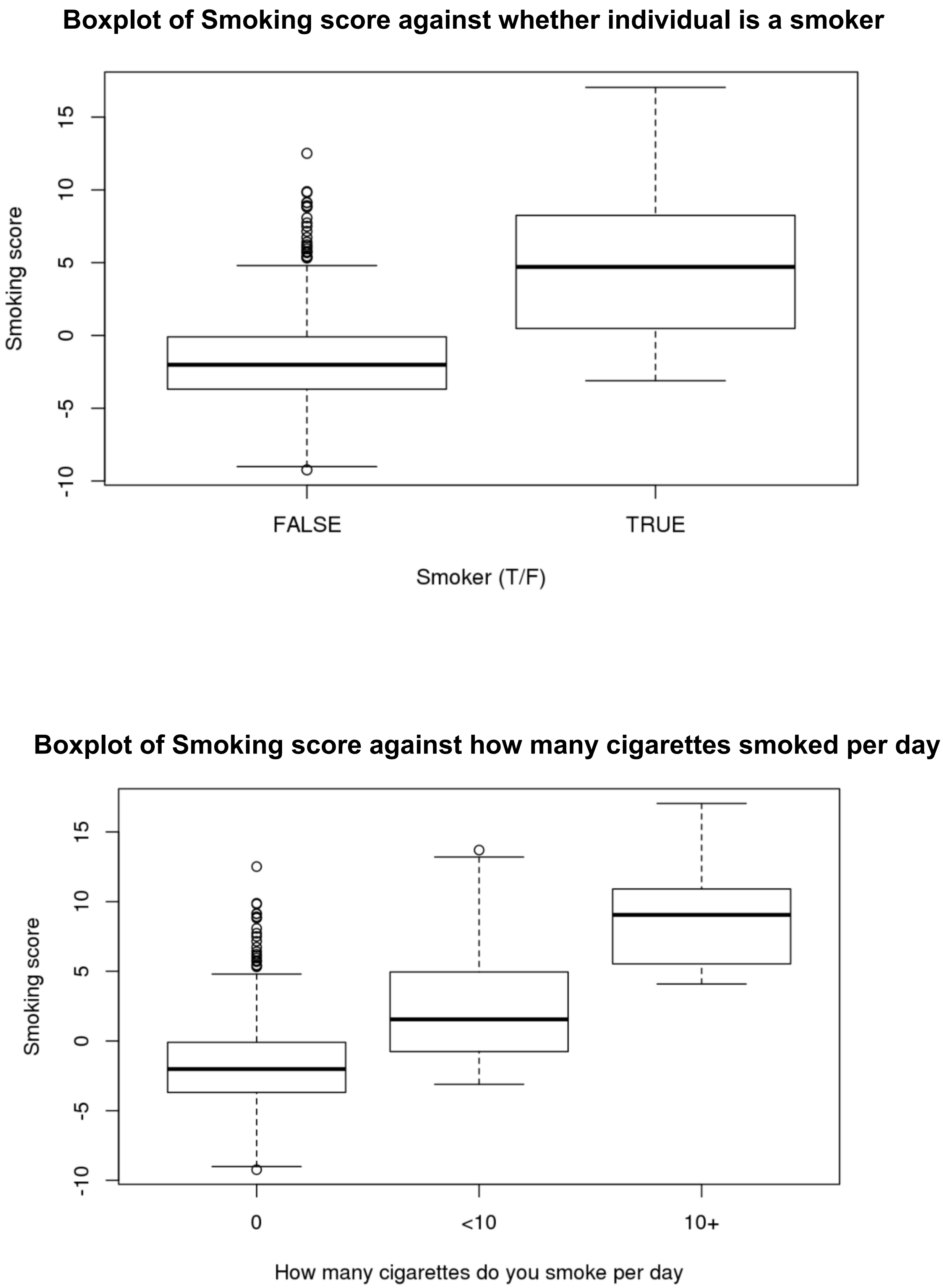


Figure 2.9: Correlations against other samples in the EXTEND cohort.

g) Calculate smoking scores (for whole blood datasets only)

Smoking is robustly associated with DNAm at sites across the genome in whole blood. Using a method developed by Elliot *at al.* (2014) we can calculate a smoking score based on DNAm at a combination of sites that can accurately predict smoking status. Where phenotypic smoking data is available it is possible to compare smoking scores between non-smokers and smokers (see **Figure 2.10**). No samples are excluded based on smoking score, but it is routinely included as a covariate in analyses.



Boxplot of Smoking score against whether individual is a smoker

Boxplot of Smoking score against how many cigarettes smoked per day

Figure 2.10: Smoking score differentiates cases and controls in the EXTEND cohort. *Top: boxplot of smokers versus non-smokers. Bottom: boxplot of how many cigarettes smokes per day.*

h) Cell type composition

DNAm varies between cell types (Mendizabal *at al.*, 2019) and therefore heterogeneity in cellular proportions can significantly influence DNAm estimates generated on bulk tissue such as whole blood. Estimations of cell-type composition are important variables to consider when analysing DNAm data. Computational methods exist to derive estimates of cellular proportions in whole blood DNAm data, using reference datasets of sorted samples. The method developed by Houseman *at al.* (Houseman *at al.*, 2012), for example, infers cell proportions based on a regression calibration technique which uses an external reference dataset to calibrate the model and correct for any bias. The function *estimateCellCounts* in the *minfi* package is used to estimate the proportion of CD8 T-cells, CD4 T-cells, natural killer cells, B-cells, monocytes and granulocytes (see **Figure 2.11**). To identify cell proportions in brain samples an algorithm has been developed using fluorescence activated nucleic sorting (FANS) data generated by our group where cortical nuclei were stained with markers for neurons (NeuN+), oligodendrocytes (Sox10+) and the remaining cells (double negative) and profiled to generate reference data (https://www.protocols.io/view/fluorescence-activated-nuclei-sorting-fans-on-huma-bmh2k38e). These data were used to estimate cell proportions in the BDR dataset.

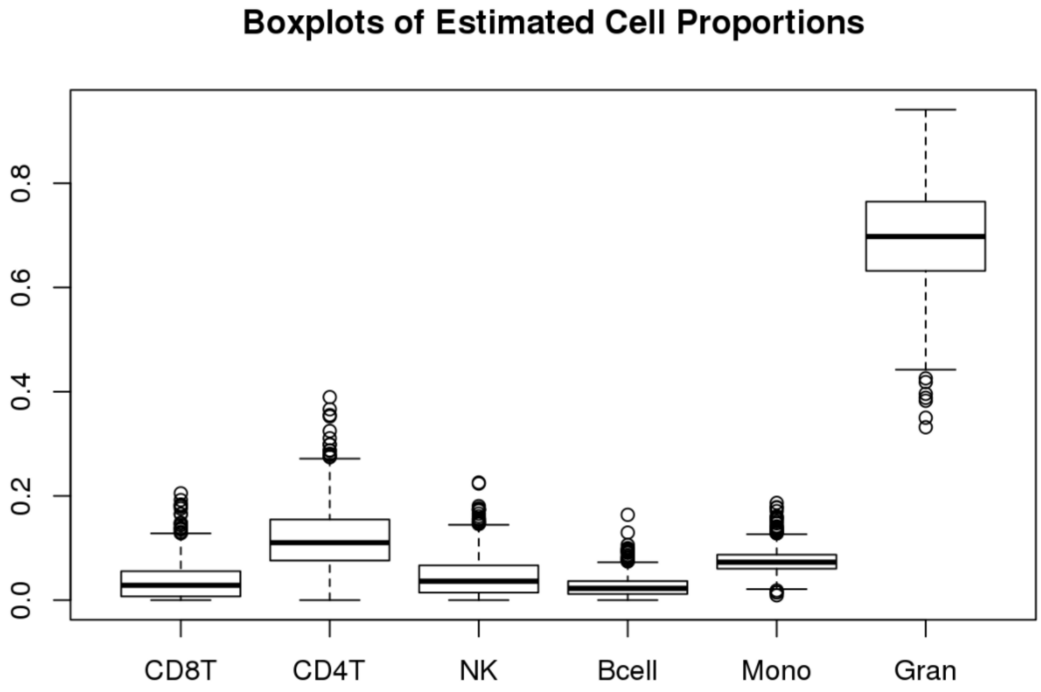


Figure 2.11: Estimated cell types in the EXTEND cohort.

i) Filter data based on beadcounts and detection p-values.

Illumina BeadChip microarrays contains several thousand sequence specific oligonucleotide coated beads to which DNA hybridises. In general, the more beads which are present on an array for a probe sequence the higher the reliability of the signal for that probe. Detection p-values are a measure of error in regard to the signals obtained from a probe in comparison to the background signal. The *pfilter* function in the *wateRmelon* package filters datasets based on beadcounts and detection p-values. If the percentage of samples with a beadcount less than 3 is greater than 5% for any probe, the probe is removed. If the percentage of probes with a detection p-value less than 0.05 is greater than 1% for any sample, the sample is removed. Similarly, if the percentage of samples with a detection p-value less than 0.05 is greater than 1% for any probe, the probe is removed.

j) Detection of outliers using the *Outlyx* function

The *outlyx* function uses PCA and Mahalanobis distances in order to determine samples which are outliers in a DNAm dataset. It considers the first principal component as the largest source of variation in a DNAm dataset. The Mahalanobis distances are calculated using the *pcout* function from the *mvoutlier* package (Filzmoser, Hron, & Reimann, 2012). Samples are identified as outliers if their values are <0.25 (out of a 0-1 range) (Filzmoser *at al.*, 2012). See **Figure 2.12** for results of the *Outlyx* function as applied to samples in the BDR dataset.

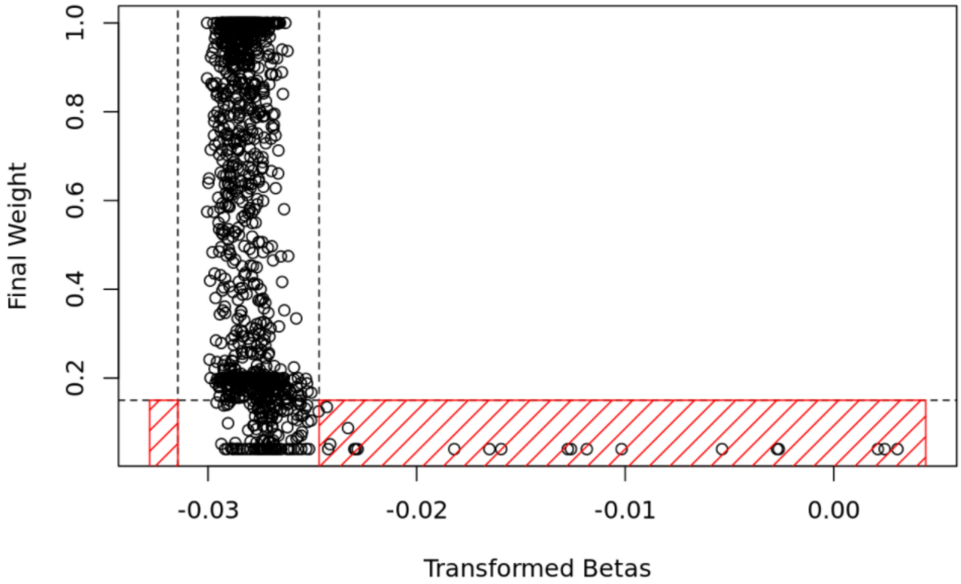


Figure 2.12: Outlyx function applied to the BDR dataset. *Outliers are those represented in the red hashed box.*

k) Remove cross hybridising probes and SNP probes

The presence of SNP variation within the vicinity (~10 base pairs (bp)) of a CpG site can interfere with probe binding and confound the measurement of DNAm (Chen *at al.*, 2013; M. E. Price *at al.*, 2013). In addition, a number of probes have been identified to cross-hybridise to other location across the genome, leading to inaccurate estimations of DNAm at the targeted site (Chen *at al.*, 2013; M. E. Price *at al.*, 2013). Probes which have been identified to be influenced by SNPs or are cross-hybridising are excluded from analyses based on a list derived by Chen *at al.* (2013) and Price *at al.* (2013).

l) Normalisation of beta values

The performance of the two probe types (I and II) differs (Dedeurwaerder *at al.*, 2011) and therefore the data needs to be normalised before DNAm can be compared between sites. For example, the type II probes cannot accurately detect extreme levels of DNAm. In addition, the increase in type II probe measurements on the EPIC array is associated with a shifted distribution of methylation values in comparison to the 450K array. A number of normalisation methods have been developed to correct for this. Our group previously developed the *wateRmelon* package (Pidsley *at al.*, 2013) which has several functions for normalisation; *dasen* normalisation is the advised function to use as it performs consistently well for both probe types (Pidsley *at al.*, 2013). See **Figure 2.13** for a comparison of un-normalised to normalised betas.

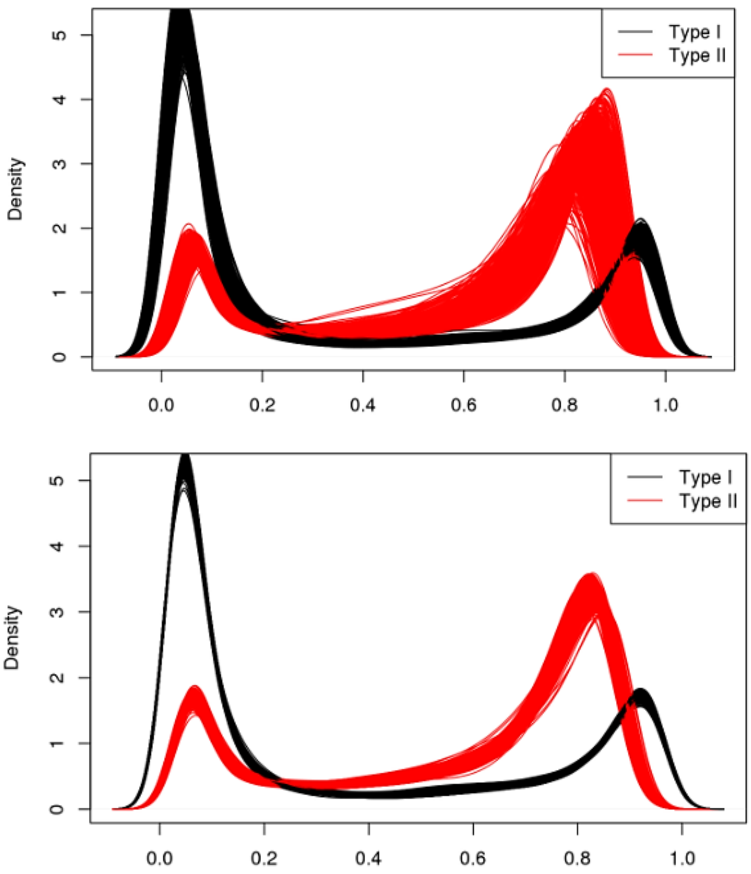


Figure 2.13: Dasen normalised betas in the EXTEND cohort. *Top: density plot of un-normalised data. Bottom: density plot of dasen normalised data.*

2.2 Genome-wide SNP profiling

Genotype data were generated for several datasets used in my thesis. They were profiled using an assortment of Illumina and Affymetrix SNP microarrays. The specific arrays used for each cohort are described in the relevant chapters. Although Illumina and Affymetix use different chemistries they both rely on the principal of complementary pairing of nucleotide bases (LaFramboise, 2009). Briefly, SNP arrays work through the hybridisation of fragmented single-stranded DNA to arrays which contain unique nucleotide probe sequences. Each probe binds to a target DNA sequence. The array is scanned to quantify the amount of sample bound to each sequence based on the signal intensity which is associated with each probe and its target after hybridisation (LaFramboise, 2009).

2.2.1 Quality control and imputation of Genotype Data

I conducted quality control and imputation for multiple genetic datasets included in **Chapter 4** and **Chapter 5**. QC was completed using *PLINK1.9* (Chang *at al.*, 2015) unless otherwise reported. There are several QC steps used in the processing of genotype data and the steps I followed are based on the pipeline suggested by Marees and colleagues (2018), described in detail below.

a) Remove samples and SNPs with high levels of missing data

High levels of missing data can be an indication of poor DNA quality or technical problems (Marees *at al.*, 2018). We exclude SNPs with missing data across many samples. In addition, samples which have high rates of missing genotype data are removed. I set the threshold to exclude SNPs and individuals with 5% missing data.

b) Checking expected sex

Samples are removed if their actual sex and genotype predicted sex (estimated using X chromosome homozygosity) are discordant, which may indicate sample mix-ups. In population genetics F-statistics are used to describe the statistically expected level of heterozygosity in a population. Females with an F value > 0.2 and males with and F value < 0.8 are removed as they have excess heterozygosity (Marees *at al.*, 2018).

c) Minor allele frequency (MAF)

SNPs with a low MAF are rare and therefore there is reduced power to detect associations with these variants. In addition, rare SNPs are more prone to genotyping errors. This threshold should be set taking into consideration samples size. Typically a threshold of 0.01 (variants present in 1% of the population) is used for larger samples of >100,000 individuals and a threshold of 0.05 (variants present in 5% of the population) is used for smaller samples (Marees *at al.*, 2018). I set the threshold to 0.05 for datasets used in this thesis.

d) Hardy-Weinberg equilibrium (HWE)

This step removes markers which deviate from HWE. The HWE law assumes there is no selection, mutation or migration within a population and also states that genotype and allele frequencies are consistent over generations (Wigginton, Cutler, & Abecasis, 2005). Deviation from HWE might indicate genotyping error or evolutionary selection. SNPs with a HWE p<1e-03 were removed from the datasets used in this thesis.

e) Heterozygosity

This relates to having inherited two different alleles at a specific SNP and heterozygosity rate is the proportion of heterozygous genotypes an individual may have. High levels may result from poor quality data and low levels may indicate inbreeding. Samples deviating ±3 SD from the heterozygosity rate mean are excluded.

f) Removal of related samples

Relatedness refers to the strength of the genetic relationship between two individuals. In general, most genetic analyses assume samples are unrelated, which is usually defined as no pair being more closely related than a 2nd degree relative. If related individuals are included and it is not accounted for in analysis models it can lead to biased estimates of effect size and standard error (Marees *at al.*, 2018). This step requires the use of independent autosomal SNPs, therefore pruning is recommended. Pruning removes SNPs in high linkage disequilibrium (LD; the correlation structure between SNPs) with each other so that the remaining variants are approximately uncorrelated. This reduces the influence of SNP clusters (Laurie *at al.*, 2010).

g) Limit samples to European ancestry and removal of population outliers

Population stratification is the concept that diversity between ancestries leads to differing allele frequencies which can confound genetic associations (A. L. Price, Zaitlen, Reich, & Patterson, 2010). Therefore most current genetic studies are limited to Europeans or a single ethnic group. To identify and exclude non-European samples, genotypes can be merged with data from HapMap Phase 3, which has known ethnicities, (http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html), followed by LD pruning the overlapping SNPs. Principal components are then calculated using the merged datatset. The first two PCs can be visually inspected along with the known ethnicities of the HapMap sample to define European samples. An example PCA plot used to determine ethnicity in the BDR cohort is shown in **Figure 2.14.**

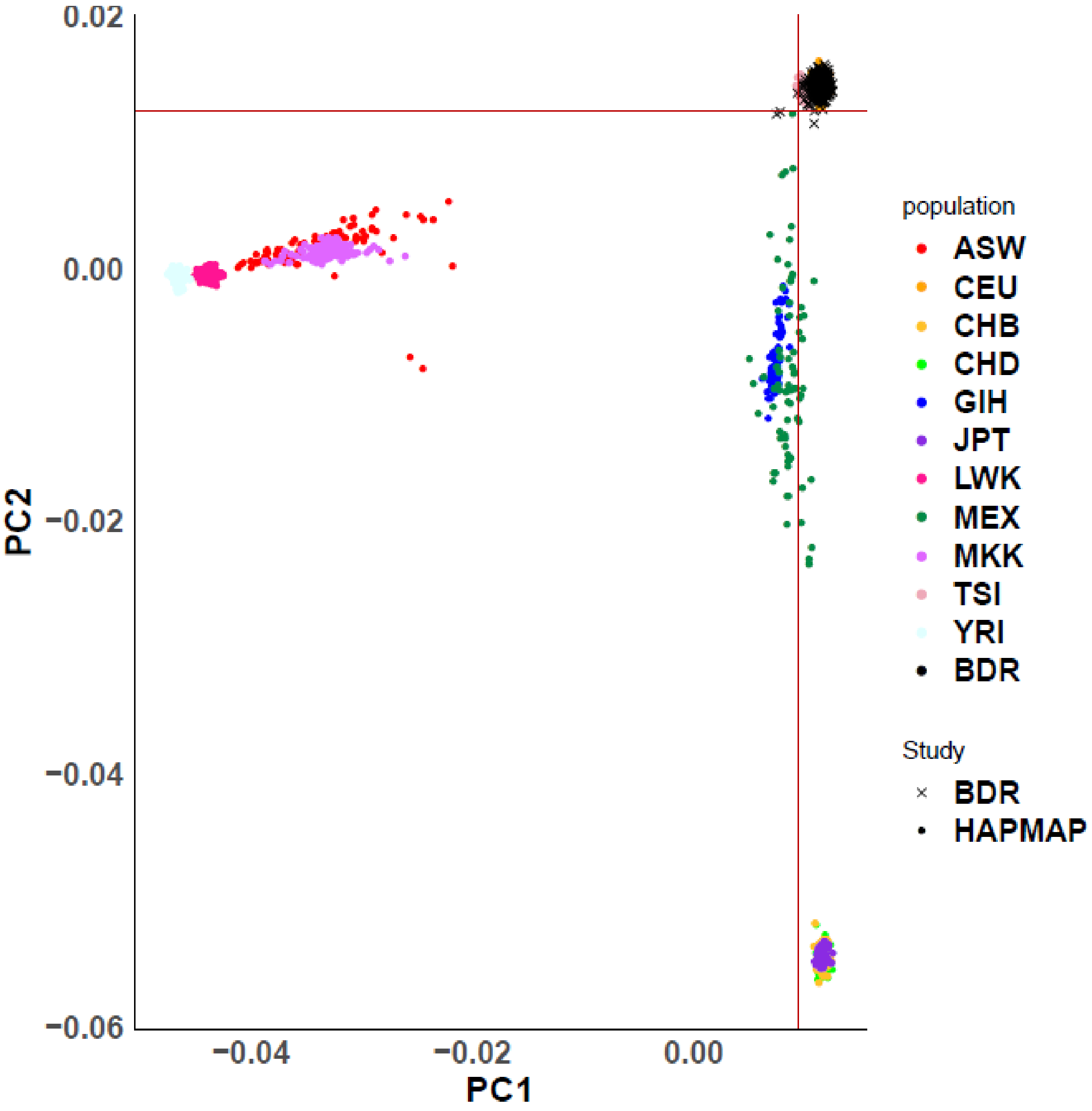


Figure 2.14: Visually ascertaining ancestry outliers in the BDR sample using principal components. *BDR was merged with HapMap3 and ancestry was ascertained based on the first two principal components. ASW = African ancestry in Southwest USA; CEU = Utah residents with Northern and Western European ancestry from the CEPH (The Centre d'Etude du Polymorphism Humain) collection; CHB = Han Chinese in Beijing, China; CHD = Chinese in Metropolitan Denver, Colorado; GIH = Gujarati Indians in Houston, Texas; JPT =Japanese in Tokyo, Japan; LWK = Luhya in Webuye, Kenya; MXL = Mexican ancestry in Los Angeles, California; MKK = Maasai in Kinyawa, Kenya; TSI = Toscani in Italia; YRI = Yoruba in Ibadan, Nigeria; BDR = Brains for Dementia Research; and HAPMAP = The International HapMap Project.*

2.2.2 Imputation

Imputation allows for the prediction of unmeasured genotypes in low-density datasets (e.g. those from SNP arrays) using densely genotyped datasets as references such as HapMap or 1000 Genomes (Jostins, Morley, & Barrett, 2011)). This results in a substantial increase in power and allows for the meta-analysis of studies genotypes on difference SNP arrays. Several tools are available for imputation. Within this thesis genotype data was imputed using the Michigan Imputation Server (Das *at al.*, 2016) (https://imputationserver.sph.umich.edu/index.html#!) which uses *Eagle2* (Loh *at al.*, 2016) to phase haplotypes, and *Minimac4* (https://genome.sph.umich.edu/wiki/Minimac4) with the most recent 1000 Genomes reference panel (phase 3, version 5) (1000 Genomes Project Consortium *at al.*, 2015).

2.3 Statistical methods for Epigenome-wide association studies

2.3.1 Linear regression

Regression is a statistical process which is used to estimate the relationship between variables. A linear regression takes the assumption that this relationship is linear. The most commonly used model in epigenome wide association studies (EWAS) is a multiple linear regression model which allows for the inclusion of biological, environmental and technical confounders. The equation for a multiple linear regression in relation to an EWAS is shown below:

$$\mathbf{DNA\ methylation_i\ = \beta_1 trait + \beta covariates}$$

DNA methylation$_i$ represents DNA methylation at each probe, $\beta_1$trait = phenotype of interest (e.g. disease status) and βcovariates are confounders such as age, sex and experimental batch (i.e. $\beta_2$age $\beta_3$sex… β_x).

There are several assumptions for linear regression:

- There must be a linear relationship between the outcome variables and the independent variables. This can be assessed using scatter plots.
- The residuals must be normally distributed. This can be assessed using quantile-quantile plots, which enables the visualisation of theoretical quantiles against standardised residuals. If the residuals are normally distributed if they follow the diagonal line.
- There is no multi-collinearity. This assumes independent variables are not highly correlated with each other.

- The variances of error terms are similar across the values of the independent variables. This is known as the assumption of homoscedasticity. A quantile-quantile plot can show whether points are equally distributed across all values of the independent variables.

Recently, Mansell and Colleagues (Mansell *at al.*, 2019) tested the assumptions of linear regression in EWAS as there have been discussions in the literature suggesting that DNAm at many sites in the genome violates these assumptions (Du *at al.*, 2010; Laird, 2010). To test the assumptions they performed an EWAS of age, which is known to be associated with DNAm variation at many loci across the genome. They found that 70% of sites rejected the null hypothesis for at least one assumption, with many sites exhibiting evidence of non-normal distributions of residuals, either due to skewness (an asymmetrical distribution) or kurtosis (having a none bell shaped distribution). Very few sites rejected the hypothesis in favour of a non-linear model or heteroscedasticity. Violations in the assumption can lead to false positives or false negatives, however Mansell and colleagues (Mansell *at al.*, 2019) found that even very significant rejections of the linear regression assumptions did not bias EWAS results in terms of false positives and negatives. Therefore, linear regression is a valid statistical methodology for DNAm studies.

2.3.2 Mixed effects models

Mixed effect regression models allow for the inclusion of both fixed effects (effects that are constant across individuals) as well as random effects (effects that vary across individuals) and are used when the data have global and group-level trends. Examples of fixed effects in EWAS regression models include age, sex and experimental batch (e.g. the specific 96 well plate the samples were run on). An example of a random effect in EWAS is individual ID, where there might be two samples from the same person (e.g. in BDR we profiled OCC and PFC tissue samples for each individual). The random intercept allows the line to cross the y-axis at a different position, however the relationship between the variables remains the same. Random slopes change the gradient of the line, changing the relationship between the variables for each group. This model accounts for the fact the samples are not fully independent.

The equation for a mixed linear regression in relation to an EWAS is shown below:

$$\textbf{DNA methylation}_{\mathbf{i}} = \boldsymbol{\beta}_{\mathbf{1}}\textbf{trait} + \boldsymbol{\beta}\textbf{covariates} + (\mathbf{1}|\textbf{individual ID})$$

DNA methylation$_i$ represents DNA methylation at each probe, $\beta_1$trait = phenotype of interest (e.g. disease status) and βcovariates are confounders to include such as age, sex and technical artefacts (i.e. $\beta_2$age $\beta_3$sex… β_x), (1|individual ID) represents a random effect, and in this case is individual ID.

2.3.3 Identifying differentially methylated regions

In order to identify differentially methylated regions (DMRs) – i.e. genomic regions in which DNA methylation across multiple sites is consistently associated with a phenotype – I applied the *dmrff* package (Suderman *at al.*, 2018) to my EWAS results generated in **Chapter 4** and **Chapter 6**. *Dmrff* identifies regions by combining summary statistics from proximally located DNAm sites. Since meta-analysis methods like Fisher's method assume independence between sites they cannot be used for DMR analysis as this is rarely true of neighbouring DNAm sites. If the assumption of independence is violated it can lead to inflation in association statistics (i.e. false positives). Burgess and colleagues developed a method to account for this proximal correlation which uses an extension of an inverse-variance weighted (IVW) meta-analysis (Burgess, Dudbridge, & Thompson, 2016). *Dmrff* is based on this IVW methodology and is computationally efficient (i.e. fast) and unlike most other DMR methods it controls false positive rates (Suderman *at al.*, 2018). *Dmrff* identifies candidate regions (regions containing ≥3 probes) as sequences of DNAm sites with EWAS values that reach a certain P_t. I used the threshold $p< 0.05$. *Dmrff* shrinks the regions based on calculations for each of the sub-regions and uses a greedy algorithm (an algorithm that takes the best immediate/ local solution) to select sub-regions which cover the candidate regions using the strongest statistics. P-values were Bonferroni corrected for multiple testing burden and treats each EWAS test and sub-region as independent tests.

2.1.1 Pathway analysis

An important downstream analysis after genome-wide methylation studies (e.g. EWAS and SMR) is gene set enrichment analysis, whereby significant genes and DNAm sites can be related to known biological functions and ontologies. The first methods developed to do this were severely biased by variation in gene length (Geeleher *at al.*, 2013), for example larger genes (which may be associated with more DNAm probes) have a higher chance of being identified as differentially methylated. In order to address these biases several methods have been developed but these methods have predominantly been developed for RNA-seq data. For example, Young and colleagues (Young, Wakefield, Smyth, & Oshlack, 2010) developed *GOseq* which uses weighted resampling in addition to Wallenius non-central hypergeometric approximation for over-representation analysis. Other methods to reduce biases in pathway-analysis include incorporating Wald test statics to adjust for length bias when ranking genes using functional class scoring (Gao, Fang, Zhang, Zhi, & Cui, 2011) or including gene length as a covariates in the logistic regression models used for the pathway analysis (S. Li, He, Pawlikowska, & Lin, 2017; Mi, Di, Emerson, Cumbie, & Chang, 2012). Recently Ren and Kuan (2019) developed *methylGSA*, which accounts for these biases in DNAm data. The Illumina UCSC gene annotation manifest was used to create test gene lists from DMPs identified in EWAS for the pathway analyses conducted in this thesis in **Chapter 4** and **Chapter 5**. Where probes were not annotated to any gene (i.e. if they are in intergenic locations), they were excluded from pathway analysis.

2.3.4 Gene annotation of EWAS

Throughout this thesis differentially methylated positions (DMPs) and regions (DMRs) were annotated using the standard Illumina UCSC gene annotation manifest, which is derived from the genomic overlap of probes with RefSeq genes or up to 1500 bp of the transcription start site of a gene (Karolchik *at al.*, 2003; Kent *at al.*, 2002).

3 Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex

This chapter is presented in the form of a peer-reviewed manuscript which has been published in Brain (Shireby *at al.*, 2020). It has been reformatted to the style of the thesis.

The Supplementary Figures and Tables are included at the end of the main manuscript (**Figures 3.4-3.10**; **Tables 3.5-3.9**)

I conducted additional analyses investigating associations between DNA methylation age and neuropathology in the BDR cohort, and these results are presented at the end of the Chapter (see Section **3.9**).

Abstract

Human DNA-methylation data have been used to develop biomarkers of ageing - referred to as ‘epigenetic clocks’ - that have been widely used to identify differences between chronological age and biological age in health and disease including neurodegeneration, dementia and other brain phenotypes. Existing DNA methylation clocks have been shown to be highly accurate in blood but are less precise when used in older samples or in tissue-types not included in training the model, including brain. We aimed to develop a novel epigenetic clock that performs optimally in human cortex tissue and has the potential to identify phenotypes associated with biological ageing in the brain. We generated an extensive dataset of human cortex DNA methylation data spanning the life-course (n = 1,397, ages = 1 to 108 years). This dataset was split into ‘training’ and ‘testing’ samples (training: n = 1,047; testing: n = 350). DNA methylation age estimators were derived using a transformed version of chronological age on DNA methylation at specific sites using elastic net regression, a supervised machine learning method. The cortical clock was subsequently validated in a novel independent human cortex dataset (n = 1,221, ages = 41 to 104 years) and tested for specificity in a large whole blood dataset (n = 1,175, ages = 28 to 98 years). We identified a set of 347 DNA methylation sites that, in combination, optimally predict age in the human cortex. The sum of DNA methylation levels at these sites weighted by their regression coefficients provide the cortical DNA methylation clock age estimate. The novel clock dramatically out-performed previously reported clocks in additional cortical datasets. Our findings suggest that previous associations between predicted DNA methylation age and neurodegenerative phenotypes might represent false positives resulting from clocks not robustly calibrated to the tissue being tested and for phenotypes that become manifest in older ages. The age distribution and tissue type of samples included in training datasets need to be considered when building and applying epigenetic clock algorithms to human epidemiological or disease cohorts.

Keywords: Cortex, age, ageing, disease, epigenetic clock, DNA methylation, post-mortem

3.1 Introduction

Advancing age is associated with declining physical and cognitive function, and is a major risk factor for many human brain disorders including dementia and other neurodegenerative diseases (Harper, 2014; Sierra, 2019). Understanding the biological mechanisms involved in ageing will be a critical step towards preventing, slowing or reversing age-associated phenotypes. Due to the substantial inter-individual variation in age-associated phenotypes, there is considerable interest in identifying robust biomarkers of 'biological' age, a quantitative phenotype that is thought to better capture an individuals' risk of age-related outcomes than actual chronological age (Jylhävä, Jiang, Foebel, Pedersen, & Hägg, 2019). Several data modalities have been used to generate estimates of biological age; these include measures of physical fitness (e.g. muscle strength) (Sosnoff & Newell, 2006), cellular phenotypes (e.g. cellular senescence) (Baker *at al.*, 2011), genomic changes (e.g. telomere length) (Jylhävä, Pedersen, & Hägg, 2017; Sanders & Newman, 2013) and epigenetic mechanisms (e.g. DNA methylation) (Horvath, 2013).

Epigenetic mechanisms act to regulate gene expression developmentally via chemical modifications to DNA and histone proteins (Bernstein, Meissner, & Lander, 2007), conferring cell-type-specific patterns of gene expression and differing markedly between tissues and cell-types (Mendizabal & Yi, 2016). There has been recent interest in the dynamic changes in epigenetic processes over the life course, and a number of 'epigenetic clocks' based on one specific epigenetic modification - DNA methylation (DNAm) - have been developed that appear to be highly predictive of chronological age (Hannum *at al.*, 2013; Horvath, 2013). The landmark DNAm clock was developed by Horvath (Horvath, 2013), who applied elastic net regression to Illumina DNAm array data from a large number of samples derived from a range of tissues (n = ~ 8,000 across 51 tissue and cell types), and generated a predictor based on DNAm at 353 CpG sites that is highly predictive of chronological age (Horvath, 2013). Given that changes in DNAm are known to index exposure to certain environmental risk factors (for example, tobacco smoking) (Elliott *at al.*, 2014; Sugden *at al.*, 2019) that are associated with diseases of old age, and variable DNAm is robustly associated with a number of age-associated disorders (Chouliaras *at al.*, 2018; Chuang *at al.*, 2017; A. R. Smith *at al.*, 2016), there has been interest in the hypothesis that DNAm clocks might robustly quantify variation in biological age.

Horvath's DNAm age clock, for example, has been widely applied to identify accelerated epigenetic ageing - where DNAm age predictions deviate from chronological age such that individuals appear older than they really are - in the context of numerous health and disease outcomes (Horvath & Ritz, 2015; Levine, Lu, Bennett, & Horvath, 2015; Marioni *at al.*, 2015; McCartney *at al.*, 2018). Although the original DNAm clocks were primarily developed to predict chronological age and are not robustly predictive of clinical health measures (e.g. blood pressure) (Quach *at al.*, 2017), more recent DNAm clocks such as Levine's 'pheno age' clock (Levine *at al.*, 2018) incorporate surrogate measures of biological age and are more directly aimed at predicting mortality and health-span. Since age is a major risk factor for dementia and other neurodegenerative brain disorders, there is particular interest in the application of epigenetic clock algorithms to these phenotypes, especially as differential DNAm in the cortex has been robustly associated with diseases including Alzheimer's disease and Parkinson's disease (Lunnon *at al.*, 2014; A. R. Smith *at al.*, 2016; Yu *at al.*, 2015). Recent studies have reported an association between accelerated DNAm age and specific markers of Alzheimer's disease neuropathology in the cortex (e.g. neuritic plaques, diffuse plaques and amyloid-β load) (Levine *at al.*, 2018, 2015). Furthermore, among individuals with Alzheimer's disease, DNAm age acceleration is associated with declining global cognitive functioning and deficits in episodic and working memory (Levine *at al.*, 2018, 2015).

A strength of several existing epigenetic clocks is that they work relatively well across different types of sample; the Horvath multi-tissue clock, for example, can accurately predict age in multiple tissues across the life-course. Importantly, as with any predictor, the composition of the training data used to develop the clock influences the generality of the model. To date, there has been limited research comparing the prediction accuracy and potential bias of existing clock algorithms across different tissues and ages. Recent analyses have highlighted potential biases when using Horvath's clock in older samples (>~60 years) and in samples derived from certain tissues, especially the central nervous system (El Khoury *at al.*, 2019). This is important for the interpretation of studies of possible relationships between accelerated epigenetic age and age-related diseases affecting the human brain (e.g. neurodegenerative phenotypes); reported associations between accelerated DNAm age and disease may actually be a consequence of fitting a suboptimal predictor to available datasets.

Potential confounders include differential changes in DNAm with age across tissues, and the age distribution of the samples used to train existing classifiers. Resolution of these biases requires the construction of specific DNAm clocks developed using data generated on the relevant tissue-type and including broad representation of the age spectrum they will be used to interrogate. Recently, a number of tissue-specific DNA methylation clocks have been described, including clocks designed for whole blood (Hannum *at al.*, 2013; Zhang *at al.*, 2019), muscle (Voisin *at al.*, 2020), bone (Gopalan, Gaige, & Henn, 2019) and paediatric buccal cells (McEwen *at al.*, 2019). Importantly, although these DNAm age estimators have increased predictive accuracy within the specific tissues in which they were built, they lose this precision when applied to other tissues (Zhang *at al.*, 2019).

In this study, we describe the development of a novel DNAm clock that is specifically designed for application in DNA samples isolated from the human cortex and is accurate across the lifespan including in tissue from older donors (aged over 60 years). We demonstrate that our clock outperforms existing DNAm-based predictors developed for other tissues, minimising the potential for spurious associations with ageing phenotypes relevant to the brain.

3.2 Materials and methods

3.2.1 Datasets used to develop the novel cortical DNAm age clock

To develop and characterise our cortical DNAm age clock ("DNAmClock$_{\text{Cortical}}$") we collated an extensive collection of DNAm data from human cortex samples (see **Supplementary Table 3.5** and **Supplementary Table 3.6)**, complementing datasets generated by our group (http://www.epigenomicslab.com) with publicly available datasets downloaded from the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) (De Jager *at al.*, 2014; Jaffe *at al.*, 2016; Lunnon *at al.*, 2014; Pidsley *at al.*, 2014; A. R. Smith *at al.*, 2019; R. G. Smith *at al.*, 2018; Wong *at al.*, 2019) (see **Supplementary Table 3.5** and **Supplementary Table 3.6**). In each of these datasets DNAm was quantified across the genome using the Illumina 450K DNA methylation array which covers >450,000 DNA methylation sites as previously described (Pidsley *at al.*, 2013). To optimise the performance of the

DNAmClock$_{\text{Cortical}}$ and to avoid reporting over-fitted statistics, the samples were split into a “training” dataset (used to determine the DNAm sites included in the model and their weighted coefficients) and a “testing” dataset (used to profile the performance of the proposed model). To reduce the effects of experimental batch in our model, we maximised the number of different datasets included in the training data by combining the ten cohorts and randomly assigning individuals within them to either the training or testing dataset in a 3:1 ratio **Table 3.1**). In total, our training dataset (age range = 1-108 years, median = 57 years; see **Supplementary Figure 3.4**) comprised DNAm data from 1,047 cortex samples (derived from 832 donors) and our testing dataset (age range = 1-108 years, median = 56 years; see **Supplementary Figure 3.4**) comprised DNAm data from 350 cortex samples (derived from 323 donors). Individuals with a diagnosis of Alzheimer’s disease and other major neurological phenotypes were excluded from our analysis given the previous associations between them and deviations in DNAm age (Levine *at al.*, 2018, 2015).

3.2.2 Cortex independent test dataset

An “independent test” cortical dataset was generated using post-mortem occipital (OCC) and prefrontal cortex (PFC) samples from the Brains for Dementia Research (BDR) cohort. BDR was established in 2008 and is a UK-based longitudinal cohort study with a focus on dementia research (Francis, Costello, & Hayes, 2018) coordinated by a network of six dementia research centres based around the UK. Post-mortem brains underwent full neuropathological dissection, sampling and characterisation using a standardised protocol (Bell *at al.*, 2008; Samarasekera *at al.*, 2013). DNA was isolated from cortical tissue samples using the Qiagen AllPrep DNA/RNA 96 Kit (Qiagen, cat no.80311) following tissue disruption using BeadBug 1.5 mm Zirconium beads (Sigma Aldrich, cat no.Z763799) in a 96-well Deep Well Plate (Fisher Scientific, cat no.12194162) shaking at 2500rmp for 5 minutes. Genome-wide DNA methylation was profiled using the Illumina EPIC DNA methylation array (Illumina Inc), which interrogates >850,000 DNA methylation sites across the genome (Moran, Arribas, & Esteller, 2016). After stringent data quality control (see below) the final independent test dataset consisted of DNAm estimates for 800,916 DNAm sites profiled in 1,221 samples (632 donors; 610 PFC; 611 OCC; see Table 1 for more details). This dataset consists of predominantly older samples (age range = 41-104 years, median = 84 years; see **Supplementary Figure 3.4**).

3.2.3 Whole blood dataset

We recently generated DNAm data from whole blood obtained from 1,175 individuals (age range = 28-98 years; median age = 59 years; see **Table 3.1** for more details) included in the UK Household Longitudinal Study (UKHLS) (https://www.understandingsociety.ac.uk/) (Hannon *at al.*, 2018).The UKHLS was established in 2009 and is a longitudinal panel survey of 40,000 UK households from England, Scotland, Wales and Northern Ireland (Buck & McFall, 2011). For each participant, non-fasting blood samples were collected through venipuncture; these were subsequently centrifuged to separate plasma and serum, and samples were aliquoted and frozen at −80°C. DNAm data were generated using the Illumina EPIC DNA methylation array as described previously (Hannon *at al.*, 2018). After stringent QC (see **3.2.4**) the whole blood dataset consisted of data for 857,071 DNAm sites profiled in 1,175 samples (Hannon *at al.*, 2018).

3.2.4 DNA methylation data pre-processing

Unless otherwise reported, all statistical analysis was conducted in the R statistical environment (version 3.5.2; https://www.r-project.org/). Raw data for all datasets were used, prior to any QC or normalisation, and processed using either the *wateRmelon* (Pidsley *at al.*, 2013) or *bigmelon* (Gorrie-Stone *at al.*, 2019) packages. Our stringent QC pipeline included the following steps: (1) checking methylated and unmethylated signal intensities and excluding poorly performing samples; (2) assessing the chemistry of the experiment by calculating a bisulphite conversion statistic for each sample, excluding samples with a conversion rate <80%; (3) identifying the fully methylated control sample was in the correct location (where applicable); (4) multidimensional scaling of sites on the X and Y chromosomes separately to confirm reported sex; (5) using the 65 SNP probes present on the Illumina 450K array and 59 on the Illumina EPIC array to confirm that matched samples from the same individual (but different brain regions) were genetically identical and to check for sample duplications and mismatches; (6) use of the *pfilter* function in *wateRmelon* to exclude samples with >1 % of probes with a detection P value > 0.05 and probes with >1 % of samples with detection P value > 0.05; (8) using principal component analysis on data from each tissue to exclude outliers based on any of the first three principal components; (9) removal of cross-hybridising and SNP probes (Chen *at al.*, 2013).

The subsequent normalisation of the DNA methylation data was performed using the *dasen* function in either *wateRmelon* or *bigmelon* (Gorrie-Stone *at al.*, 2019; Pidsley *at al.*, 2013).

3.2.5 Deriving a novel cortical DNAm age classifier

To build the DNAmClock$_{\text{Cortical}}$ we implemented an elastic net (EN) regression model, using the methodology described by Horvath (2013). The EN model is designed for high dimensional datasets with more features than samples and where the features are potentially highly correlated (Zou & Hastie, 2005). As part of the methodology, the model selects the subset of features (i.e. DNAm sites) that cumulatively produce the best predictor of a provided outcome. EN was implemented in the R package *GLMnet* (Friedman, Hastie, & Tibshirani, 2010). It uses a combination of Ridge and LASSO (Least Absolute Shrinkage and Selection Operator) regression. Ridge regression penalises the sum of squared coefficients and has an (alpha) parameter of zero. LASSO regression penalises the sum of the absolute values of the coefficients and has an $\boldsymbol{\alpha}$ parameter of one. EN is a convex combination of ridge and LASSO and, therefore, the elastic net $\boldsymbol{\alpha}$ parameter was set to 0.5. The lambda value (the shrinkage parameter) was derived using 10-fold cross-validation on the training dataset (lambda = 0.0178). DNAm probes included in the analysis were limited to sites which were present on both the Illumina EPIC and Illumina 450K arrays, with no missing values across the training datasets (n probes = 383547). Previous analyses have shown that the relationship between DNAm age (predicted age from epigenetic age estimators) and chronological age is logarithmic between 0-20 years and linear from 20 years plus (Horvath, 2013). Our data revealed a similar pattern and therefore chronological age was transformed (**Supplementary Figure 3.5**). A transformed version of chronological age was regressed on DNAm levels at all included DNAm sites.

Table 3.1: Sample characteristics of the training (cortex), testing (cortex), independent test (cortex) and whole blood datasets used in the development and evaluation of our novel cortical DNA methylation clock.

Dataset		**Age (years)**					**Sex (n)**		**Illumina methylation array**	**References**
	N	**Mean**	**1st Quartile**	**Median**	**3rd Quartile**	**Range**	**Female**	**Male**		
Training	1047	56.53	38.56	57	78	1-108	362	685	450K	(De Jager *at al.*, 2014; Jaffe *at al.*, 2016; Lunnon *at al.*, 2014; Pidsley *at al.*, 2014; A. R. Smith *at al.*, 2019; R. G. Smith *at al.*, 2018; Wong *at al.*, 2019)
Testing	350	55.87	39.05	56	78	1-108	144	206	450K	(De Jager *at al.*, 2014; Jaffe *at al.*, 2016; Lunnon *at al.*, 2014; Pidsley *at al.*, 2014; A. R. Smith *at al.*, 2019; R. G. Smith *at al.*, 2018; Wong *at al.*, 2019)
Independent Test	1221	83.49	78	84	90	41-104	577	644	EPIC	-
Blood	1175	57.96	46	59	69	28-98	686	489	EPIC	Hannon *at al.* (2018)

3.2.6 Implementing DNAm Age prediction

We applied the DNAmClock$_{\text{Cortical}}$ (comprising 347 DNAm sites) to the testing, independent test and whole blood DNAm datasets. We then compared its performance to a number of existing DNAm clocks which have previously shown biases when applied to brain tissue and older samples (El Khoury *at al.*, 2019; Horvath & Raj, 2018; Zhang *at al.*, 2019): Horvath's original multi-tissue clock ("DNAmClock$_{\text{Multi}}$"; 353 DNAm sites) (Horvath, 2013), Zhang's EN blood and saliva-based DNAm clock ("DNAmClock$_{\text{Blood}}$": 514 DNAm sites) (Zhang *at al.*, 2019) and Levine's second generation 'pheno age' DNAm Clock ("DNAmClock$_{\text{Pheno}}$"; 513 DNAm sites) (Levine *at al.*, 2018). Briefly, to predict DNAm age using the DNAmClock$_{\text{Multi}}$ we applied the *agep* function in *wateRmelon* (Pidsley *at al.*, 2013). Although this function does not contain the custom normalisation method applied at the DNAm age calculator website (https://DNAmClock.genetics.ucla.edu/), both methods work similarly in brain and blood studies, providing the data have been pre-processed adequately (El Khoury *at al.*, 2019).To predict age using the DNAmClock$_{\text{Pheno}}$ (Levine *at al.*, 2018), we also applied the *agep* function, inputting a vector of the coefficients and the intercept using the data provided in the supplementary material of Levine et al's manuscript. To predict DNAm age with the DNAmClock$_{\text{blood}}$, we used the authors' published code (available on GitHub https://github.com/qzhang314/DNAm-based-age-predictor) (Zhang *at al.*, 2019).

3.2.7 Determining the predictive accuracy of different DNAm clocks

DNAm age was estimated in the testing dataset (n = 350), independent test dataset (n = 1,221) and whole blood dataset (n = 1,175) using the DNAmClock$_{\text{Cortical}}$, DNAmClock$_{\text{Multi}}$, DNAmClock$_{\text{Blood}}$ and the DNAmClock$_{\text{Pheno}}$. To compare and evaluate the predictive accuracy of these DNAm age predictors, estimates were assessed using two measures: Pearson's correlation coefficient (r; a measure indicating the strength of the linear relationship between the actual (chronological age) and predicted (DNAm age) variables) and the root mean squared error (RMSE; square root of the mean differences between the actual and predicted variables) which quantifies the precision of the estimator.

3.2.8 Analysis against age

To test associations between DNAm age and chronological age, we fitted regression models to each dataset. As a subset of donors included in the testing and independent test datasets contributed data from multiple cortical regions, we used mixed effects linear regression models, implemented with the *lme4* and *lmerTest* packages, where DNAm age was regressed against chronological age as a fixed effect and individual was included as a random effect. In the blood cohort, as there was only one sample per individual, we applied standard linear regression models. A second regression model was also fitted which additionally tested for associations with an age-squared term. In the whole blood dataset, we ran these analyses again in the subset of samples over 55 years old to make the results more comparable to those generated using the independent test dataset.

3.2.9 Analysis against biological and technical factors

To test associations between DNAm age and sex, post-mortem interval (PMI), experimental batch and neuronal cell proportion estimates (derived using the CETS algorithm (Guintivano *at al.*, 2013)) we fitted regression models to the independent test dataset (n = 1,221 cortical samples). We used mixed effects regression models implemented as described above. DNAm age was regressed against each variable in turn with age, age squared and derived cell proportion estimates (excluding the model looking for associations with cell proportions) as fixed effects and individual as random effect. In the analysis with PMI we included brain bank as a fixed effect.

3.2.10 Data availability

The datasets used for the training and testing samples are available for download from GEO (https://www.ncbi.nlm.nih.gov/geo/) using the following accession numbers: GSE74193; GSE59685; GSE80970; GPL13534 and GSE43414. The independent test data are available from the authors upon request or via the Dementias Platform UK (DPUK) data portal (https://portal.dementiasplatform.uk/). The whole blood DNA methylation data are available upon application through the European Genome-Phenome Archive under accession code EGAS00001001232. Analysis scripts used in this manuscript and code to run the clock are available on GitHub (https://github.com/gemmashireby/CorticalClock).

3.3 Results

3.3.1 Existing DNAm clock algorithms work sub-optimally in the human cortex, systematically underestimating age in older individuals

The performance of DNAm clocks is influenced by the characteristics (e.g. specific tissue type and age range) of the training data used to build the prediction algorithm. Applying predictors to datasets that differ in terms of these characteristics may lead to biases when estimating DNAm age, and confound phenotypic analyses using these variables (El Khoury *at al.*, 2019). We found that existing DNAm clocks (i.e. the $\text{DNAmClock}_{\text{Multi}}$ (Horvath, 2013) the $\text{DNAmClock}_{\text{Blood}}$ (Zhang *at al.*, 2019) and the $\text{DNAmClock}_{\text{Pheno}}$ (Levine *at al.*, 2018)) do not perform optimally in human cortex tissue (**Supplementary Figure 3.6**), with notable differences between derived DNAm age and actual chronological age (i.e. the derived values do not lie along the y = x line, see **Figure 3.1**). In our testing dataset (n = 350 cortex samples; age range = 1 - 108 years; median age = 57 years), the $\text{DNAmClock}_{\text{Multi}}$ systematically underestimated DNAm age in individuals over ~60 years old, and systematically overestimated it in individuals below ~60 years old (**Figure 3.1A**(**ii**) and **Figure 3.2A**(**ii**)). In the older group (aged over 60 years), around 80% of samples had lower predicted DNAm ages than their actual chronological age. These deviations were also observed when looking at the mean differences between actual age and predicted DNAm age (referred to as Δ (delta) age), such that Δ age was positive for younger ages and vice versa for the older group (**Supplementary Figure 3.7A**). Use of the $\text{DNAmClock}_{\text{Blood}}$ produced even more pronounced systematic underestimation of DNAm age in adults, starting around 30 years (**Figure 3.1A**(**iii**) and **Figure 3.2A**(**iii**)*),* and this trend was mirrored for Δ age (see **Supplementary Figure 3.7A**)*.* Finally, the $\text{DNAmClock}_{\text{Pheno}}$ severely under predicted age in the cortex, with 100% of samples being assigned a lower DNAm age than the actual chronological age (**Figure 3.1A**(**iv**)**, Figure 3.2A**(**iv**) **and Supplementary Figure 3.7A** (**iv**)**).** Similar biases in age prediction were seen in our independent test dataset (n = 1,221 cortex samples; age range = 41 years to 104 years; mean age = 83.49 years), confirming the systematic underestimation of DNAm age in older donor**s (see Figure 3.1B** and **Figure 3.2B**). As with the other clocks, Δ age captured these biases, with particularly poor performance evident when applying the DNAmClock$_{\text{Pheno}}$ and the DNAmClock$_{\text{Blood}}$ to this dataset, in which Δ age was consistently below zero (where zero would represent perfect prediction; see **Supplementary Figure 3.7B**).

3.3.2 Developing a novel DNAm clock for the human cortex based on 347 DNA methylation sites

To alleviate the biases observed when applying existing DNAm clocks to data generated on older human cortex samples, we focussed on building a DNAm clock using relevant tissue samples from donors that spanned a broad range of ages and included a large number of samples from older donors (**Supplementary Figure 3.4**). We developed our novel cortical DNAm clock (DNAmClock$_{\text{Cortical}}$) using an elastic net regression, regressing chronological age against DNAm levels across 383,547 sites quantified in 1,047 cortex samples (see **Methods 3.2**). This approach identified a set of 347 DNAm sites which in combination optimally predict age in the human cortex. The sum of DNAm levels at these sites weighted by their regression coefficients provides the DNAmClock$_{\text{Cortical}}$ age estimate (see **Supplementary Table 3.7**). Of note, the majority of sites selected for our cortex clock were novel and not present in existing DNAm clock algorithms; only 5 of the sites overlap with the DNAmClock$_{\text{Multi}}$ (composed of 353 DNAm sites), 15 with the DNAmClock$_{\text{Blood}}$ (comprising 514 DNAm sites), and 5 with the DNAmClock$_{\text{Pheno}}$ (comprising 513 DNAm sites) (see **Supplementary Table 3.8**).

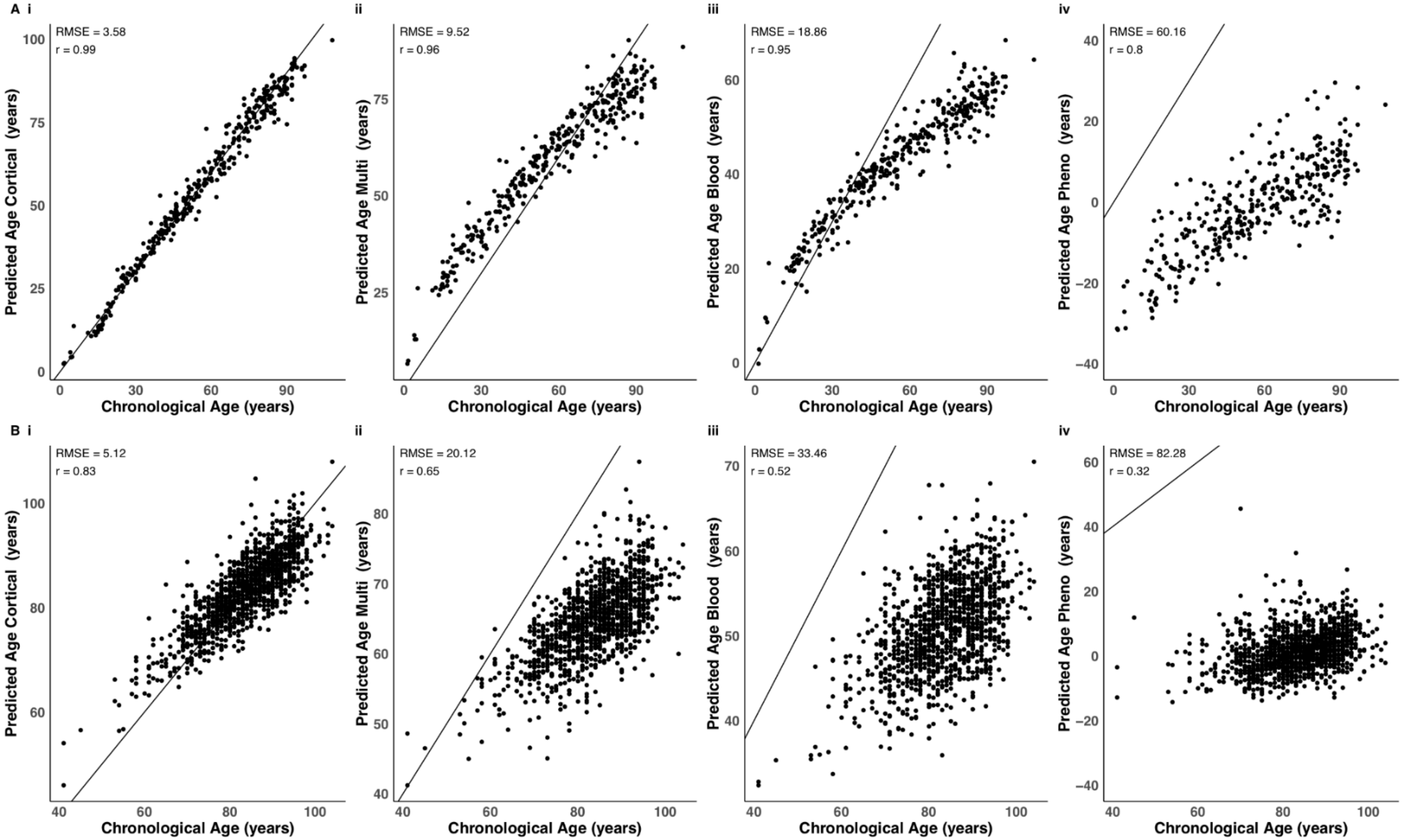


Figure 3.1: Comparison of chronological age with DNA methylation age derived using four DNA methylation age clocks. *Shown are comparisons of chronological age with predicted age in **(A)** the testing dataset (n = 350 cortical samples) and **(B)** the independent test dataset (n = 1221 cortical samples). DNAm age was predicted using four DNA methylation age clocks: **(i)** our novel DNAmClock$_{Cortical}$; **(ii)** Horvath's DNAmClock$_{Multi}$; **(iii)** Zhang's DNAmClock$_{Blood}$ and **(iv)** Levine's DNAmClock$_{Pheno}$. The x-axis represents chronological age (years) and the y-axis represents predicted age (years). Each point on the plot represents an individual sample. Our cortical clock out-performed the three alternative DNAm clocks across all accuracy statistics. DNA methylation age estimates derived using the DNAmClock$_{Multi}$ **(A(ii)** testing and **B(ii)** independent test) and the DNAmClock$_{Blood}$ (**A(iii)** testing and **B(iii)** independent test) appear to have a non-linear relationship with chronological age.* DNAmClock$_{Cortical}$ = Cortical DNA methylation age Clock; DNAmClock$_{Multi}$ = Multi-tissue DNA methylation age clock; DNAmClock$_{Blood}$ = Blood DNA methylation age clock and DNAmClock$_{Pheno}$ = Pheno Age DNA methylation age clock.*

3.3.3 Increased prediction accuracy of the novel cortex clock in cortical tissue compared to existing DNAm clocks

We used the $\text{DNAmClock}_{\text{Cortical}}$ to estimate DNAm age in both the testing (n = 350 cortex samples) and independent test (n = 1,221 cortex samples) datasets, and compared the estimates to those derived using $\text{DNAmClock}_{\text{Multi}}$, $\text{DNAmClock}_{\text{Blood}}$ and $\text{DNAmClock}_{\text{Pheno}}$. The $\text{DNAmClock}_{\text{Cortical}}$ predicted age accurately in the testing dataset and there was a strong correlation between DNAm age and age (r = 0.99; **Table 3.2** and **Figure 3.1**(**i**)). In the independent test dataset, which consisted predominantly of older samples, our clock also performed well and was highly correlated with age (r = 0.83), outperforming $\text{DNAmClock}_{\text{Multi}}$ (r = 0.65), $\text{DNAmClock}_{\text{Blood}}$ (r = 0.52), and $\text{DNAmClock}_{\text{Pheno}}$ (r = 0.32) (see **Table 3.2**; **Figure 3.1B**(**i**)).The most striking differences were in the accuracy of the $\text{DNAmClock}_{\text{Cortical}}$ in comparison to previously developed DNAm clocks; it outperformed the three other clocks we tested across all accuracy statistics in both cortical datasets (**Table 3.2**). The biggest differences in accuracy can be seen in the independent test dataset (see **Figure 3.1B**; **Figure 3.1B** and **Supplementary Figure 3.7B**), in which the RMSE was 15 years more accurate when using the $\text{DNAmClock}_{\text{Cortical}}$ (RMSE: 5 years) than the $\text{DNAmClock}_{\text{Multi}}$ (RMSE: 20 years), 28 years more accurate than the $\text{DNAmClock}_{\text{Blood}}$ (RMSE: 33 years) and 77 years more accurate than the $\text{DNAmClock}_{\text{Pheno}}$ (RMSE: 82 years). This is further supported when looking at the how much of the variation in DNAm age is explained by age, where the $\text{DNAmClock}_{\text{Cortical}}$ was the best fitting model in both cortical datasets (testing dataset R^2 = 0.98 independent test sample R^2 = 0.65) in comparison to the three other clocks, with age explaining the least variance in DNAm age estimated using the $\text{DNAmClock}_{\text{Pheno}}$ (testing dataset R^2 = 0.65; independent test sample R^2 = 0.10) (see **Table 3.2** for more details). The $\text{DNAmClock}_{\text{Pheno}}$ was consistently the most inaccurate at estimating age in the cortical data sets (RMSE: testing 60 years; independent test 82 years), followed by $\text{DNAmClock}_{\text{Blood}}$ (RMSE: testing 19 years; independent test 33 years) and the $\text{DNAmClock}_{\text{Multi}}$ (RMSE: testing: 10 years; independent test 20 years) (see **Table 3.2** for more details).

Table 3.2: Our novel cortex clock outperforms existing DNA methylation age algorithms in human cortex samples. *Accuracy statistics between DNAm age estimates and chronological age using our novel cortical clock, Horvath's multi-tissue clock (Horvath, 2013), Zhang's elastic net blood clock (Zhang at al., 2019) and Levine's Pheno Age clock (Levine at al., 2018) in both the testing (n = 350 cortical samples) and the independent test (n = 1,221 cortical samples) datasets. * RMSE = root mean squared error (years).*

	Testing dataset (n =350)				**Independent test dataset (n = 1221)**			
	Cortical Clock	**Multi-tissue Clock**	**Blood Clock**	**Pheno Age Clock**	**Cortical Clock**	**Multi-tissue Clock**	**Blood Clock**	**Pheno Age Clock**
Correlation (r)	0.99	0.96	0.95	0.8	0.83	0.65	0.52	0.32
RMSE (years)	3.58	9.52	18.86	60.16	5.12	20.12	33.46	82.28

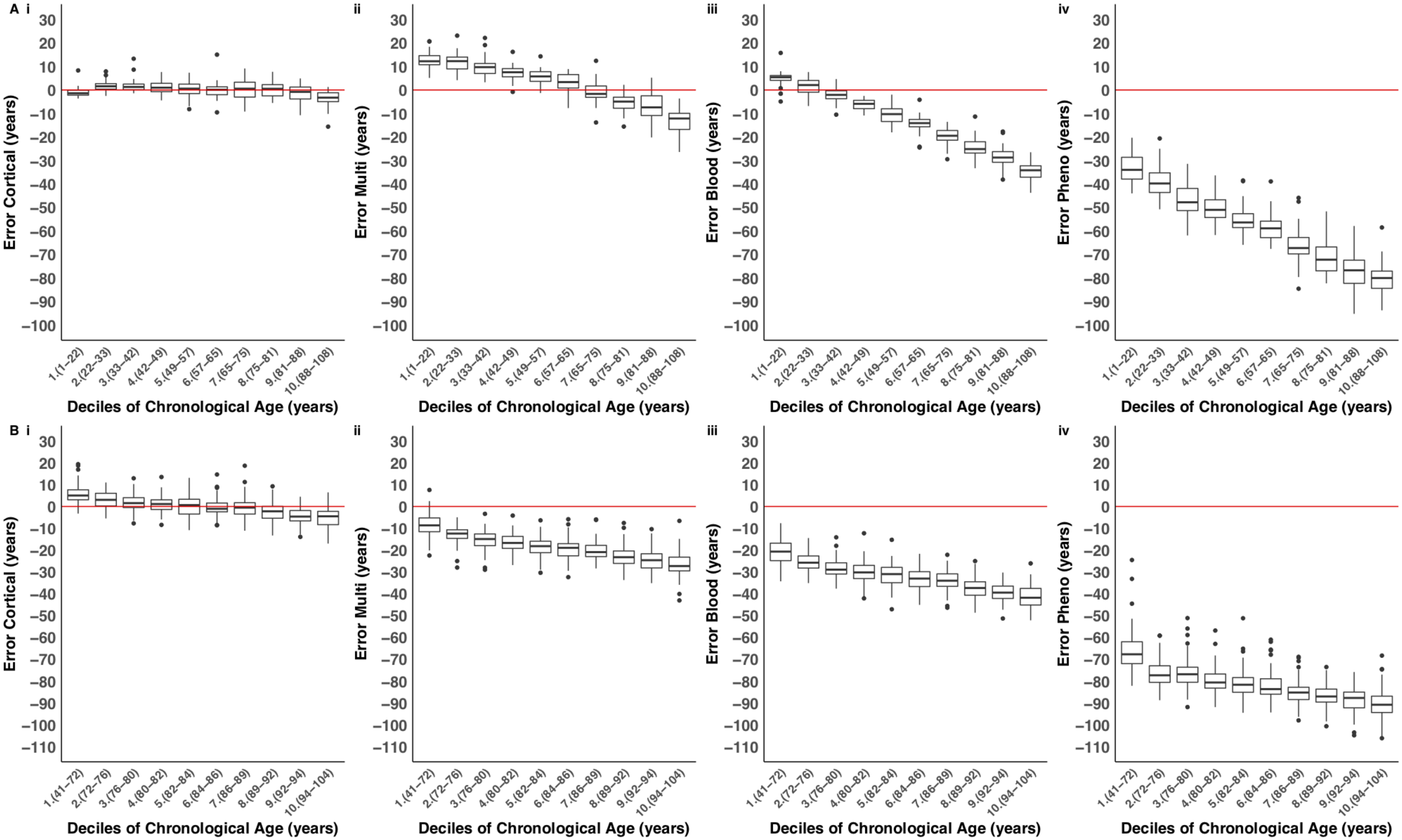


Figure 3.2: The cortical DNA methylation age clock has elevated accuracy in human cortex samples across the lifespan. *Shown is the distribution of the error (DNA methylation age - chronological age) for each age decile in (A) the testing dataset (n = 350 cortical samples) and (B) the independent test dataset (n = 1221 cortical samples) for each of the four DNA methylation age clocks: (i) our novel DNAmClock$_{Cortical}$; (ii) Horvath's DNAmClock$_{Multi}$; (iii) Zhang's DNAmClock$_{Blood}$*

and (iv) Levine's DNAmClock$_{Pheno}$. Deciles were calculated by assigning chronological age into ten bins and are represented along the x-axis by the numbers one to ten, followed by brackets which display the age range included in each decile. The ends of the boxes are the upper and lower quartiles of the errors, the horizontal line inside the box represents the median deviation and the two lines outside the boxes extend to the highest and lowest observations. Outliers are represented by points beyond these lines. The red horizontal line represents perfect prediction (zero error). Our novel DNAmClock$_{Cortical}$ (A(i) testing and B(i) independent test) consistently had the smallest error across the age groups, shown by the tightness of the boxplot distributions along the zero-error line. The DNAmClock$_{Multi}$ over-predicted younger ages (deciles 1-5 in A(ii)), shown by boxplots distributions which are above the zero-error line, and under predicted older ages (deciles 8-10 in A(ii) and deciles 1-10 in B(ii)), shown by boxplot distributions below the zero-error line. The DNAmClock$_{Blood}$ (A(iii) testing and B(iii) independent test) and the DNAmClock$_{Pheno}$ (A(iv) testing and B(iv) independent test) consistently under predicted age, with under prediction of DNA methylation age increasing with chronological age.

** DNAmClock$_{Cortical}$ = Cortical DNA methylation age Clock; DNAmClock$_{Multi}$ = Multi-tissue DNA methylation age clock; DNAmClock$_{Blood}$ = Blood DNA methylation age clock and DNAmClock$_{Pheno}$ = Pheno Age DNA methylation age clock.*

3.3.4 The relationship between age and DNAm age plateaus in old age

By definition, DNAm age is correlated with chronological age, meaning age is a potential confounder for analyses of Δ age; not adequately controlling for age increases the likelihood that false positive associations will be identified (El Khoury *at al.*, 2019). To assess associations between DNAm age and chronological age we used a mixed effects regression model (see **Methods 3.2**) and found that estimates from all four DNAm age clocks were significantly associated with age in the testing dataset (Bonferroni p< 0.005, see **Table 3.3**). Many studies of Δ age in health and disease control for age by using a linear model to regress out its effect (Marioni *at al.*, 2015; McKinney, Lin, Ding, Lewis, & Sweet, 2018) although one of the assumptions of this approach is that the prediction accuracy of the DNAm clock is consistent across the life course. If the accuracy varies non-linearly with chronological age, then simply including age as a linear covariate in association analyses will not sufficiently negate the confounding effect of age. We therefore sought to formally test the extent to which the prediction accuracy of the four clocks correlates with age by including an age squared term in the regression model. In the testing dataset all four clocks had a significant age squared term (**Table 3.3**), indicating that their predictive accuracy varies as a function of age. Specifically, all clocks were associated with a plateau where the difference between DNAm age and chronological age becomes larger as actual age increases (**Figure 3.2**). Importantly, however, out of the three first generation clocks the coefficient for the age squared term was smallest for the DNAmClock$_{\text{Cortical}}$ (beta = -1.64E-03, p= 1.94E-07), again highlighting that bespoke clocks can be used to minimise bias in subsequent analyses.

Table 3.3: The relationship between DNAm age and age (age and age^2) using different DNAm clock algorithms. *DNAm age was estimated using our novel cortical clock, Horvath's multi-tissue clock (Horvath, 2013), Zhang's elastic net blood clock (Zhang at al., 2019) and Levine's Pheno Age clock (Levine at al., 2018) in the "testing" dataset (n = 350 cortical samples), the "independent test" dataset (n =1221 cortical samples) and the blood dataset (n =1175 whole blood samples).*

		Testing dataset				**Independent test dataset**				**Blood dataset**			
		Beta	**SE**	**P**	**R^2**	**Beta**	**SE**	**P**	**R^2**	**Beta**	**SE**	**P**	**R^2**
Cortical Clock	**DNAm age vs age**	1.14	3.39E-02	2.86E-108	0.98	1.03	0.17	5.31E-09	0.65	0.58	0.06	5.37E-20	0.78
	DNAm age vs age^2	-1.64E-03	3.08E-04	1.94E-07	1.57E-03	-2.39E-03	1.08E-03	2.80E-02	1.47E-03	-2.05E-04	5.34E-04	0.70	1.61E-04
Multi-tissue clock	**DNAm age vs age**	1.08	4.14E-02	3.17E-83	0.93	0.68	0.16	3.51E-05	0.42	0.75	0.06	6.01E-30	0.80
	DNAm age vs age^2	-3.81E-03	3.75E-04	2.45E-21	0.02	-1.76E-03	1.02E-03	8.50E-02	1.39E-03	-1.15E-03	5.49E-04	0.04	5.66E-04
Blood Clock	**DNAm age vs age**	0.82	3.41E-02	1.30E-74	0.90	0.64	0.18	3.00E-04	0.26	1.14	0.05	9.50E-111	0.94
	DNAm age vs age^2	-3.16E-03	3.09E-04	1.81E-21	0.02	-2.08E-03	1.09E-03	5.70E-02	2.30E-03	-2.26E-03	3.90E-04	8.47E-09	1.61E-03
Pheno Age Clock	**DNAm age vs age**	0.57	6.89E-02	3.19E-15	0.65	-0.35	0.23	1.27E-01	0.10	0.63	0.08	1.86E-13	0.75
	DNAm age vs age^2	-1.79E-03	6.25E-04	4.47E-03	0.01	3.53E-03	1.42E-03	1.40E-02	0.01	6.22E-04	7.21E-04	0.39	5.41E-05

3.3.5 Higher cortical DNAm age is associated with decreased neuronal cell proportions

Many sample-related and technical factors can influence analyses of DNAm in post-mortem cortex tissue including sex, neuronal cell proportions, PMI and experimental batch effects. To assess associations between DNAm age and these variables we used a mixed effects regression model (see **Methods 3.2**) and after correcting for multiple comparisons ($p < 0.005$) found no association between DNAm age and sex ($p=0.03$), PMI ($p=0.54$) or batch ($p=0.38$) (see **Supplementary Table 3.9**).In contrast there was a significant association between neuronal cell proportion estimates derived from the DNAm data (beta = -8.72, p= 9.57E-36; see **Supplementary Table 3.9**) and DNAm age, indicating that individuals who are predicted as older using the $\text{DNAmClock}_{\text{Cortical}}$ have lower neuronal cell proportions. This correlation is not surprising as other clocks have been widely reported to associate with differences in cell-type proportions (Horvath & Ritz, 2015; Levine *at al.*, 2018) and it is known that the proportion of neuronal cells in the cortex changes with age. This result highlights the importance of, where possible, including cellular proportion variables as a covariate in any downstream analyses performed using DNAm clocks.

3.3.6 The cortical clock loses accuracy when applied to non-cortical tissues

To assess the specificity of the novel cortex clock we next applied each of the DNAm age clocks to a large whole blood DNAm dataset (n = 1175; age range = 28 - 98 years; mean age = 57.96 years). Although the $\text{DNAmClock}_{\text{Cortical}}$ actually performed remarkably well on whole blood (r = 0.88), with a similar predictive ability to the $\text{DNAmClock}_{\text{Multi}}$ (r = 0.90) (**Figure 3.3** and **Supplementary Figure 3.8**), there was a non-linear relationship between DNAm age and age estimated using this clock and a systematic under prediction of DNAm age in samples from people aged over 60 years (**Figure 3.3A**(**i**) and **Figure 3.3B**(**i**)). The $\text{DNAmClock}_{\text{Blood}}$ performed best on the blood dataset (r = 0.97), with age explaining the highest proportion of variation in DNAm age ($R^2 = 0.94$), outperforming the three other clocks (**Table 3.4; Figure 3.3; Supplementary Figure 3.8 and Supplementary Figure 3.9**), and providing further support for the notion that epigenetic clocks work optimally for the tissue-type on which they are calibrated. Of note, when limiting the age range of samples included in the blood cohort to be more comparable to the independent test dataset (age range limited to >55 years), the relationship between estimated and actual age is considerably lower for the three non-blood-specific clocks (r ~ 0.7) and the DNAmClock$_{\text{Blood}}$ (r = 0.88), reflecting the lower variability of age across samples in the dataset (see **Supplementary Figure 3.10**).

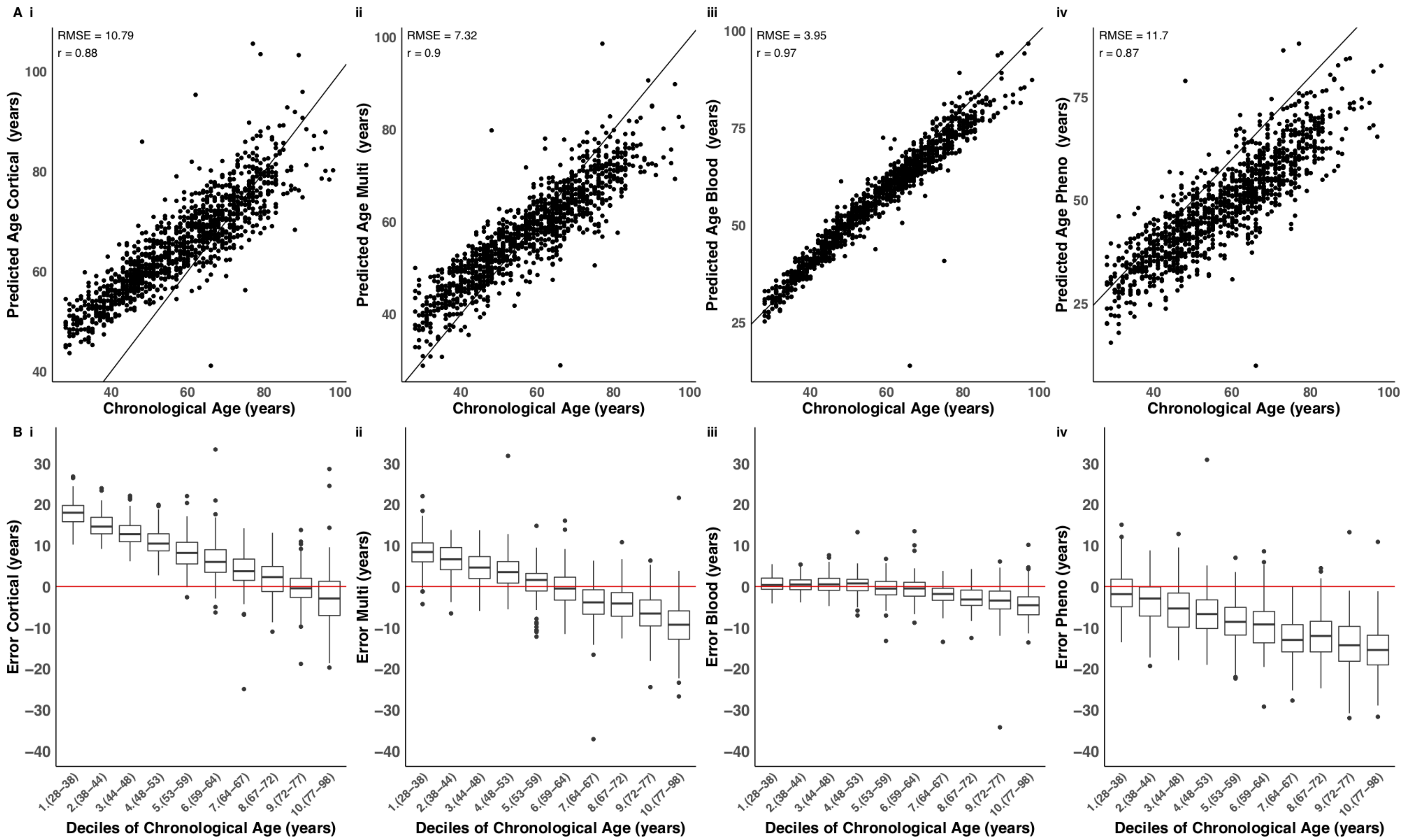


Figure 3.3: The blood based DNA methylation clock performs best in data derived from whole blood samples. (A) *Shown is a comparison of DNA methylation age estimates against chronological age in a large whole blood dataset (n = 1175), where DNAm age derived using four DNA methylation age clocks:* ***(i)*** *our novel DNAmClock$_{Cortical}$;* ***(ii)*** *Horvath's DNAmClock$_{Multi}$;* ***(iii)*** *Zhang's DNAmClock$_{Blood}$ and* ***(iv)*** *Levine's DNAmClock$_{Pheno}$. The x-axis represents chronological age (years),*

the y-axis represents predicted age (years). Each point on the plot represents an individual in the whole blood dataset. Our novel clock does not predict as well in blood compared to the cortex, although it has a similar predictive ability to Horvath's clock. The distribution of the error (DNA methylation age - chronological age) is presented in ***(B)*** *for each decile for each of the four DNA methylation clocks. Deciles were calculated by assigning chronological age into ten bins and are represented along the x-axis by the numbers one to ten, followed by brackets which display the age range included in each decile. The ends of the boxes are the upper and lower quartiles of the errors, the horizontal line inside the box represents the median deviation and the two lines outside the boxes extend to the highest and lowest observations. Outliers are represented by points beyond these lines. The red horizontal line represents perfect prediction (zero error).*

** $DNAmClock_{Cortical}$ = Cortical DNA methylation age Clock; $DNAmClock_{Multi}$ = Multi-tissue DNA methylation age clock; $DNAmClock_{Blood}$ = Blood DNA methylation age clock and $DNAmClock_{Pheno}$ = Pheno Age DNA methylation age clock.*

Table 3.4: The cortex clock is less accurate at estimating DNA methylation age algorithms in blood compared to cortex tissue samples. *Although still compares well to existing clock algorithms. Accuracy statistics between DNAm age estimates and chronological age using our novel cortical clock, Horvath's multi-tissue clock (Horvath, 2013), Zhang's elastic net blood clock (Zhang at al., 2019) and Levine's Pheno Age clock (Levine at al., 2018) in our blood dataset (n = 1,175 whole blood samples). RMSE = root mean squared error (years).*

	Cortical Clock	**Multi-tissue Clock**	**Blood Clock**	**Pheno Age Clock**
Correlation (r)	0.88	0.90	0.97	0.87
RMSE (years)	10.79	7.32	3.95	11.70

3.4 Discussion

Existing DNAm age clocks have been widely utilised for predicting age and exploring accelerated ageing in disease, although there is evidence of systematic underestimation of DNAm age in older samples, particularly in the brain (El Khoury *at al.*, 2019). We developed a novel epigenetic age model specifically for human cortex - the cortical DNAm clock (DNAmClock$_{\text{Cortical}}$) - built using an extensive collection of DNAm data from >1000 human cortex samples. Our model dramatically outperforms existing DNAm-based biomarkers for age prediction in data derived from the human cortex.

There are several potential causes of the systematic underestimation of DNAm age in the cortex, especially in samples from older donors (aged over 60 years), when using existing DNAm clocks such as Horvath's DNAmClock$_{\text{Multi}}$ (Horvath, 2013), Zhang's DNAmClock$_{\text{Blood}}$ (Zhang *at al.*, 2019) and Levine's DNAmClock$_{\text{Pheno}}$ (Levine *at al.*, 2018). First, it may be a consequence of the distribution of ages in the training data used in existing clocks; these clocks were derived using samples containing a relatively small proportion of samples from human brain and/or from older people. Second, as there is evidence for cell-type and tissue-specific patterns of DNAm (Mendizabal *at al.*, 2019), the observed imprecision may reflect a consequence of underfitting the model across tissues. Third, the relationship between DNA methylation and age may not be linear across the lifespan, and a non-linear model is needed to capture attenuated effects in older samples. This would be comparable to the transformation required to accurately predict DNAm age for younger samples (0-20 years), where the association between age and with DNA methylation is of larger magnitude.

Our data suggest that both tissue-specificity and the age of samples included in the training dataset influence the precision of DNAm age estimators, as shown by the increase in accuracy when using our cortical clock relative to existing clocks in human cortex tissue samples. This notion is further supported by the accuracy we found using the blood-based estimators on a large blood dataset. Our observations suggest that tissue type has a major influence on the accuracy of DNAm age clocks, and to accurately predict age it is important to use a clock calibrated specifically for the tissue from which samples have been derived. Our data demonstrate that the performance of existing

DNAm clocks varies considerably across ages and is diminished in samples from older donors. This is particularly important to consider when assessing DNAm age in the context of diseases and phenotypes that are associated with older age such as dementia and other types of neurodegenerative disease. Our results show that it is important to use a clock that has been trained using samples from the relevant age group; the training data used in the development of the $\text{DNAmClock}_{\text{Cortical}}$ included a good representation of older samples, meaning it overcomes the systematic underestimation of DNAm age in the older that was observed with existing clocks. It is also important to consider the distribution of ages in the training dataset (e.g. minimum, maximum, median, first and third quartiles), as this can influence the predictor and lead to biases if not representative of the datasets it will be applied to.

The importance of developing tissue-specific estimators is supported by other recently developed tissue-specific clocks including DNAm age predictors for whole blood (Zhang *at al.*, 2019), human skeletal muscle (Voisin *at al.*, 2020) and human bone (Gopalan *at al.*, 2019), which all out perform pan-tissue clocks in samples from the specific tissues in which they were trained. It is known that DNA methylation patterns are distinct between tissue and cell types (Mendizabal *at al.*, 2019), and it is therefore not surprising that DNAm age estimation models would differ in accuracy across tissue types. As technologies for profiling DNAm in purified cell populations from bulk tissue become more accessible, future clocks should be developed for purified populations of individual cell-types to overcome issues of cellular heterogeneity in complex tissues such as the brain. Furthermore, our finding that the $\text{DNAmClock}_{\text{Cortical,}}$ like other clocks, is associated with the proportion of specific cell-types in a given tissue sample highlights the importance of covarying for cellular heterogeneity in all subsequent analyses using values derived from epigenetic clocks.

Although a pan-tissue estimator such as Horvath's $\text{DNAmClock}_{\text{Multi}}$ has clear general utility, the trade-off between accuracy and practicality needs to be taken into consideration depending on the hypothesised question being tested. Applying one model across multiple tissues may lead to a suboptimal fit (for example, when applying a linear model where there is non-linearity), and the performance of such a clock would need to be tested in individual tissue-types. To assess the linearity of DNAm age predictors we investigated the association between DNAm age, and age squared. Of note, as age explains less of the variation in DNAm age in the second generation clocks (where the primary aim is to predict health outcomes) including the DNAmClock$_{\text{Pheno}}$, adding an age-squared term may be an unsuitable measure to address non-linearity where these predictors are applied. Adding the squared variable allowed us to more accurately model the effect of age in the three first generation clocks (where the primary aim is to predict age), which could have a non-linear relationship with DNAm age. The DNAmClock$_{\text{Cortical}}$ was the most linear in terms of fitting DNAm age against actual age. Although age squared terms were significantly associated with DNAm age in the testing data using all estimators, the higher significance of the age squared term in the cortex-specific clock suggests that of all the clocks, our model is the least biased. However, as indicated by the relationship between DNAm and age squared, we need to consider the possibility that fitting a linear model might not be the best approach, and to account for this possibility we recommend that future age-acceleration analyses control for age squared terms. Due to the nature of DNAm clocks, Δ age estimated using existing clocks is highly correlated with chronological age (El Khoury *at al.*, 2019). If age is not controlled for it could lead to spurious associations with health outcomes, which are driven by age and not the variable of interest. Furthermore, as the prediction is less precise in older individuals, even where DNAm is regressed on chronological age, the residual may still be associated with age, potentially leading to false positive associations. Recent studies have found associations between accelerated DNAm age in human brain and neurodegenerative phenotypes (Levine *at al.*, 2018, 2015). Our findings suggest that previous associations with age-associated phenotypes may have been confounded by a lack of robust calibration to estimate DNAm age in human cortex from older donors; caution is warranted in interpreting reported results that have been generated using a non-tissue specific predictor. Future work will focus on applying our novel DNAmClock$_{\text{Cortical}}$ to existing cohorts with DNAm data and detailed measures of neuropathology. While DNAm age is a useful indicator of age, it may not be the best indicator of health disparities between individuals with brain disorders.

In summary, we show that previous epigenetic clocks systematically underestimate age in older samples and do not perform as well in human cortex tissue. We developed a novel epigenetic age model specifically for human cortex. Our findings suggest that previous associations between predicted DNAm age and neurodegenerative phenotypes may represent false positives resulting from suboptimal calibration of DNAm clocks for the tissue being tested and for phenotypes that manifest at older ages. The age distribution and tissue type of samples included in training datasets need to be considered when building and applying epigenetic clock algorithms to human epidemiological or disease cohorts.

3.5 Acknowledgements

We would like to gratefully acknowledge all donors and their families for the tissue provided for this study. Human post-mortem tissue was obtained from the South West Dementia Brain Bank, London Neurodegenerative Diseases Brain Bank, Manchester Brain Bank, Newcastle Brain Tissue Resource and Oxford Brain Bank, members of the Brains for Dementia Research (BDR) Network. We wish to acknowledge the neuropathologists at each centre and BDR Brain Bank staff for the collection and classification of the samples.

3.6 Funding

G.S. was supported by a PhD studentship from the Alzheimer’s Society. E.H., J.M., and L.C.S. were supported by Medical Research Council grant K013807. M.K. was supported by the University of Essex and ESRC (grant RES-596-28-0001). Data analysis was undertaken using high-performance computing supported by a Medical Research Council (MRC) Clinical Infrastructure award (M008924). DNA methylation data generated in the Brains for Dementia Research cohort was supported by the Alzheimer’s Society and Alzheimer’s Research UK (ARUK). Measurement of DNA methylation in The UK Household Longitudinal Study was funded through an enhancement to Economic and Social Research Council (ESRC) grant ES/N00812X/1. The BDR is jointly funded by Alzheimer's Research UK and the Alzheimer's Society in association with the Medical Research Council. The South West Dementia Brain Bank is part of the Brains for

Dementia Research program, jointly funded by Alzheimer's Research UK and Alzheimer's Society, and is also supported by BR*ACE* (Bristol Research into Alzheimer's and Care of the Elderly) and the Medical Research Council.

3.7 Competing interests

The authors declare that they have no competing interests.

3.8 Supplementary data

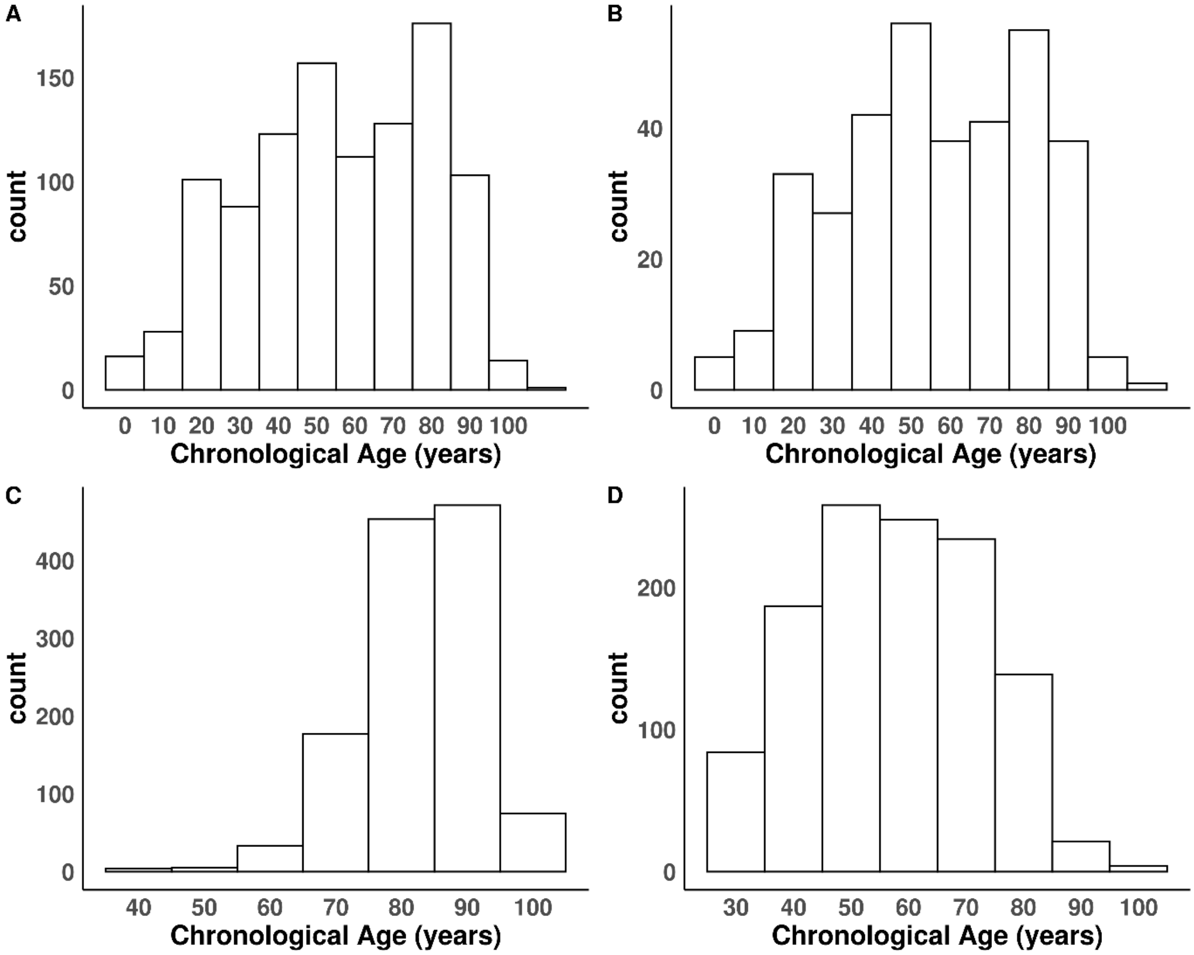


Figure 3.4: Histograms showing the distribution of chronological age in the datasets used in the study. (A) *The training dataset (n = 1,047 cortical samples);* ***(B)*** *the testing dataset (n = 350 cortical samples); (C) the independent test dataset (n = 1221 cortical samples) and (D) the whole blood dataset (n = 1175 whole blood samples).*

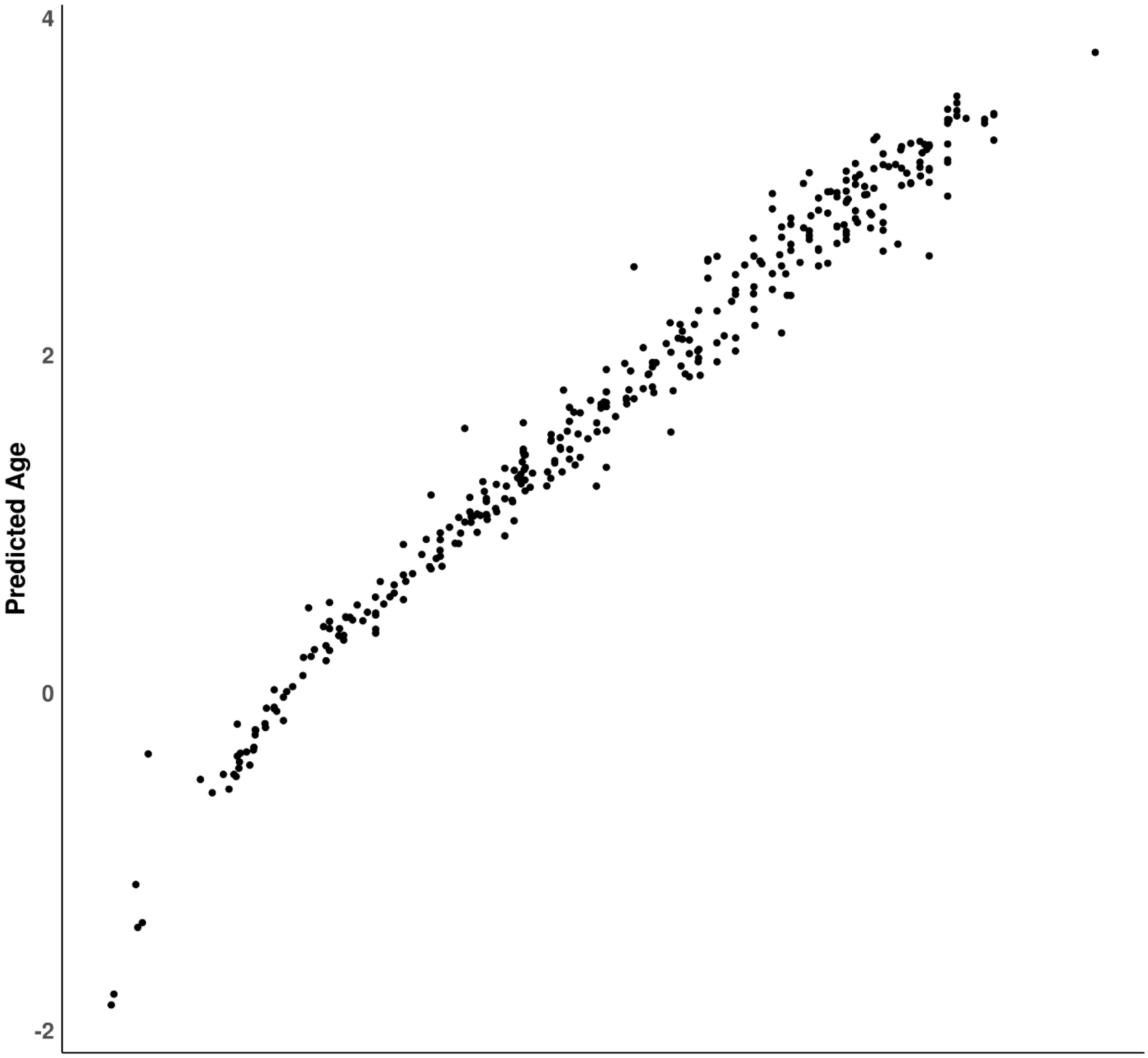


Figure 3.5: DNA methylation age has a logarithmic relationship with chronological age between the ages of 0-20 years. *From 20 years onward there is a linear relationship with chronological age. The x-axis represents chronological age (years), the y-axis represents predicted age prior to applying the anti-transformation function, whereby age between 0-20 years is log transformed, and ages 20+ are transformed to account for this. Each point on the plot is a sample from the testing dataset (n = 350 cortical samples).*

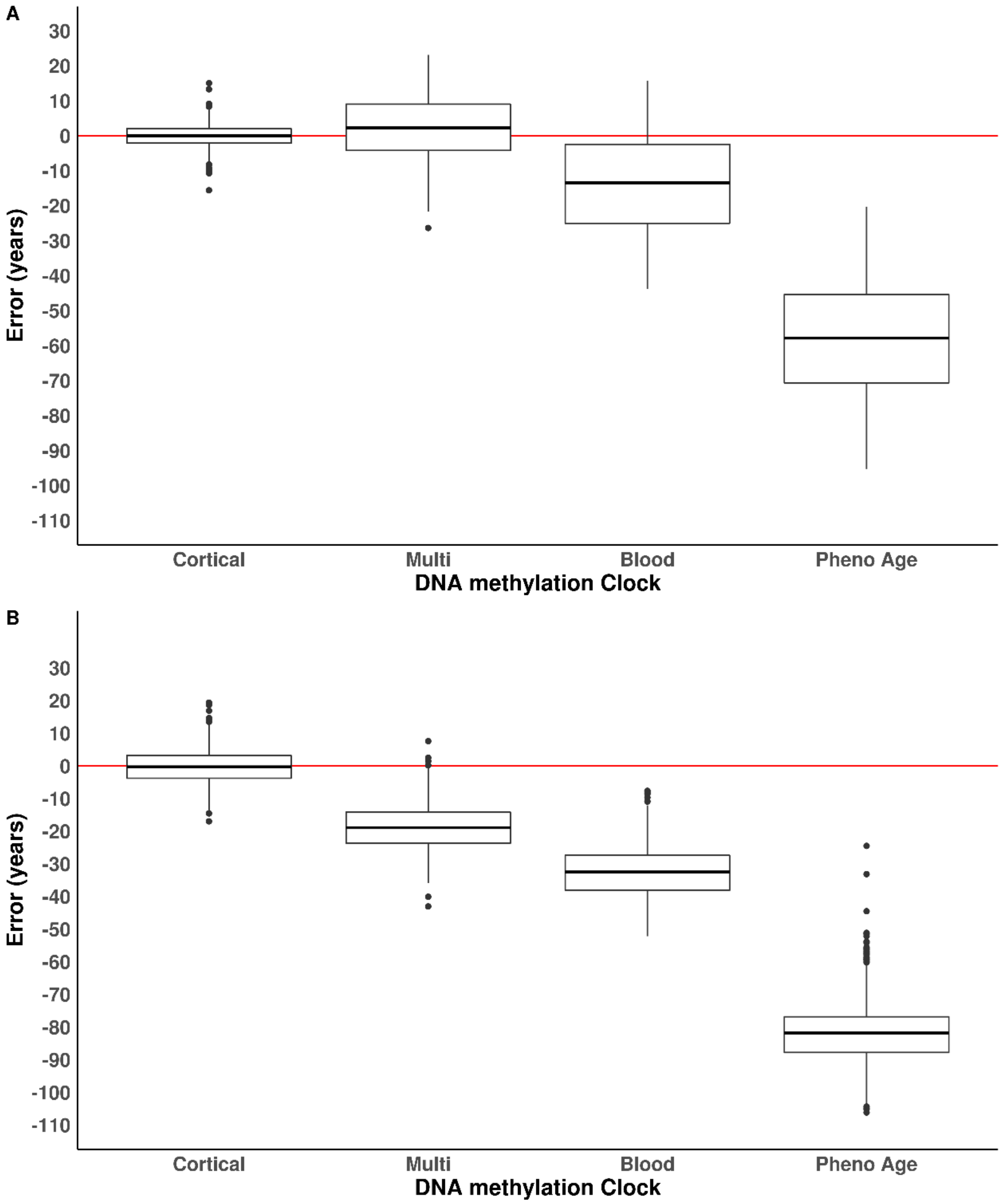


Figure 3.6: The cortical DNAm age clock has elevated accuracy in human cortex samples compared to existing DNAm clocks. *Shown is the distribution of the error (DNA methylation age - chronological age) for each of the four DNA methylation age clocks in* ***(A)*** *the testing dataset (n = 350 cortical samples) and* ***(B)*** *the independent test dataset (n = 1,221 cortical samples). The ends of the boxes are the upper and lower quartiles of the errors, the horizontal line inside the box represents the median deviation and the two lines outside the boxes extend to the highest and lowest observations. Outliers are represented by points beyond these lines. The red horizontal line represents perfect prediction (zero error).*

**DNAmClock$_{Cortical}$ = Cortical DNA methylation age Clock; DNAmClock$_{Multi}$ = Multi-tissue DNA methylation age clock; DNAmClock$_{Blood}$ = Blood DNA methylation age clock and DNAmClock$_{Pheno}$ = Pheno Age DNA methylation age clock.*

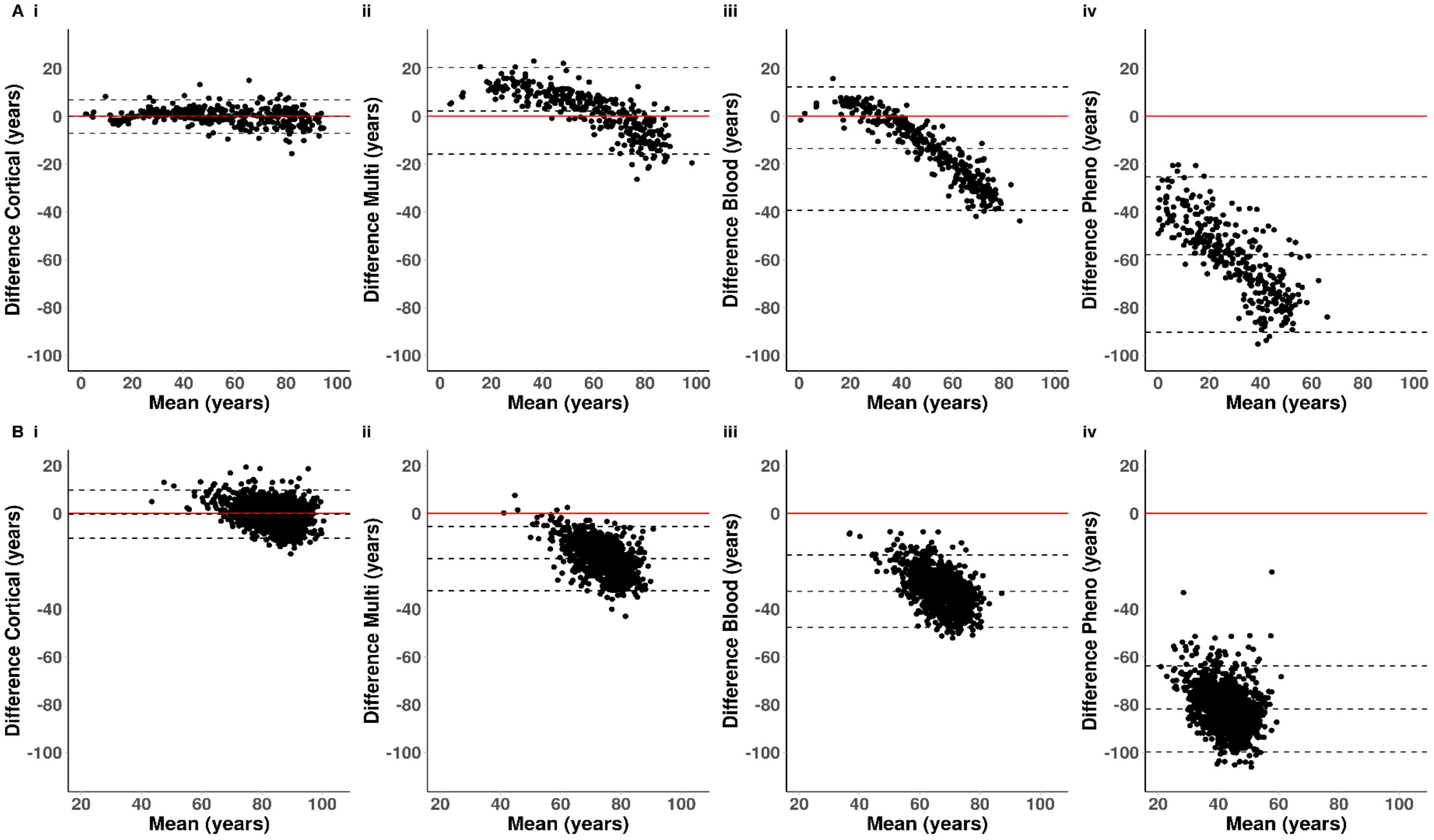


*Figure 3.7: **Bland-Altman plots highlighting enhanced performance of the cortical DNAm clock in human cortex tissue across the lifespan.** Shown is the mean difference between actual chronological age and estimated DNAm ages derived in **(A)** the testing dataset (n= 350 cortical samples) and **(B)** the independent test dataset (n = 1221 cortical samples), where DNAm age derived using four DNA methylation age clocks: **(i)** our novel DNAmClock$_{Cortical}$; **(ii)** Horvath's DNAmClock$_{Multi}$; **(iii)** Zhang's DNAmClock$_{Blood}$ and **(iv)** Levine's DNAmClock$_{Pheno}$ The dashed horizontal lines in each case are the differences between actual and chronological age +/- 1.96 * standard deviation; for normally distributed difference due to error 5% points would lie outside these. The solid horizontal line represents where the difference would be zero.*

**DNAmClock$_{Cortical}$ = Cortical DNA methylation age Clock; DNAmClock$_{Multi}$ = Multi-tissue DNA methylation age clock; DNAmClock$_{Blood}$ = Blood DNA methylation age clock and DNAmClock$_{Pheno}$ = Pheno Age DNA methylation age clock.*

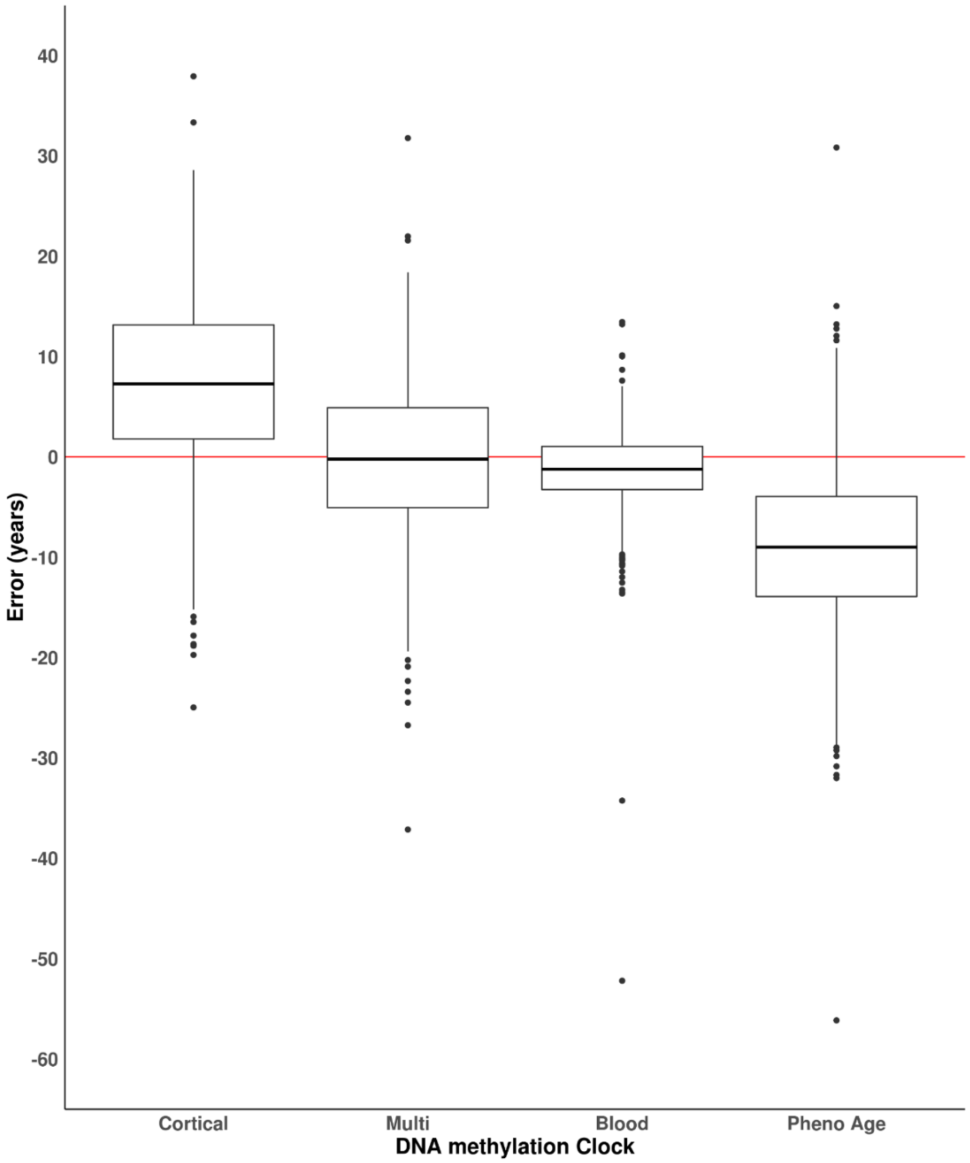


Figure 3.8: The blood DNAm age clock has better accuracy in human whole blood samples compared to non-tissue specific DNAm clocks. *Distribution of the error in years (DNA methylation age - chronological age) comparing four DNA methylation age clocks: our novel DNAmClock$_{Cortical}$, the DNAmClock$_{Multi}$, the DNAmClock$_{Blood}$ and the DNAmClock$_{Pheno}$ in the whole blood dataset (n = 1175). The ends of the boxes are the upper and lower quartiles of the errors, the horizontal line inside the box represents the mean absolute deviation and the two lines outside the boxes extend to the highest and lowest observations. Outliers are represented by points beyond these lines.*

**DNAmClock$_{Cortical}$ = Cortical DNA methylation age Clock; DNAmClock$_{Multi}$ = Multi-tissue DNA methylation age clock; DNAmClock$_{Blood}$ = Blood DNA methylation age clock and DNAmClock$_{Pheno}$ = Pheno Age DNA methylation age clock.*

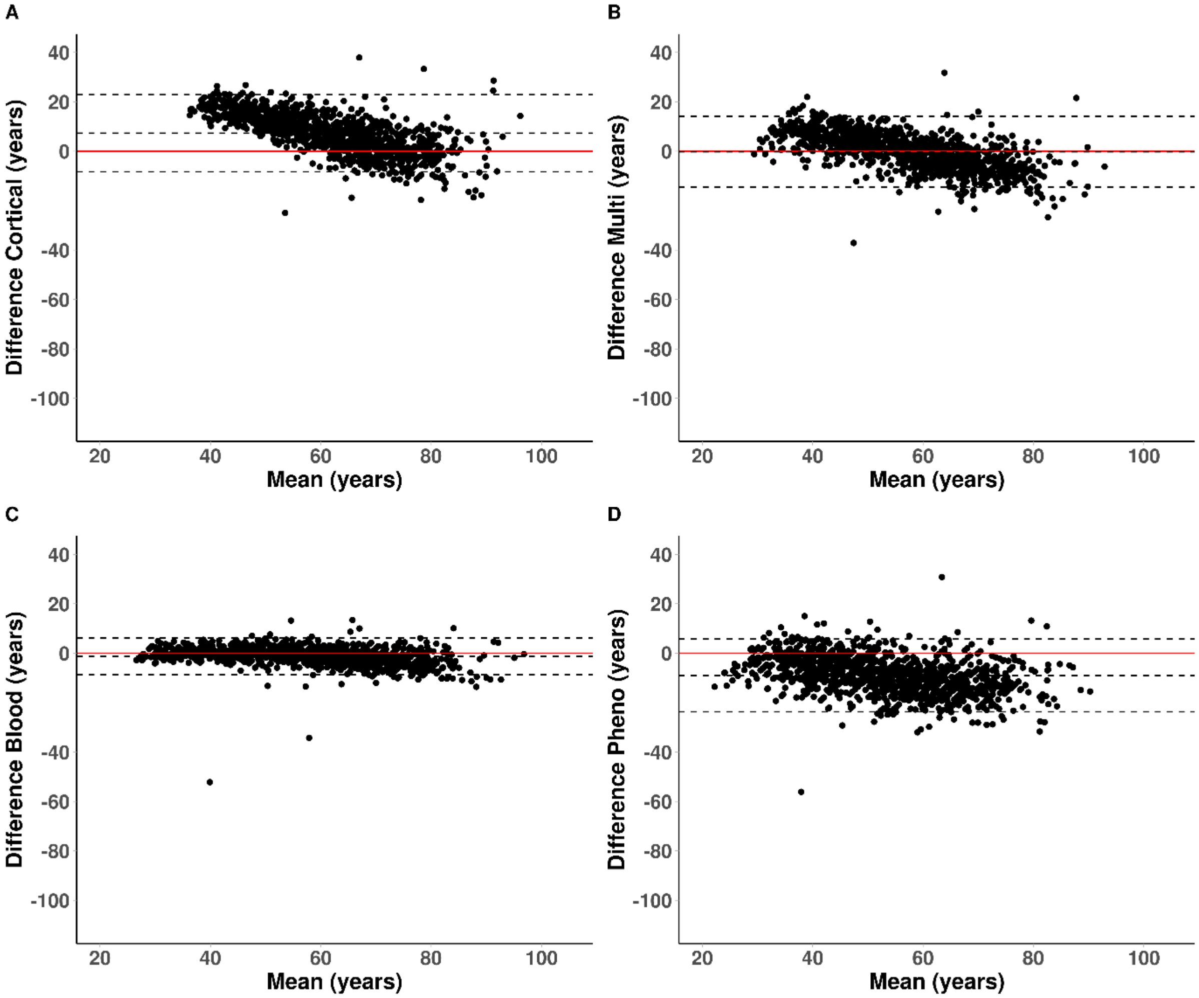


Figure 3.9: Bland-Altman plots highlighting elevated performance of blood based DNAm clocks in whole blood samples. *Mean-difference (Bland-Altman) plots showing the difference between DNA methylation age estimates against chronological age using **(A)** our novel DNAmClock$_{Cortical}$, **(B)** the DNAmClock$_{Multi}$, **(C)** the DNAmClock$_{Blood}$ and **(D)** the DNAmClock$_{Pheno}$ in the whole blood cohort (n = 1175). The dashed horizontal lines in each case are the differences between actual and chronological age +/- 1.96 * standard deviation; for normally distributed difference due to error 5% points would lie outside these. The solid horizontal line represents where the difference would be zero.*

**DNAmClock$_{Cortical}$ = Cortical DNA methylation age Clock; DNAmClock$_{Multi}$ = Multi-tissue DNA methylation age clock; DNAmClock$_{Blood}$ = Blood DNA methylation age clock and DNAmClock$_{Pheno}$ = Pheno Age DNA methylation age clock.*

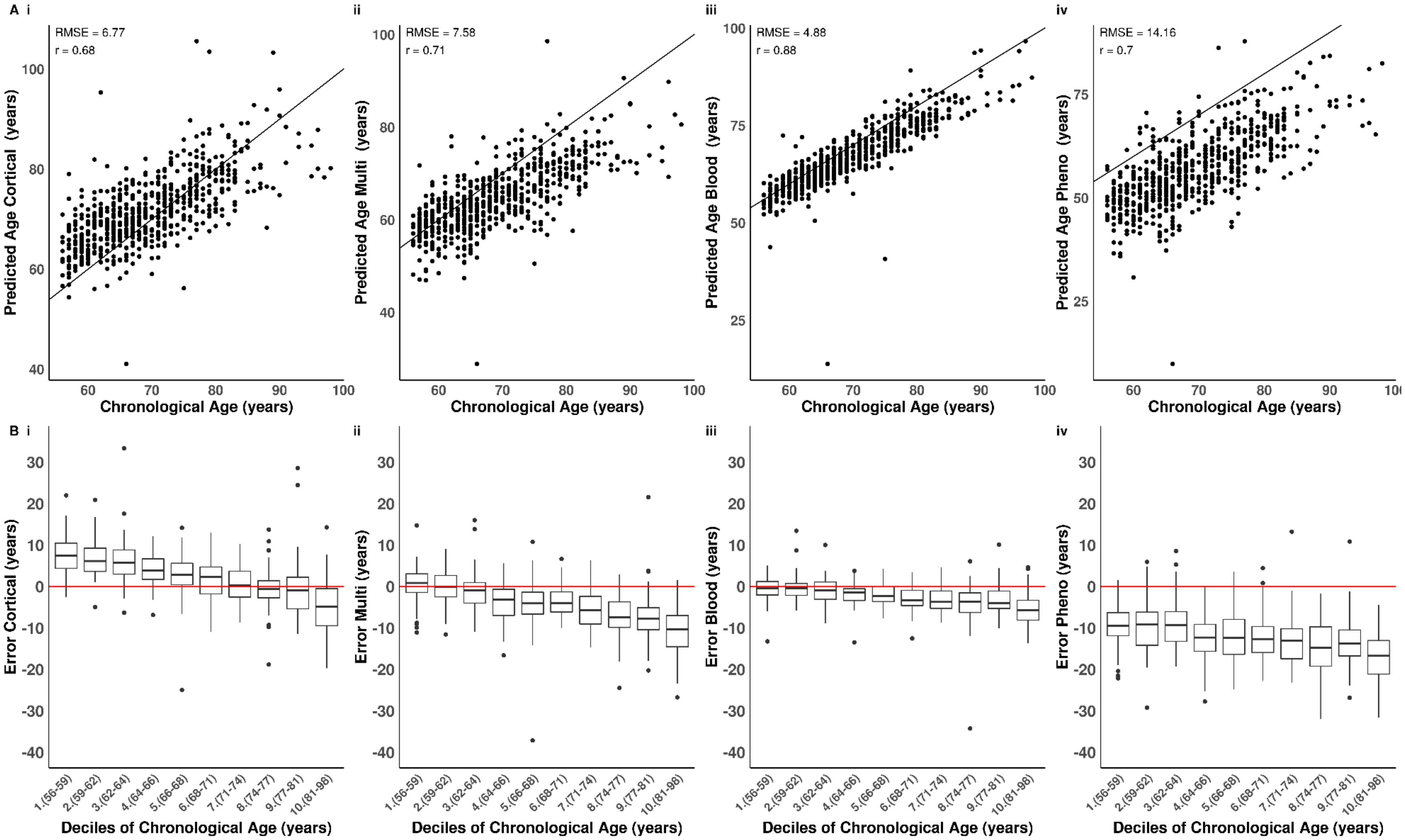


Figure 3.10: The blood-based DNA methylation clock performs best in data derived from whole blood samples. ***(A)*** *Shown is a comparison of DNA methylation age estimates against chronological age in samples >55 years old from a large whole blood dataset (n = 646), where DNAm age is derived using four DNA methylation age clocks:* ***(i)*** *our novel DNAmClock$_{Cortical}$;* ***(ii)*** *Horvath's DNAmClock$_{Multi}$;* ***(iii)*** *Zhang's DNAmClock$_{Blood}$ and* ***(iv)*** *Levine's DNAmClock$_{Pheno}$. The x-axis represents chronological age (years), the y-axis represents predicted age (years). Each point on the plot represents an individual in the whole blood dataset.*

Our novel clock does not predict as well as in the cortex, although it has a similar predictive ability to Horvath's clock. The distribution of the error (DNA methylation age - chronological age) is presented in ***(B)*** *for each decile for each of the four DNA methylation clocks. Deciles were calculated by assigning chronological age into ten bins and are represented along the x-axis by the numbers one to ten, followed by brackets which display the age range included in each decile. The ends of the boxes are the upper and lower quartiles of the errors, the horizontal line inside the box represents the median deviation and the two lines outside the boxes extend to the highest and lowest observations. Outliers are represented by points beyond these lines. The red horizontal line represents perfect prediction (zero error).*

**$DNAmClock_{Cortical}$ = Cortical DNA methylation age Clock; $DNAmClock_{Multi}$ = Multi-tissue DNA methylation age clock; $DNAmClock_{Blood}$ = Blood DNA methylation age clock and $DNAmClock_{Pheno}$ = Pheno Age DNA methylation age clock.*

Table 3.5: Sample characteristics of the training (cortex), testing (cortex), independent test (cortex) and whole blood datasets used in the development and evaluation of our novel cortical DNA methylation clock including the number of samples from each brain region. ** BA = Brodmann area; EC = Entorhinal cortex; HIP = Hippocampus; PFC = Prefrontal cortex; OCC = Occipital lobe; STG = Superior temporal gyrus; STR = Striatum; SD = Standard deviation; GEO = Gene Expression Omnibus*

Dataset				**Age (years)**				**Sex (n)**		**Illumina methylation array**	**Reference**	**GEO Accession number**
	Tissue Type	**Brain regions (n)**	**N**	**Mean**	**Median**	**Range**	**SD**	**Female**	**Male**			
Training	Cortical	**BA[9,11,17,25,41]** (284); **EC**(59); **HIP** (31); **PFC** (513); **STG** (83); **STR** (77)	1047	56.53	57	1.34-108	24.13	362	685	450K	(Jaffe *at al.*, 2016; De Jager *at al.*, 2014; Lunnon *at al.*, 2014; Pidsley *at al.*, 2014; Smith *at al.*, 2018, 2019; Wong *at al.*, 2019)	GSE74193; GSE59685; GSE80970; GPL13534; GSE43414;
Testing	Cortical	**BA[9,11,17,25,41]** (97); **EC**(18); **HIP** (11); **PFC** (167); **STG** (37); **STR** (20)	350	55.87	56	1.34-108	24.25	144	206	450K	(Jaffe *at al.*, 2016; De Jager *at al.*, 2014; Lunnon *at al.*, 2014; Pidsley *at al.*, 2014; Smith *at al.*, 2018, 2019; Wong *at al.*, 2019)	GSE74193; GSE59685; GSE80970; GPL13534; GSE43414;
BDR	Cortical	**PFC** (610) **OCC** (611)	1221	83.49	84	41-104	9.10	577	644	EPIC	-	
Understanding Society	Whole Blood	-	1175	57.96	59	28-98	14.97	686	489	EPIC	Hannon *at al.* (2018)	

Table 3.6: Age statistics for cohorts included in the training (cortex) and testing (cortex) datasets used in the development and evaluation of our novel cortical DNA methylation clock including the number of samples from each cohort. ** LBB2 = London Brain Bank 2; MS = Mount Sinai; NICHD = National Institute of Child Health and Human Development; ROSMAP = Religious Orders Study and Rush Memory and Aging Project; SD = Standard Deviation*

Cohort / Brain Bank	**N**	**Age (years)**						**Sex (n)**	
		1st quartile	**Median**	**Mean**	**3rd quartile**	**Range**	**SD**	**Female**	**Male**
Edinburgh	45	37	42	43.22	51	18-69	13.45	7	38
Harvard	54	16.25	26.5	28.39	39	2-60	15.79	11	43
Lieber Institute	566	28.42	45.14	42.45	55.42	1.34-96.98	18.6	197	369
LBB2	27	67.5	80	78.56	88.5	58-99	11.32	14	13
Maudsley	282	64	79	73.15	86	25-105	17.06	123	159
Montreal	137	30	41	44.47	57	18-90	18.28	28	109
MS	125	76	82	82.89	89	70-108	7.88	64	61
NICHD	18	15	21.5	26.61	42.75	4-67	18.55	0	18
Oxford	29	62	64	62.38	68	41-71	8.51	9	20
ROSMAP	114	77.32	81.82	81.45	86.18	65.99-89.73	5.45	53	61

Table 3.7: The Cortical Clock DNA methylation probes and their coefficients. *The Cortical Clock DNA methylation probes and their coefficients. Using elastic net regression we identified 347 DNA methylation sites which in combination optimally predict age in the human cortex. The sum of DNAm levels at these sites weighted by their regression coefficients provides the Cortical Clock DNA methylation age estimate. These coefficient values relate to a transformed version of age and therefore the linear combination of CpGs needs to be suitably transformed.*

Illumina Probe ID	**Coefficient**
(Intercept)	0.57768257
cg00059225	0.24595921
cg00088042	0.18271556
cg00252534	0.16822336
cg00297950	0.14175773
cg00384539	0.17820188
cg00491255	0.14450779
cg00521255	-0.2443627
cg00648582	-0.0391462
cg00771642	0.12772573
cg00924265	-0.0685163
cg00935119	-0.038119
cg00940577	-0.0612173
cg01091514	-0.0322298
cg01122755	0.06571091
cg01162920	-0.1973567
cg01194538	-0.1516094
cg01264729	-0.1524237
cg01311102	-0.1092163
cg01529637	0.02680964
cg01532168	0.33465593
cg01616394	-0.0187101
cg01639032	0.5150527
cg01641432	-0.0013768
cg01655150	-0.3276135
cg01745370	-0.0058643
cg01899542	-0.0509626
cg02046143	-0.0077791
cg02047661	-0.0171169
cg02357838	0.23797342
cg02361903	-0.5452234
cg02397497	0.07620684

cg02448244	0.03955555
cg02504211	0.13446075
cg02524236	-0.058474
cg02546818	0.01616236
cg02583546	-0.0359673
cg02795151	-0.0056271
cg02983424	0.05696411
cg03001484	-0.0258987
cg03025830	0.13631721
cg03040821	-0.0574853
cg03048488	-0.1256683
cg03591753	-0.4520824
cg03594801	-0.0691302
cg03613618	-0.1393547
cg03633073	0.24465057
cg03639603	0.07133576
cg03710354	0.00012886
cg03717616	-0.1918336
cg03854598	0.19543214
cg03864215	-0.3501439
cg04031134	0.02070123
cg04036898	-0.0001322
cg04055913	-0.0657139
cg04060163	-0.0111752
cg04078896	-0.0120007
cg04235075	0.20720766
cg04315771	-0.052222
cg04370442	0.07671797
cg04432319	0.37983386
cg04604946	-0.1166149
cg04684267	0.56103948
cg04686264	0.09996322
cg04704414	-0.0249742
cg04834794	0.86355274
cg04880546	0.03041421
cg04913265	0.05506598

cg05006304	-0.099641
cg05030953	-0.012153
cg05149386	-0.1053181
cg05213896	0.01844346
cg05280698	0.25970942
cg05333146	0.19252122
cg05362168	-0.0134173
cg05396044	0.9444237
cg05404236	0.09491394
cg05444541	-0.0118706
cg05634040	-0.0761376
cg05724492	0.03479139
cg05756780	0.06888371
cg05785488	0.58125152
cg05795849	-0.0377685
cg05839636	-0.0709078
cg06069616	-0.0128466
cg06144905	0.05548133
cg06236737	-0.4656742
cg06385324	0.08016557
cg06570818	-0.0040639
cg06645033	0.23463392
cg06648759	0.40674287
cg06711656	0.15663601
cg06751446	0.08353432
cg06852461	-0.0063935
cg07011538	-0.0337366
cg07120479	-0.4293158
cg07181374	-0.1310245
cg07197480	0.04978597
cg07461572	-0.0014446
cg07544187	0.19887884
cg07547549	0.01343572
cg07550554	0.2486126
cg07581257	-0.1444611
cg07806886	0.01131

cg08097417	0.55988863
cg08193650	0.07922242
cg08370996	0.124238
cg08594681	-0.0976416
cg08606497	-0.0136992
cg08708711	0.03540909
cg08727193	0.11187629
cg08786136	0.13167062
cg08952475	-0.0018342
cg08995871	0.0004663
cg09058748	0.00609294
cg09124496	-0.0726601
cg09178970	-0.0244848
cg09254686	-0.2087676
cg09363564	-0.1487783
cg09372546	-0.018437
cg09407967	0.07651973
cg09468836	0.02747043
cg09664474	-0.711323
cg09731141	-0.002565
cg09810078	0.03982331
cg09888620	0.22004768
cg09906752	-0.015249
cg09935271	-0.0195523
cg10027085	-0.0038364
cg10104252	-0.0272235
cg10225362	0.04192172
cg10232140	-0.0178474
cg10359006	-0.1063366
cg10378521	-0.2043615
cg10389229	0.00641507
cg10460946	-0.2376023
cg10501210	-0.305486
cg10574566	-0.0829055
cg10790698	-0.1026348
cg10851350	-0.0098493

cg10904972	-0.0784357
cg10924085	-0.0343679
cg10961323	0.0117429
cg11018337	0.10227319
cg11071401	0.00301298
cg11120115	-0.1945644
cg11397957	-0.1287726
cg11462165	0.1817344
cg11565355	0.11512885
cg11603443	0.11675982
cg11719412	-0.0717141
cg12024906	0.05166947
cg12100751	0.08174288
cg12175729	0.07132377
cg12423733	-0.0653703
cg12591491	0.24421376
cg12597389	0.5583175
cg12637942	0.17055797
cg12950231	-0.0656435
cg12978800	-0.1818211
cg13099374	-0.0473964
cg13259357	0.03688023
cg13298199	-0.0971264
cg13308350	-0.2068517
cg13327545	0.17202187
cg13477806	0.03348578
cg13718185	-0.0024661
cg13733708	-0.0027441
cg13744194	0.09010979
cg13755546	0.16217758
cg13806267	0.02601212
cg13848598	0.31150974
cg14189141	0.14486526
cg14242024	-0.0891087
cg14343652	-0.4916114
cg14489570	-0.0668353

cg14507337	0.04659403
cg14577472	0.0484974
cg14611683	0.05482077
cg14655122	0.2417095
cg14704921	0.21510624
cg14759277	0.11132052
cg14871932	0.10868066
cg15001747	0.43298364
cg15022387	0.04441422
cg15341124	0.90243633
cg15393490	-0.8410718
cg15410236	0.21389648
cg15540623	-0.2615207
cg15593298	0.33773506
cg15638207	-0.6916681
cg15665342	-0.0605596
cg15718663	0.11706067
cg15925792	-0.307196
cg15927455	-0.3584996
cg15974867	0.02314957
cg15988970	-0.0064278
cg16121765	-0.0123589
cg16142349	-0.2568507
cg16148593	-0.0981528
cg16206504	-0.0083124
cg16339238	0.00143445
cg16340422	-0.0292647
cg16359034	-0.0461923
cg16408865	-0.1234816
cg16604975	0.01666265
cg16643261	-0.1185516
cg16703882	0.09662046
cg16865965	-0.0417045
cg16867657	0.19738643
cg16909962	0.26618242
cg17040471	0.01800027

cg17086398	-0.2146144
cg17117277	0.41446544
cg17341113	1.44006579
cg17411994	-0.0762994
cg17435266	0.06161905
cg17523253	-0.0670554
cg17592231	0.68799031
cg17640485	-0.3129539
cg17693222	0.00318175
cg18077971	0.0867302
cg18125865	0.65840184
cg18247055	0.01999414
cg18279094	0.35656711
cg18427787	-0.0441448
cg18449120	0.00500354
cg18468088	0.22981384
cg18480946	0.00979905
cg18504218	-0.008073
cg18514820	0.22286731
cg18540328	0.02261323
cg18549036	0.00040885
cg18584803	-0.1976921
cg18604199	-0.0521892
cg18626323	-0.0303841
cg19028706	-0.0901704
cg19056004	-0.043383
cg19142026	0.24687437
cg19230755	0.00276837
cg19242851	-0.1708169
cg19399220	0.23167516
cg19416570	0.21729811
cg19699893	0.13548726
cg19724470	-0.0496928
cg19802138	0.01144858
cg19807317	-0.0546721
cg19955173	-0.0675901

cg20185454	0.18559061
cg20198242	-0.0058134
cg20429250	0.02217301
cg20495179	-0.3422078
cg20516262	-0.0036741
cg20583430	0.04465635
cg20594982	0.20060143
cg20603637	0.04123459
cg20627572	-0.037941
cg20697204	0.04133236
cg20747577	-0.0067257
cg20773033	-0.205063
cg20818778	0.03589532
cg21010435	0.00127567
cg21052766	-0.0232065
cg21186299	0.05465318
cg21218687	0.12544553
cg21415530	0.00975912
cg21581504	0.10314205
cg21801378	0.3019194
cg21826815	0.14908739
cg21851534	-0.0365721
cg21865150	-0.0024381
cg21967909	0.13778356
cg22007809	-0.0919512
cg22131013	-0.0213741
cg22285878	0.49092431
cg22320999	-0.0141227
cg22344793	0.00901416
cg22363327	0.0637605
cg22531668	0.03617343
cg22661556	0.10352047
cg22714290	-0.0507599
cg22884541	0.14417822
cg22900415	0.04351403
cg23051272	-0.0698149

cg23091758	0.11685302
cg23124451	-0.3676633
cg23163283	-0.1166281
cg23166770	-0.4055405
cg23174607	0.47022718
cg23201812	-0.0043506
cg23352942	-0.0709361
cg23474190	-0.0992123
cg23606718	0.25215709
cg23636833	0.05409979
cg23661344	-0.0953841
cg23662138	-0.1453696
cg23684204	0.36656376
cg23813012	0.09645841
cg23896431	0.10834819
cg23939875	-0.0027469
cg23995914	0.13059076
cg24085039	-0.0266476
cg24231804	-0.0610665
cg24287218	-0.057864
cg24346776	0.03688965
cg24377285	-0.0034626
cg24420164	0.01149875
cg24483655	-0.1180724
cg24567591	0.00663023
cg24715767	0.03440595
cg24718465	0.09807228
cg24933925	-0.0626709
cg24990808	-0.0417352
cg25007705	-0.2476066
cg25018458	0.51960078
cg25090514	0.5000666
cg25114611	-0.1558847
cg25201359	-0.3919139
cg25459323	-0.0912804
cg25922329	-0.0449031

cg26092675	0.09283213
cg26115633	0.27801401
cg26161329	0.64983683
cg26242531	-0.3438033
cg26306636	0.67247185
cg26332630	-0.0217055
cg26377000	0.06837371
cg26384036	0.08320489
cg26472036	-0.2159495
cg26490949	0.39372896
cg26542283	0.06460673
cg26645401	0.23806474
cg26660754	-0.3879276
cg26726230	0.08107185
cg26782108	0.09393526
cg26856080	0.25484991
cg26885220	0.15739762
cg26952697	-0.0575663
cg26952796	-0.0916588
cg27013696	0.18890756
cg27043838	0.00732047
cg27134767	-0.1281932
cg27230784	0.11078553
cg27388680	-0.1024366
cg27414487	-0.0433871
cg27529628	0.01778693
ch.2.1904845F	-0.1201514
ch.2.71774667F	-0.2699354

Table 3.8: Overlap between the probes which collectively make up the Cortical DNA methylation clock and three other DNA methylation age clocks: Horvath's Multi tissue DNA methylation Clock, Zhang's Blood DNA methylation Clock and Levine's Pheno Age DNA methylation Clock.

Overlap Horvath Multi Tissue	Overlap Zhang Blood	Overlap Levine Pheno Age
cg06144905	cg02046143	cg06144905
cg08370996	cg03025830	cg19724470
cg19724470	cg04604946	cg21801378
cg21801378	cg04684267	cg23124451
cg23124451	cg06648759	cg25459323
	cg07547549	
	cg08097417	
	cg09935271	
	cg15393490	
	cg16867657	
	cg18468088	
	cg21186299	
	cg23174607	
	cg23606718	
	cg23995914	

***Table 3.9: The relationship between DNAm age and technical and biological variables**. DNAm age was estimated using our novel cortical clock in the independent test dataset (n=1,221). DNAm age was regressed against biological and technical variables (cell proportions, sex, post-mortem interval and batch). *DNAm = DNA methylation; PMI = Post-mortem interval*

	Beta	SE	P
Cell proportions	-8.72	0.67	9.57E-36
Sex	0.60	0.28	0.03
PMI	-3.90E-03	0.01	0.54
Batch	-0.03	0.03	0.38

3.9 Associations between age acceleration and neuropathology

3.9.1 Methods

I derived a measure of epigenetic age acceleration (EAA) for each of the four clocks tested in Chapter 3 by regressing DNAm age on chronological age, sex and cell proportions and extracting the residuals. I then investigated if EAA was associated with five neuropathology measures (Braak NFT stage, Thal phase, CERAD density, Braak LB stage and TDP-43 state) as well as AD status (defined as cases with Braak >4 and control with Braak <2) in the BDR cohort (see sections **3.2.2** and **4.4.1**). For Braak NFT stage, Thal phase, CERAD density, Braak LB stage I used mixed effects regression models using the *lmer and* package in r, including EAA as the independent variable, neuropathology as the dependent variable and individual as a random effect. Mixed effects logistic regression models were used for TDP-43 status and AD status (binary variables) but the framework remained the same.

3.9.2 Results and discussion

After controlling for multiple testing (Bonferroni $p < 0.05/24 = 0.002$) no associations between EAA and any of the four epigenetic clocks were identified (see **Table 3.10**). Although there is no evidence to suggest there is a relationship between EAA and neuropathology we cannot definitively conclude that there is no association, since these analyses were limited by power. However, these results suggest that while 1st generation epigenetic clocks are good predictors of age, they may not be directly associated with neuropathology. Previous studies reporting EAA associations with neuropathology using epigenetic clocks not built for the cortex should be interpreted with caution.

In contrast to the results in BDR, I recently collaborated on a study using the ROSMAP PFC dataset (excluding samples included in the training of the cortex clock). We found modest significant associations between EAA estimated using the cortical clock and neuropathology (Grodstein *at al.*, unpublished). The strength of the association was strongest when using the cortical clock in comparison to the multi-tissue clock, strengthening the hypothesis that the most optimal clock for a dataset should be trained in the relevant tissue type.

Table 3.10 Associations between epigenetic age acceleration and neuropathology.

	Beta	SE	P
Cortical clock - Braak NFT Stage	-2.31E-03	7.73E-03	0.77
Multi-tissue clock - Braak NFT Stage	2.10E-02	1.63E-02	0.20
Blood clock - Braak NFT Stage	3.95E-02	2.40E-02	0.10
Pheno Age clock - Braak NFT Stage	-3.50E-03	1.14E-02	0.76
Cortical clock - Thal Phase	9.22E-03	2.08E-02	0.66
Multi-tissue clock - Thal Phase	3.56E-02	1.60E-02	0.03
Blood clock - Thal Phase	3.42E-02	2.36E-02	0.15
Pheno Age clock - Thal Phase	3.56E-02	1.60E-02	0.03
Cortical clock - CERAD density	1.02E-02	2.74E-02	0.71
Multi-tissue clock - CERAD density	9.82E-03	2.08E-02	0.64
Blood clock - CERAD density	1.02E-02	2.74E-02	0.71
Pheno Age clock - CERAD density	9.82E-03	2.08E-02	0.64
Cortical clock - Braak LB Stage	4.01E-03	1.48E-02	0.79
Multi-tissue clock - Braak LB Stage	7.11E-03	1.13E-02	0.53
Blood clock - Braak LB Stage	1.55E-02	1.68E-02	0.35
Pheno Age clock - Braak LB Stage	7.11E-03	1.13E-02	0.53
Cortical clock – TDP-43 Status	5.27E-03	4.80E-03	0.27
Multi-tissue clock - TDP-43 Status	1.68E-03	3.58E-03	0.64
Blood clock - TDP-43 Status	-2.41E-03	5.36E-03	0.65
Pheno Age clock - TDP-43 Status	1.63E-03	3.56E-03	0.65
Cortical clock - AD status	-2.45E-03	6.35E-03	0.70
Multi-tissue clock - AD status	5.90E-03	4.95E-03	0.23
Blood clock - AD status	1.07E-02	7.12E-03	0.13
Pheno Age clock - AD status	-4.30E-04	3.46E-03	0.90

4 Epigenome-wide association study of neuropathology in the Brains for Dementia Research Cohort

4.1 Introduction

As described in **Chapter 1** (section **1.1**), dementia is an umbrella term used to describe a group of symptoms associated with global cognitive impairment. Dementia encompasses a number of neurodegenerative diseases, including Alzheimer's disease (AD), vascular dementia (VaD), dementia with Lewy bodies (DLB), Parkinson's disease (PD) and frontotemporal dementia (FTD) (Lobo *at al.*, 2000). Most neurodegenerative diseases are characterised by the aggregation of specific proteins in intracellular inclusions or extracellular aggregates within the brain (Ross & Poirier, 2004). Although the proteins and specific brain regions involved in the aetiology of neurodegenerative diseases differ, in all cases the progressive accumulation of these deposits ultimately leads to neuronal cell death and brain atrophy (Ross & Poirier, 2004).

4.1.1 Neuropathology of Alzheimer's disease

AD is the most common form of dementia, accounting for ~60-80% of cases (Alzheimer's Association, 2019). AD is characterised by two histopathological hallmarks: the accumulation of extracellular amyloid-beta (Aβ) plaques and deposit of intracellular neurofibrillary tangles of tau (NFT) (Braak, Alafuzoff, Arzberger, Kretzschmar, & Del Tredici, 2006; Thal, Rüb, Orantes, & Braak, 2002). There are several different measures used to quantify the burden and progression of neuropathology in AD brain. Braak NFT staging (Braak, Alafuzoff, *at al.*, 2006) provides a measure of NFT pathology. There are seven Braak NFT 7 stages (0-VI), starting with the accumulation of NFTs in the entorhinal cortex (Braak NFT stages I-II) and ending with the accumulation of NFTs in the neocortex (Braak NFT stages V-VI) (see **Figure 1.1**). Thal Phasing provides a measure of Aβ deposits (both diffuse and dense-core)(Thal *at al.*, 2002). There are six Thal phases (0-5). In phase 1 Aβ deposits are found in the neocortex and by phase 5 Aβ deposits have started to accumulate in the cerebellum and other brain regions (see **Figure 1.1**). Consortium to Establish a Registry for AD (CERAD) score (Mirra *at al.*, 1991) describes the density of neuritic Aβ plaques (dense-core) in three areas of the isocortex (frontal, temporal and parietal) and is measured on a four point scale ranging from no/ none to frequent/ high. A detailed description of neuropathology in AD is given in section **1.1.4.1.**

4.1.2 Neuropathology of other neurodegenerative diseases

Lewy-body (LB) pathology is involved in the pathogenesis of DLB and PD and involves the accumulation and of α-synuclein in neuronal cell bodies as Lewy bodies (LBs) and in neuronal cell processes (e.g. axons) as Lewy neurites (Outeiro *at al.*, 2019). LB body pathology is measured using Braak LB staging (Braak, Bohl, *at al.*, 2006). There are 7 Braak LB stages (0-6). In stage 1 α-synuclein starts to accumulate in the motor nucleus of the medulla oblongata and by stage 6 α-synuclein has spread through all of the neocortex and is detected in the premotor and motor regions. FTD diseases are primary TAR DNA-binding protein 43 (TDP-43) proteinopathies, where TDP-43 is the main driver in disease pathogenesis (Mackenzie, Rademakers, & Neumann, 2010). TDP-43 can be found in different parts of the brain depending on the FTD-subtype. There is currently no standardised measure for TDP-43 and its presence if often recorded as a binary yes/no. A detailed description of neuropathology in other neurodegenerative diseases is given in section **1.1.6.**

4.1.3 Neuropathological comorbidities in dementia

It has been recognised that dementia in older people is usually a consequence of multiple pathologies (Kapasi, DeCarli, & Schneider, 2017; Thomas *at al.*, 2020). Most frequently, for example, LB pathology is present in addition to the core AD pathologies, affecting ~50% of AD cases (Thomas *at al.*, 2020). Another common neuropathological comorbidity is the aggregation of TDP-43 (Thomas *at al.*, 2020) and individuals with AD usually have some level of TDP-43 (Wilson, Dugger, Dickson, & Wang, 2011). These comorbidities have been shown to increase cognitive impairment in non-AD cases and evidence suggests they contribute to the declining cognition in AD cases beyond that of Aβ and NFT pathology (Thomas *at al.*, 2020). This hypothesis is supported by recent research conducted by Thomas and colleagues (2020), who looked at the associations between neuropathology and cognitive decline (measured using the mental state exam and the clinical dementia rating) in ~40,000 individuals and found that TDP-43 and cerebral amyloid angiopathy (where amyloid builds up on the walls of the arteries in the brain) are associated with cognitive impairment to a similar magnitude as AD neuropathology; 63% of the individuals in their study diagnosed with AD had TDP-43 or cerebral angiopathy severe enough to explain the cognitive deficits independently of AD. It is therefore important to consider multiple pathologies in combination to better understand the aetiology of AD and other neurodegenerative diseases.

4.1.4 Standardised brain banking

The definitive presence of the neuropathological hallmarks of dementia can only be confirmed via post-mortem brain examinations. This led the development of standardised procedures in brain banking for dementia, including the semi-quantitative classification schemes for the different neuropathology features described in detail in section **1.1.4** (Braak, Alafuzoff, *at al.*, 2006; Braak, Bohl, *at al.*, 2006; Thal *at al.*, 2002). This has facilitated harmonisation across brain banks, enabling consistent classifications and increasing reliability and validity across studies. Recently, the Brains for Dementia Research (BDR) cohort was established with the aim of generating a large comprehensive neuropathological dataset from multiple brain banks using these standardised procedures, enabling the investigation of dementia through detailed phenotypic and multi-omics datasets (Francis, Costello, & Hayes, 2018). Both dementia patients and unaffected controls > 65 years of age were recruited to partake in routine longitudinal assessments collecting cognitive, clinical, lifestyle and psychometric data, prior to post-mortem brain donation (Francis *at al.*, 2018).

4.1.5 Epigenetic dysregulation in Alzheimer’s disease and other neurodegenerative diseases

Multiple studies suggest epigenetic dysregulation is associated with the progression of pathology seen in AD and other dementias (Roubroeks *at al.*, 2020; A. R. Smith *at al.*, 2016, 2019; R. G. Smith *at al.*, 2018; R. Smith *at al.*, 2021; Vasanthakumar *at al.*, 2020). Genome-wide changes in DNA methylation (DNAm) associated with disease and pathology have been explored using epigenome-wide association studies (EWAS), where DNAm at each site is independently tested for association with the trait of interest (Q. S. Li, Sun, & Wang, 2020; Lunnon *at al.*, 2014; Roubroeks *at al.*, 2020; R. Smith *at al.*, 2021) (see section **1.3.3** for more details). Unlike genetic variation, epigenetic signatures are tissue specific (Mendizabal & Yi, 2016) and therefore the tissue type used in EWAS is important to consider; most EWAS of neurodegenerative diseases have been conducted in regions of the cortex. In addition, the majority of AD EWAS have been conducted using the Illumina Infinium 450K Beadarray (450K array) (see **section 2.1.1**) which quantifies DNAm at > 450,000 DNAm sites (Bibikova *at al.*, 2011). Recently, Smith, Pishva and colleagues conducted a meta-analysis of AD EWAS studies (R. Smith *at al.*, 2021), combining data from six 450K array analyses of AD (N=1,453 unique individuals) to identify differential methylation associated with Braak NFT across multiple cortical regions. In their cross-cortex meta-analysis (N=1,408 donors) they identified 220 DMPs associated with Braak NFT stage, annotated to 121 genes (see **Figure 4.1**). These results support a role for differential DNAm across many genes in AD.

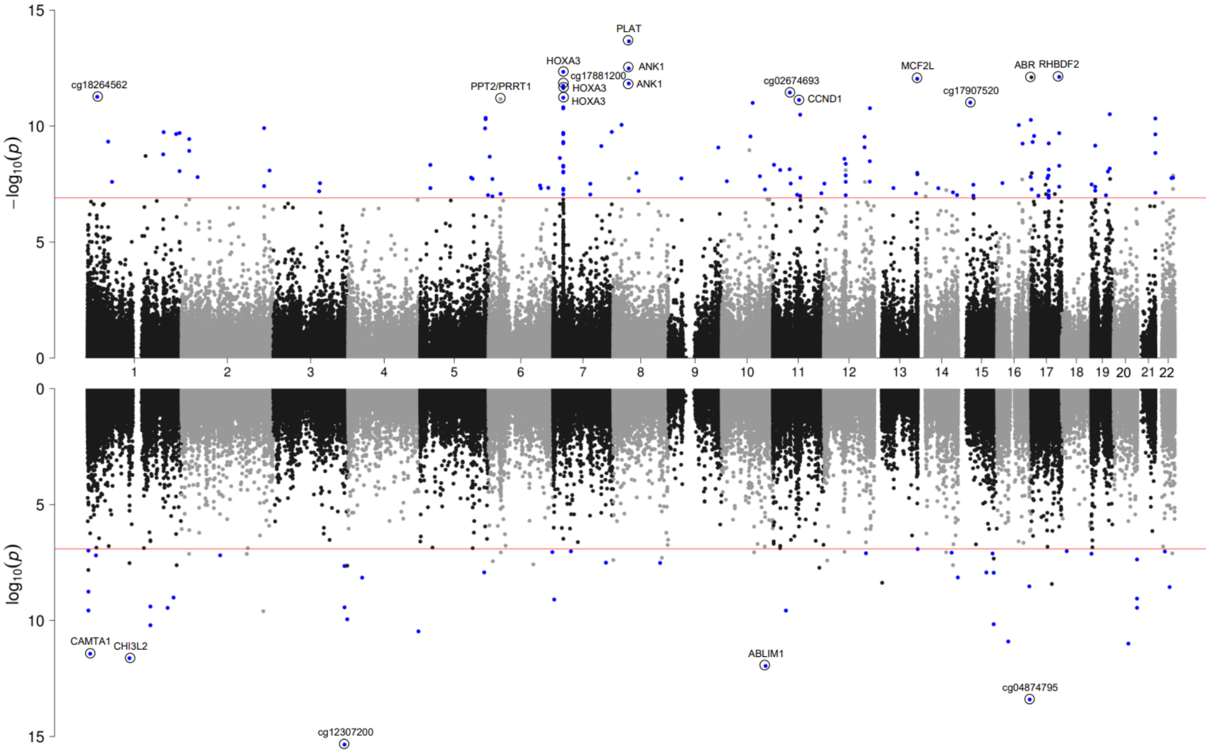


Figure 4.1: A Miami plot of the cross-cortex meta-analyses of Braak NFT Stage. *Probes shown above the x-axis indicate hypermethylation with higher Braak stage, whilst probes shown below the x-axis indicate hypomethylation with higher Braak stage. The chromosome and genomic position are shown on the x-axis. The Y-axis shows −log10(p). The red horizontal lines indicate the Bonferroni significance level of p< 1.238 x 10−7. Probes with a methylation (beta) effect size (ES: difference between Braak 0-Braak VI) ≥ 0.01 and p< 1.238 x 10−7 are shown in blue. The 20 most significant DMPs are circled on the plot and Illumina UCSC gene name is shown if annotated, or CpG ID if unannotated. Figure and legend taken from (R. Smith et al., 2021).*

It can be challenging to understand how DNAm influences the development of AD due to clinical and neuropathological heterogeneity and the comorbidities associated with diagnosing AD. Previous EWAS analyses have predominantly looked at a singular pathology measures such as Braak NFT stage (see section **1.3.3**). Ideally molecular studies of dementia brain should simultaneously consider multiple neuropathology measures in order to better understand the molecular mechanisms leading to disease.

4.1.6 Utilising the BDR dataset for neuropathology EWAS

There have been few EWAS studies of dementia utilising the new Illumina Infinium EPIC Beadarray (EPIC array; https://emea.illumina.com/), which quantifies DNAm at > 850,000 sites (Pidsley *at al.*, 2016)(see section **2.1.2**).The BDR DNAm dataset generated as part of my thesis represents the largest dementia EPIC datasets with an extensive range of neuropathological and phenotypic data to accompany it. In this chapter I quantified DNAm in 611 prefrontal frontal cortex (PFC) samples and 610 occipital cortex (OCC) samples from 631 individuals in the BDR cohort to identify if variable DNAm across the genome is associated with multiple neuropathology measures.

4.2 Chapter aims

The primary aim of this chapter is to integrate the detailed BDR neuropathological measures with Illumina EPIC array DNAm data to better understand the aetiology of AD and related dementias. The specific aims of this chapter are:

1. to identify if variable DNAm in the cortex is associated with five measures of neuropathology: Braak NFT Stage, Thal Phase, CERAD density, Braak LB stage and TDP-43
2. to identify common and unique changes in DNAm across two regions of the cortex
3. to identify similarities and differences in the patterns of DNAm associated with various types of neuropathology and explore the extent variation is driven by the progression of pathology

4.3 Methods

4.3.1 Brains for Dementia research cohort description

The Brains for Dementia research (BDR) cohort was established in 2008 and consists of a network of six dementia research centres across England and Wales (based at Bristol, Cardiff, King's College London, Manchester, Oxford and Newcastle Universities) and five brain banks (the Cardiff brain donations were banked in London). Participants >65 years of age were recruited using both national and local press (e.g. newspapers, newsletters, leaflets), TV and radio coverage as well as at memory clinics and support groups. There was no exclusion or inclusion criteria for individuals with specific diagnoses or those carrying genetic variants associated with neurodegenerative diseases; the cohort includes those with and without dementia and covers the full range of dementia diagnoses. Participants >65 years of age (including dementia cases and cognitively normal controls) underwent a series of longitudinal cognitive and psychometric assessments and registered for brain donation.

4.3.1.1 Longitudinal cognitive and clinical assessments

Longitudinal cognitive and clinical assessments were conducted by a trained psychologist or a research nurse. A series of exclusion criteria was put in place prior to assessments which included: 1) factors preventing brain donation such as brain injury and major stroke; 2) being < 65 years of age for healthy controls (apart from partners of participants with dementia); 3) not having the English language skills necessary for completing assessments; and 4) living to far geographically from an assessment centre or too remotely for a home visit. Baseline assessments were conducted face-to-face either in the participant's place of residence or a BDR centre. Follow-up assessments were predominantly face-to-face with the exception telephone interviews used for some of the healthy control participants. Follow-up interviews were annually conducted for participants with cognitive impairment, and every 1 to 5 years (depending on age) for cognitively healthy participants. Clinical assessment was performed using the Clinical Dementia Rating (CDR), which is measured on a 5 point scale (0,0.5,1,2,3) (Morris, 1997).

4.3.1.2 Post-mortem neuropathological assessment

Post-mortem brains underwent full neuropathological dissection, sampling and characterisation by experienced neuropathologists in each of the five network brain banks using a standardised BDR protocol which was based on the BrainNet Europe initiative (Alafuzoff *at al.*, 2008; Bell *at al.*, 2008). This protocol was used to generate a description of the regional pathology within the brain together with standardised scoring (see section **1.1.4**). In this chapter I considered five variables representing four neuropathological features:

1. Braak NFT stage which captures the progression of NFT pathology (Braak and Braak, 1991; Braak *at al.*, 2006)
2. Thal phase which captures the regional distribution of Aβ plaques (Thal *at al.*, 2002)
3. CERAD stage/ density which profiles neuritic plaque density (Mirra *at al.*, 1991; Montine *at al.*, 2012)
4. Braak LB stage which captures the progression of α-synuclein throughout the brain (Braak *at al.*, 2003)
5. TDP-43 status - a binary indicator of the TDP-43 inclusions, which was assessed using immunohistochemistry to identify the presence of phosphorylated TDP-43 in the amygdala, hippocampus and adjacent temporal cortex.

Braak NFT stage, Thal phase, CERAD density and Braak LB stage were analysed as continuous variables, utilising the semi-quantitative nature of these measures to identify dose-dependent relationships of increasing neuropathological burden. TDP-43 status was analysed as a binary variable.

4.3.2 DNA methylation data

After stringent data quality control (see section **3.2.2**) the BDR dataset consisted of DNAm estimates for 800,916 DNAm sites profiled in 1,221 samples (632 donors [53% male]; 610 PFC; 611 OCC; age range = 41-104 years, median = 84 years, mean = 83.49 years). Tissue samples from the PFC and OCC were selected since the PFC is one of the first areas affected in AD (the entorhinal cortex stage; Braak NFT Stages I-II) and the OCC is only affected at the latest stages (the neocortex stage; Braak NFT Stages V-VI)(see **Figure 1.1**).

4.3.3 *APOE* genotyping

In order to determine *APOE* status, samples were genotyped for *APOE* ε2, ε3 and ε4 alleles using an optimised TaqMan assay for SNPs rs7412 and rs429358 (Applied Biosystems). The genotype call rate was 99.7% (Brookes *at al.*, 2018). *APOE* status was modelled as two numeric variables counting the number ε2 and ε4 alleles each individual had. ε2 is rare and only 3 donors had an ε2/ε2 genotype, therefore the ε2/ε2 individuals were combined with the individuals with one ε2 allele. In addition to *APOE* genotype, samples were also profiled using a SNP array as described in **Chapter 5** section **5.3.2.2**, although these data were not used in the analysis in this chapter.

4.3.4 Investigating the effects of confounding variables on EWAS

DNAm measurements are potentially influenced by technical (e.g. plate on which the sample were run) and biological (e.g. cellular heterogeneity) confounding influences (van Iterson, van Zwet, BIOS Consortium, & Heijmans, 2017). It is essential to control for test-statistic inflation in EWAS as it can lead to false positives (van Iterson *at al.*, 2017). In GWAS, the genomic inflation factor (denoted lambda; λ) is used to quantify inflation by comparing test statistics across all SNPs compared to those under the null hypothesis (van Iterson *at al.*, 2017). In EWAS, test-statistics can be impacted by both inflation and bias (Leek, Johnson, Parker, Jaffe, & Storey, 2012). If there is bias in the test statistics this can lead to a shift in the effect size distribution resulting from confounding variables (Leek *at al.*, 2012). To identify if there was test-statistic inflation within the BDR DNAm dataset I ran several analyses with the aim of finding the optimal regression model in order to control for inflation without including excessive covariates and reducing power. The addition of too many covariates can lead to overfitting, where the model ends up measuring random noise as opposed to the relationship between the variables.

4.3.4.1 Calculating Principal Components

Principal component analysis (PCA) is a mathematical procedure that transforms a number of potentially correlated variables into a smaller number of uncorrelated variables called principal components (PCs). The first PC accounts for as much of the variability in the data as possible, and each succeeding PC accounts for as much of the remaining variability as possible. I used the *‘prcomp’* package in R to generate

PCs for the BDR DNAm data. I then calculated the correlation coefficient and the amount of variation explained in in PC1-PC10 by known covariates, which was calculated using regression models, where PC was the dependent variable and the selected covariates (age, sex, derived cell proportions, 96-well plate the samples were run, brain bank and PMI) were iteratively included as independent variables. I coded brain bank and plate as dummy variables (coded 0 or 1 to indicate if they are within that grouping variable).

Other methods have been suggested to remove the effects of unknown confounders and batch effects in DNAm data, including Surrogate Variable Analysis (SVA).The *sva* R package (Leek *at al.*, 2012) contains functions for removing batch effects and other unwanted variation in high-throughput experiments. Surrogate variables are covariates constructed directly from high-dimensional data (e.g. DNAm data / gene expression data) that can be used in subsequent analyses to adjust for unknown (or latent) sources of noise and are therefore very similar to PCs. I repeated the analyses described above incorporating SV's generated using *sva* rather than PCs generated using *prcomp* to identify which method is optimal to use with the BDR dataset. The difference between the two methods is that to calculate the SVs, the analysis takes into consideration the variable of interest. I created two model matrices using the *model.matrix* function: a null model and full model. The null model contained the known covariates. The full contained the known covariates and the variable of interest which I chose to be Braak NFT stage. The *sva f*unction was then used to calculate the SVs and I investigated how much variance was explained in PC1 by the SVs.

4.3.4.2 Measuring inflation and applying Bacon

I ran a preliminary EWAS using the PFC samples, associating variable DNAm against Braak NFT stage to identify if there was any evidence for systematic inflation in the results. I ran two mixed effects linear regression models: 1) a model including age, sex, experimental batch (i.e. the 96 well plate the samples were run on) and derived cell proportions as covariates, and 2) a model including the same covariates but additionally including PC1 (as derived as the optimal covariate to include to reduce inflation; see section **4.4.2**). I used the λ genomic inflation factor as a measure of inflation. If λ is > 1, this suggests there is inflation in the data. Of note, λ is not necessarily the best statistic to capture inflation in EWAS as λ overestimates inflation in the presence of a moderate proportion of true associations (van Iterson *at al.*, 2017). An additional measure to determine if there is inflation in the results is to identify bias in the test statistics. I assessed the t-statistic which is the calculated difference represented in units of standard error. I plotted the t-statistics and visually inspected them for evidence of bias (i.e. if the results were skewed in either direction this suggests there is bias).

4.3.5 Regression against neuropathology

To identify associations between variable DNAm and neuropathology, I fitted regression models using the optimal model as established in **section 4.4.2**. As data for each donor was derived from two cortical regions I used a mixed effects linear regression model (for more information see Chapter 2 section **2.3.2**), implemented with the *lme4 (Bates, Mächler, Bolker, & Walker, 2015)* and *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2017) packages.

To identify DNAm sites associated with dementia neuropathology, I conducted an EWAS in which DNAm at each probe was regressed against all five neuropathology measures (Braak NFT stage, CERAD density, Thal Phase, Braak LB stage and TDP43-status). To identify the p-value I used an ANOVA comparing the full model (see full model below) including the five neuropathology measures to a null model (see null model below) in which all five neuropathology measures were excluded. If the ANOVA provides evidence that the models are different (i.e. a significant p-value), this suggests that variable DNAm is driven by changes in neuropathology. In each model the following covariates were included as fixed effects: age, sex, experimental batch, PC1 and derived neural cell proportions. Cell proportions were derived using an algorithm developed by our lab group (Hannon *at al.*, unpublished) which classifies the cellular populations in the cortex into three proportions (neurons, oligodendrocytes, and other glial populations and the remaining cell types). Two of the three proportions (neuronal and glial populations / remaining cell types) were included in the model to eliminate the effects of multi-collinearity. Individual (donor ID) was included as a random effect to account for the fact that DNAm was quantified in multiple tissues from a single donor. Equations for the mixed effect regression models are shown below:

EWAS of all five neuropathology measures full model, where i = DNAm site:

$$DNA\ methylation[i]\sim Braak\ NFT\ stage + \ CERAD\ density + \ \ Thal\ Phase \ \ + Braak\ LB\ stage \\ + \ TDP43 + age + sex + batch + \ cell\ proportions + PC1 + (1|individual)$$

EWAS of all five neuropathology measures null model, where i = DNAm site:

$$DNA\ methylation[i]\sim age + sex + batch + \ cell\ proportions + PC1 \ + (1|individual)$$

I next conducted an EWAS for each of the five neuropathology measures separately (Braak NFT stage, CERAD density, Thal Phase, Braak LB stage and TDP43-status) and the global clinical dementia rating (CDR; a measure of cognitive decline) using mixed effect regression model where age, sex, experimental batch, PC1 and derived neural cell proportions were included as fixed effects and individual was included as a random effect. Cognitive decline has previously been shown to be correlated with neurodegeneration (Risacher *at al.*, 2017), hence I included CDR as an additional EWAS variable to further assess this hypothesis. To identify the p-value I used an ANOVA comparing the full model including the neuropathology measure or CDR to a null model in which the neuropathology measure or CDR was excluded. The numbers of samples included in each regression model are shown in **Table 4.1**. Equations for the mixed effect regression models are shown below:

EWAS of neuropathology full model, where i = DNAm site:

$$DNA\ methylation[i]\sim neuropathology\ measure + age + sex + batch + \ cell\ proportions + PC1 \\ + (1|individual)$$

EWAS of neuropathology null model, where i = DNAm site:

$$DNA\ methylation[i]\sim age + sex + batch + \ cell\ proportions + PC1 \ + (1|individual)$$

4.3.5.1 Identifying differential effects across cortical brain regions

In order to identify if there was an interaction with cortical region (i.e. different associations in the PFC compared to the OCC) for each neuropathology measure, I ran a mixed effects regression analysis including brain region as an interaction term in the regression model. To identify the p-value I used an ANOVA comparing the full model including the interaction term to a null model in which the interaction was excluded. The equation for the mixed effect regression models for the interaction term are shown below.

EWAS of neuropathology interaction with brain region, where i = DNAm site:

$$DNA\ methylation[i] \sim neuropathology\ measure + age + sex + batch + \ cell\ proportions + PC1 \\ + brain\ region + \ neuropathology\ measure * brain\ region \ + (1|individual)$$

EWAS of neuropathology interaction with brain region null model, where i = DNAm site:

$$DNA\ methylation[i] \sim neuropathology\ measure + age + sex + batch + \ cell\ proportions + PC1 \\ + brain\ region + (1|individual)$$

4.3.5.2 Identifying independent neuropathology associations

In order to confirm if the neuropathology measures were characterised by common or independent associations with DNAm I ran a mixed effects regression analysis including all neuropathology measures in the model. To identify the p-values I used ANOVAs comparing the full model including all neuropathology measures to null models where the target neuropathology measure was excluded in each case. Equation for the mixed effect regression model including all neuropathology measures is shown below:

EWAS of all neuropathology measures, where i = DNAm site:

$$DNA\ methylation[i] \sim Braak\ Tangle\ Stage + Thal\ Phase + cerad\ density + Lewy\ Body\ stage + \\ TDP - 43\ staus + age + sex + batch + \ cell\ proportions + PC1 \ + (1|individual)$$

EWAS of all neuropathology measures null model, where i = DNAm site:

$$DNA\ methylation[i] \sim four\ remaining\ neuropathology\ measures \ + age + sex + batch \\ + \ cell\ proportions + PC1 \ + (1|individual)$$

4.3.5.3 Significance threshold

I used an experiment wide significance threshold of p<9e-08 to identify differentially methylated positions associated with neuropathology. This threshold has been empirically derived by our group via the simulation of null DNAm datasets using Illumina EPIC array data (Mansell *at al.*, 2019).

Table 4.1: Samples and donors included in the EWAS for each phenotype tested.

Phenotype	Measurement	Brain Region		Total Samples	Donors
		Occipital	Prefrontal		
Braak NFT Stage	0	22	21	43	22
	1	61	59	120	64
	2	116	116	232	120
	3	73	77	150	78
	4	60	59	119	62
	5	101	101	202	103
	6	158	157	315	163
Total		**591**	**590**	**1181**	**612**
CERAD density	no	156	159	315	165
	sparse	80	80	160	82
	moderate	86	82	168	88
	high	228	230	458	237
Total		**550**	**551**	**1101**	**572**
Thal Phase	0	62	62	124	64
	1	69	69	138	73
	2	66	64	130	67
	3	81	82	163	84
	4	81	80	161	83
	5	176	181	357	186
Total		**535**	**538**	**1073**	**557**
Braak LB Stage	0	366	371	737	385
	1	5	5	10	5

	2	8	7	15	8
	3	22	20	42	22
	4	20	21	41	22
	5	27	27	54	28
	6	68	68	136	69
Total		**516**	**519**	**1035**	**539**
TDP 43 present	0	451	444	895	463
	1	122	125	247	127
Total		**573**	**569**	**1142**	**590**
Global CDR	0	131	132	263	137
	0.5	54	55	109	56
	1	75	78	153	80
	2	60	60	120	62
	3	139	136	275	141
Total		**459**	**461**	**920**	**476**
APOE-ε2	0	536	534	1070	553
	1&2	61	62	123	65
Total		**597**	**596**	**1193**	**618**
APOE-ε4	0	279	277	556	285
	1	275	278	553	290
	2	43	41	84	43
Total		**597**	**596**	**1193**	**618**

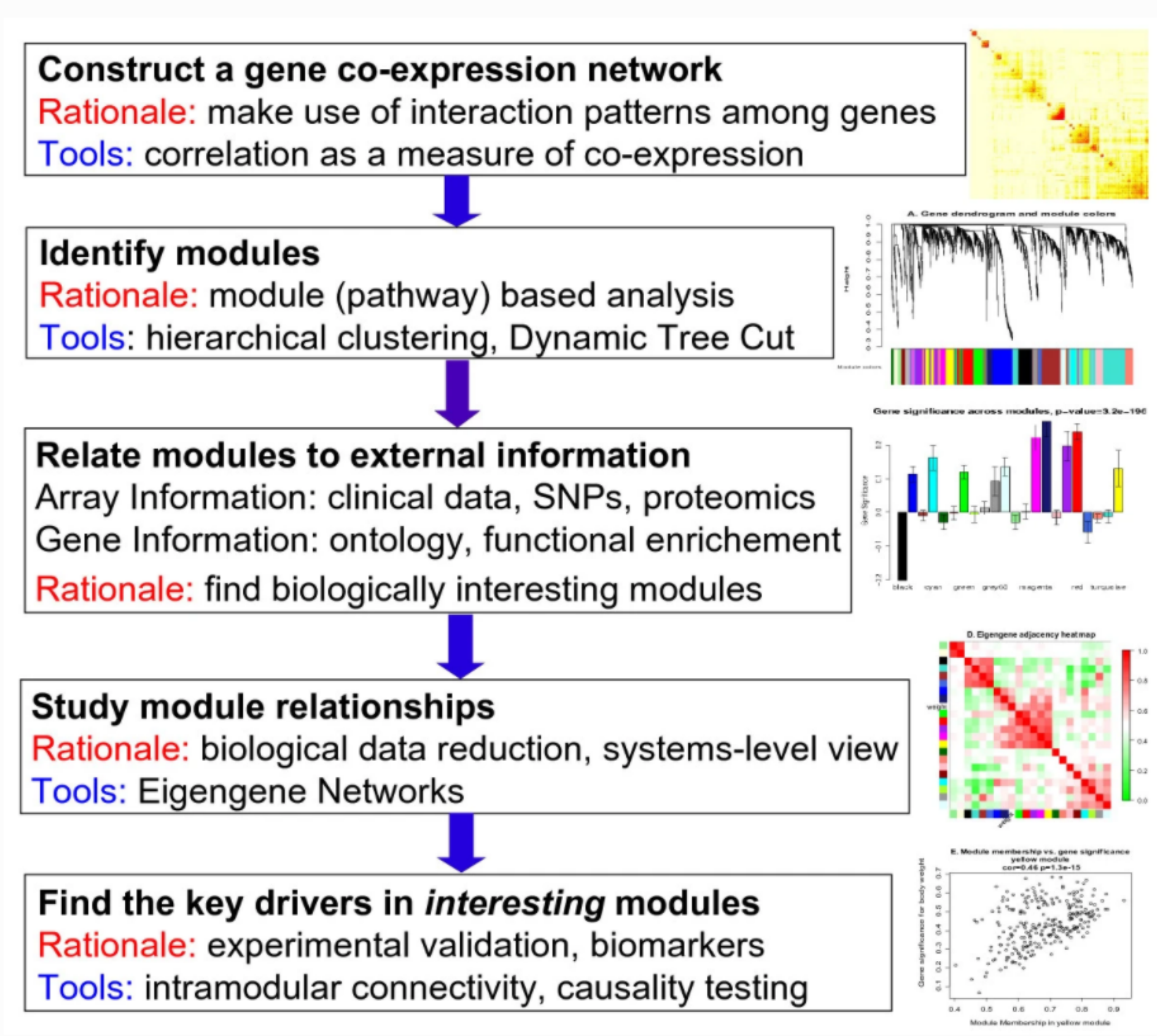
4.3.6 Identifying differentially methylated regions

In order to identify differentially methylated regions (DMRs) – i.e. genomic regions in which DNAm across multiple sites is consistently associated with a phenotype – I used *dmrff* (Suderman *at al.*, 2018). *Dmrff* identifies regions by combining summary statistics from proximally located DNAm sites. For more details on dmrff see Chapter 2 section **2.3.4.** I applied dmrff to the results from each of the individual EWAS analyses.

4.3.7 Generation of co-methylation networks

To identify modules (i.e. clusters) of highly correlated DNAm sites in the genome I used the R package weighted gene correlation network analysis (*WGCNA*) (Langfelder & Horvath, 2008). *WGCNA* can be used as a gene ranking mechanism and the underlying hypothesis of this methodology proposes that genes/loci which are highly co-regulated (i.e co-vary together) likely share underlying biological processes. The package was originally developed for gene expression data although it can be applied to DNAm data. An overview of the methodology of *WGCNA* applied to gene expression data is shown in **Figure 4.2.** The first step of the *WGCNA* process is to construct a network of DNAm sites which are co-regulated within the dataset. The network represents one system which is then sub-grouped into modules of co-regulated sites which have related functions. The modules can then be used in a range of different ways; for example they can be associated with phenotypes of interest (e.g. neuropathology, cell proportions) and pathway or gene ontology (GO) analyses can be applied to investigate the biological functions of each of the modules.

Preceding the generation of network modules in the BDR DNAm data, I first removed non-variable probes (defined as probes smaller than the median variance across all sites; median variance = 0.18%) from the normalised DNAm data. This left 400,458 (50%) probes for analysis. Samples were then clustered based on Euclidean distances and outliers were removed (n = 17 outliers). *WGCNA* uses a 'soft-thresholding' approach (i.e. weighted) for generating the biological networks (Langfelder & Horvath, 2008). The *pickSoftThreshold* function from the WGCNA package was used to analyse scale-free network topology (defined as following a power law distribution) and identify the optimal soft-thresholding power. I visually ascertained the optimal threshold by inspecting a plot showing the scale free topology, fitting index R^2 versus the different soft-thresholding powers (see **Figure 4.3**). Langfelder and Horvath (2008) suggest the optimal soft-thresholding power is where the curve flattens out after reaching a high R^2 value (>0.80). The connection strength between two DNAm sites was weighted using a soft-threshold value of 16 (see **Figure 4.3**). Network construction and module detection was performed in a block wise manner whereby I based the calculation on a maximum block size of 10,000 (the maximum number of probes included in each block when generating the networks). Of note, the modules were constructed irrespective of the direction of correlation between the probes. In order to generate networks I used the *blockwiseModules* function. The modules were arbitrarily assigned colours with the grey module being excluded as it contains unassigned probes. Module eigengenes (MEs; defined as the first principal component calculated from the DNAm values of all members of that module and represents the DNAm profile in a given module (Langfelder & Horvath, 2008)) were estimated for each module using the *moduleEigengenes* function. Each sample was assigned an ME value and this represents the shared DNAm profile of the module.



Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression

Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut

Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules

Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks

Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing

A. Gene dendrogram and module colors
Height
Module colors
Gene significance across modules, p−value=3.2e−196
Gene Significance
black
cyan
green
grey60
magenta
red
turquoise
D. Eigengene adjacency heatmap
weight
1.0
0.8
0.6
0.4
0.2
0.0
E. Module membership vs. gene significance
yellow module
cor=0.46 p=1.3e−15
Gene significance for body weight
Module Membership in yellow module
0.4
0.5
0.6
0.7
0.8
0.9

Figure 4.2: Overview of WGCNA methodology. *This flowchart presents an overview of the main steps of Weighted Gene Co-expression Network Analysis. Figure and legend reprinted from (Langfelder & Horvath, 2008).*

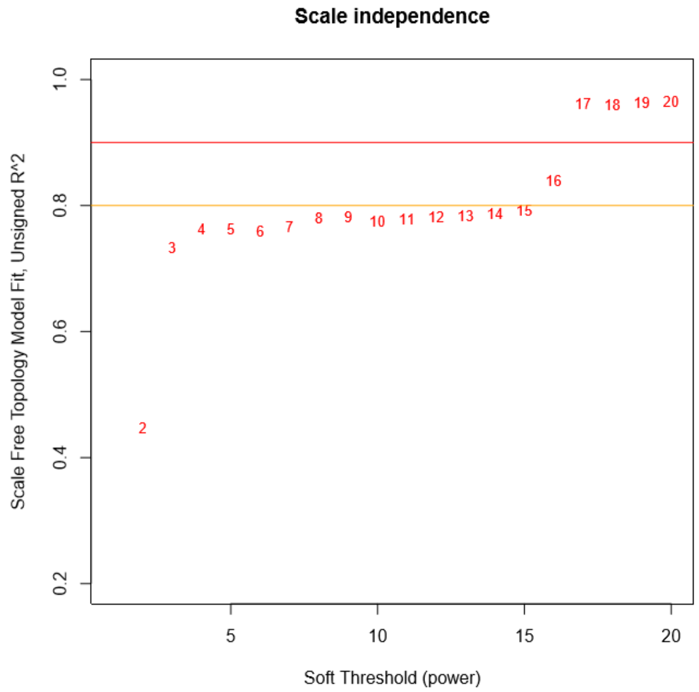


Figure 4.3: Soft power threshold for WGCNA analysis. *Analysis of network topology for various soft-thresholding powers. Figure shows the scale-free fit index R^2 (y-axis) as a function of the soft-thresholding power (x-axis) for the Brains for Dementia Research DNA methylation network.*

4.3.7.1 Associating modules to neuropathology and other traits of interest

All identified modules were correlated with each of the five neuropathology measures (using Pearson's correlation coefficient) as well as other traits including derived neural cell proportions, age, sex and *APOE* genotype. Covariates were not regressed out since differential methylation associated with them may provide biologically relevant insights. In order to identify if associations with neuropathology were driven by covariates, I subsequently ran regression models for the modules of interest (i.e. the modules which were significantly correlated with neuropathology) including the covariates in the data. Modules which remained significant were carried through to subsequent analyses.

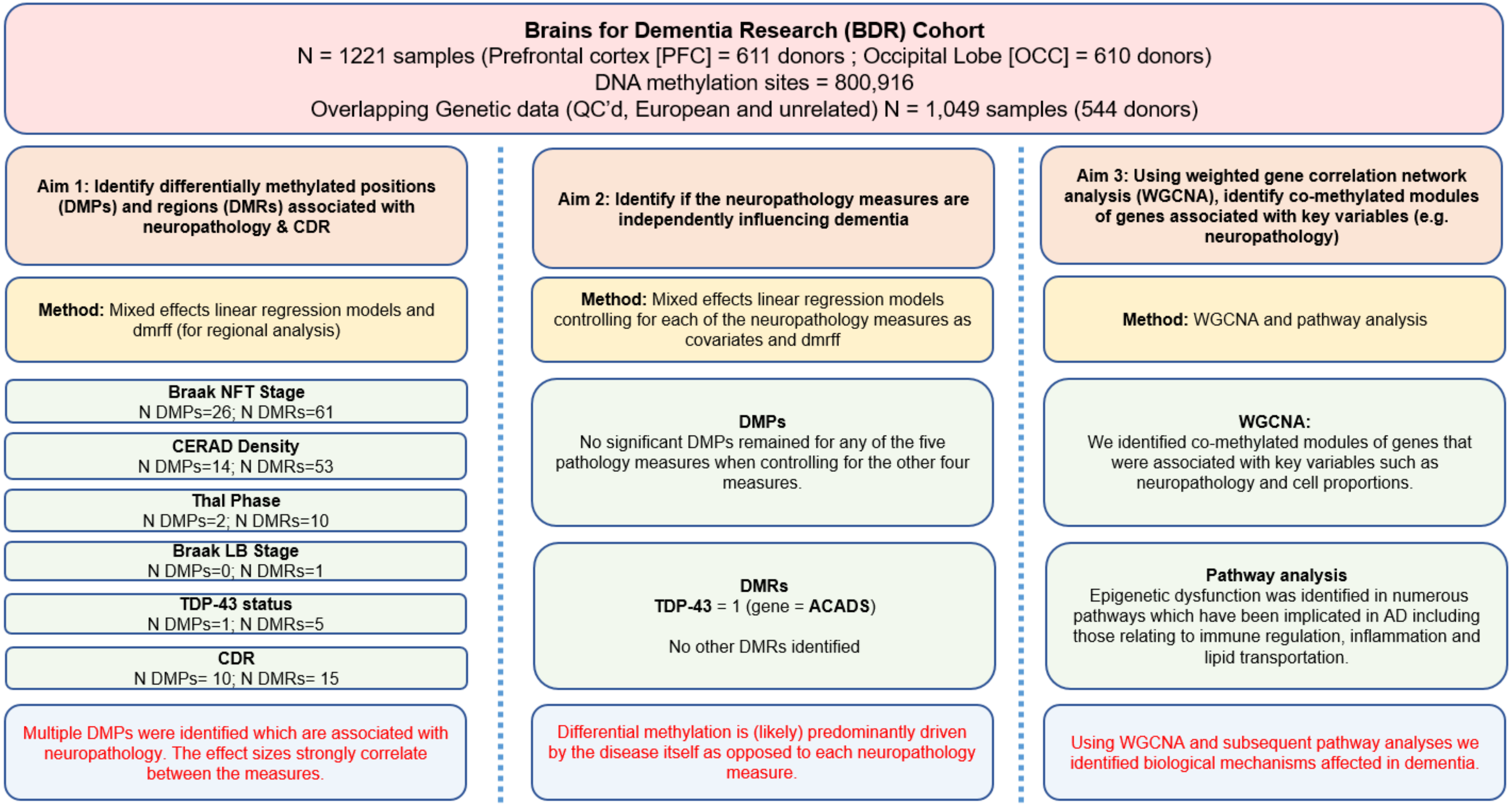
4.3.8 Module significance and Pathway analysis

For the modules showing significant associations across the 28 identified co-methylation modules (Bonferroni $p < 0.05/28 = 0.002$) with neuropathology after controlling for covariates, I calculated the module membership (MM) and probe significance. MM was calculated as the Pearson correlation coefficient between the

DNAm value of each probe and the ME values, measuring the association between a probe and the module which it was assigned to. The probe significance was then calculated to identify the strength of the correlation between the DNAm methylation value of each probe and the trait of interest: Pearson's correlation was used for continuous traits and Spearman's correlation was used for binary traits.

Gene ontology (GO) pathway analysis was applied to the list of genes annotated to the probes in each co-methylation module associated with at least one measure of neuropathology to identify enrichments for specific biological processes. The Illumina UCSC gene annotation was used to create the test gene lists for pathway analysis. Where probes were not annotated to any gene (e.g. if they are in intergenic locations), they were excluded from this analysis. Where probes were annotated to >1 gene, each gene was included in GO analysis. Analyses were performed using the *methylglm* function within the methylGSA package developed by Ren and Kuan (2019) using the default parameters. *methylglm* adjusts for the number of DNAm sites in the logistic regression model (for more details see **Chapter 2** section **2.3.4**).

4.4 Results



Brains for Dementia Research (BDR) Cohort
N = 1221 samples (Prefrontal cortex [PFC] = 611 donors ; Occipital Lobe [OCC] = 610 donors)
DNA methylation sites = 800,916
Overlapping Genetic data (QC'd, European and unrelated) N = 1,049 samples (544 donors)

Aim 1: Identify differentially methylated positions (DMPs) and regions (DMRs) associated with neuropathology & CDR

Method: Mixed effects linear regression models and dmrff (for regional analysis)

Braak NFT Stage
N DMPs=26; N DMRs=61

CERAD Density
N DMPs=14; N DMRs=53

Thal Phase
N DMPs=2; N DMRs=10

Braak LB Stage
N DMPs=0; N DMRs=1

TDP-43 status
N DMPs=1; N DMRs=5

CDR
N DMPs= 10; N DMRs= 15

Multiple DMPs were identified which are associated with neuropathology. The effect sizes strongly correlate between the measures.

Aim 2: Identify if the neuropathology measures are independently influencing dementia

Method: Mixed effects linear regression models controlling for each of the neuropathology measures as covariates and dmrff

DMPs
No significant DMPs remained for any of the five pathology measures when controlling for the other four measures.

DMRs
TDP-43 = 1 (gene = **ACADS**)

No other DMRs identified

Differential methylation is (likely) predominantly driven by the disease itself as opposed to each neuropathology measure.

Aim 3: Using weighted gene correlation network analysis (WGCNA), identify co-methylated modules of genes associated with key variables (e.g. neuropathology)

Method: WGCNA and pathway analysis

WGCNA:
We identified co-methylated modules of genes that were associated with key variables such as neuropathology and cell proportions.

Pathway analysis
Epigenetic dysfunction was identified in numerous pathways which have been implicated in AD including those relating to immune regulation, inflammation and lipid transportation.

Using WGCNA and subsequent pathway analyses we identified biological mechanisms affected in dementia.

Figure 4.4: Summary of results for Chapter 4: Epigenome wide association study of Neuropathology in the Brains for Dementia research Cohort.

4.4.1 Cohort characteristics

Donors had a mean age at death of 83.49 years (SD = 9.1) and 53% were male. Males were significantly younger at death by 2.69 years (p=2.33e-07) in comparison to females, which is consistent with epidemiological studies (Oksuzyan, Juel, Vaupel, & Christensen, 2008; Owens, 2002). In 1181 samples (612 donors) NFT pathology was quantified using Braak NFT stage (Braak and Braak, 1991; Braak *at al.*, 2006) with a mean Braak score of 3.74 (SD = 1.90; see **Figure 4.5**). Aβ was quantified using two variables: Thal phase (Thal *at al.*, 2002) with a mean value of 3.09 (SD = 1.78) across 1073 samples (557 donors) and neuritic plaque density scored using the CERAD classification (Mirra *at al.*, 1991; Montine *at al.*, 2012) with a mean value of 1.69 (SD = 1.26) across 1101 samples (539 donors; see **Figure 4.5**). α-synuclein pathology was quantified using Braak LB stage, and across 1035 samples (590 donors) the mean score was 1.36 (SD = 2.26; see **Figure 4.5**). TDP-43 status was available for 1142 samples (590 donors), with 247 samples (127 donors; 22%) being classed as being TDP-43 positive (see **Figure 4.5**). In 920 samples (476 donors) CDR was measured, with an average score of 1.38 (SD = 1.22) out of a possible 3 (0 = normal; 0.5 = very mild Dementia; 1 = mild dementia; 2 = moderate dementia; 3 = severe dementia; see **Figure 4.5**).

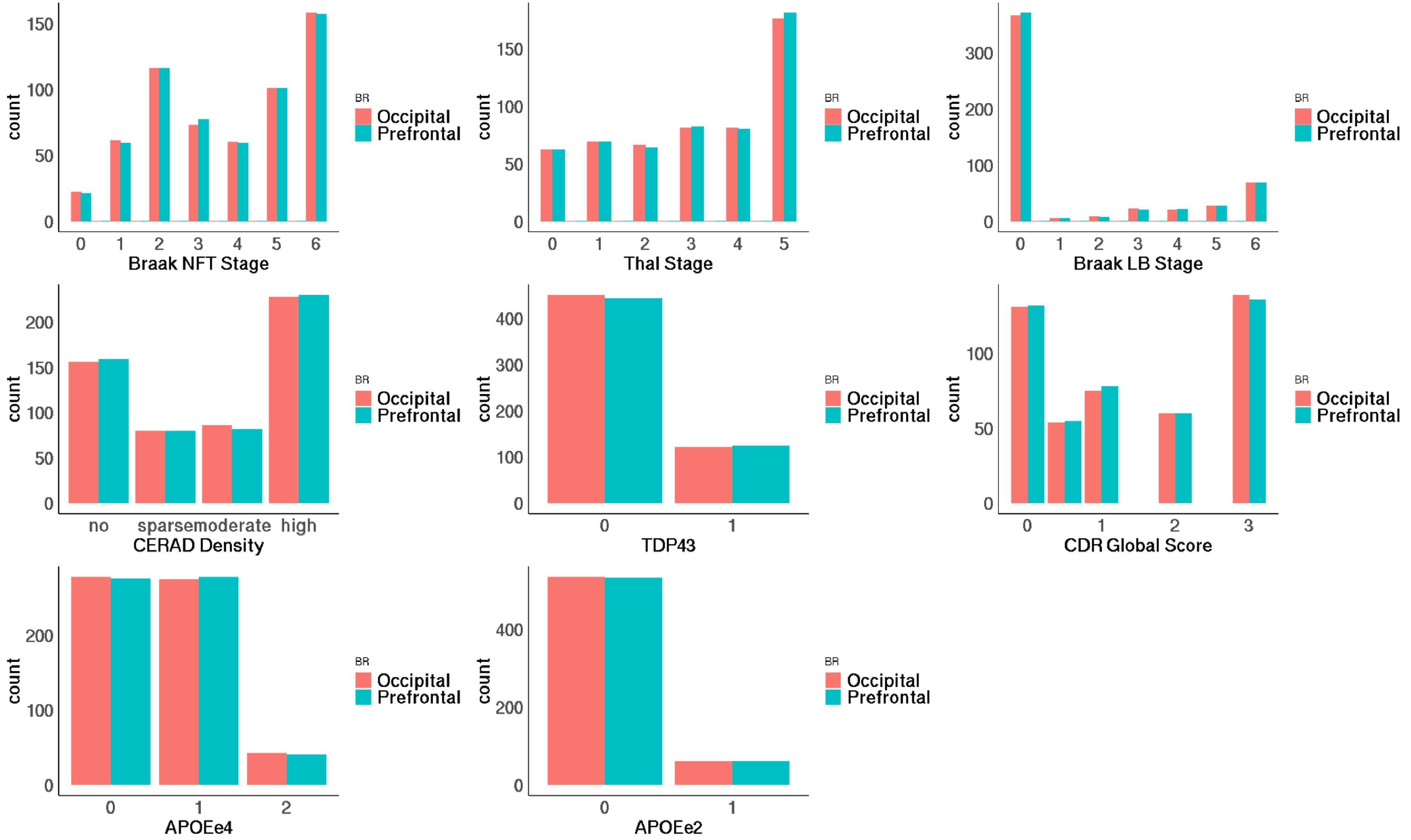


Figure 4.5: Distribution of neuropathology, CDR and APOE genotype (ε3 and ε4) in the BDR DNA methylation cohort split by brain region. *CDR = Clinical dementia rating. BR = brain region. Braak NFT Stage = Braak neurofibrillary tangle stage. Braak LB Stage = Braak Lewy body Stage.*

4.4.2 Establishing the best model to use for EWAS

I ran a series of analyses to establish the optimal EWAS model to use on the BDR data. I first generated PC's using *prcomp* finding that PC1 explained the most variation in the data (96.8%), with subsequent PC's explaining considerably less of the variation (PC2=0.76%, PC3=0.34%, PC4=0.23%, PC5=0.15%). I then associated PC1 against known traits to identify how much of the variation is explained by known covariates. From the correlation matrix and heatmap against all traits (see **Figure 4.6**) it can be seen that brain bank (r=-0.5, p=7.6e-79) and experimental plate (r=-0.4, p=2.4e-39) are both highly correlated with PC1, highlighting the importance of experimental batch effects. To further explore this, I plotted heatmaps of brain bank and plate (coded as dummy variables) against the first 10PCs, to help identify the main source of variation (see **Figure 4.6 B and C)**. The samples from Manchester had the largest influence in PC1 (r=0.4, p=1.1e-47), followed by Bristol (r=0.27, p=3.9e-22). More specifically, plate 8 has the highest correlation with PC1 (r=0.42, p=3.43-54), followed by plate 12 (r=-0.29, p=1.7e-24).

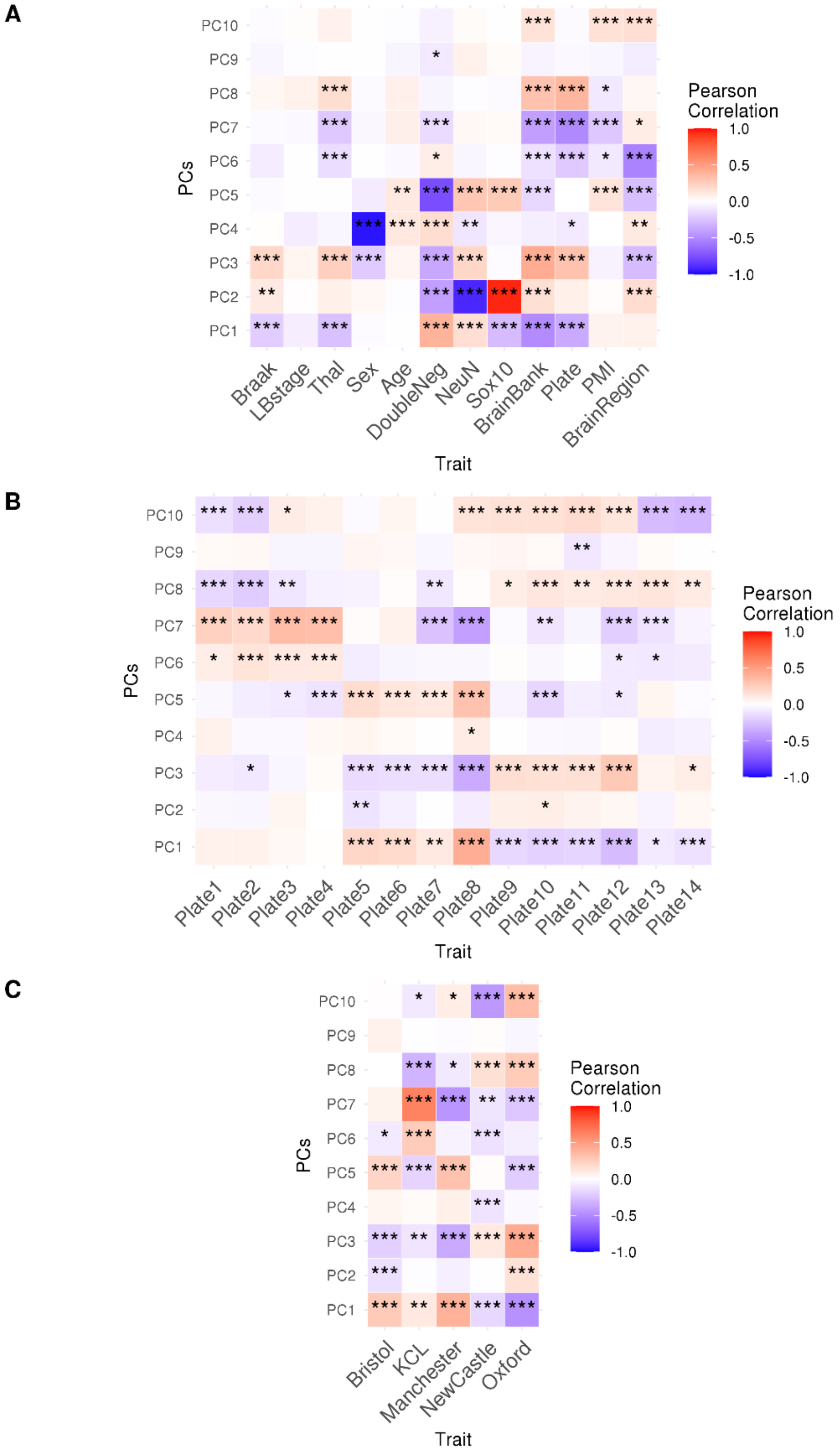


***Figure 4.6: Heatmaps of the correlation between principal components (PC) 1-10 and known covariates.** Shown are correlations **(A)** against all known covariates; **(B)** against plate – coded as a dummy variable and **(C)** against brain bank – coded as a dummy variable. NFT = neurofibrillary tangle stage; LB = Lewy body; PMI = post mortem interval; Thal = Thal phase; plate = 96 well array on which the samples were run. Cell proportions were estimated using a deconvolution algorithm incorporating cell-type-specific DNAm profiles generated using fluorescence activated nuclei sorting (FANS); cortical nuclei were stained with markers for neurons (NeuN+) and oligodendrocytes (Sox10+) and the remaining cells (double neg) before purification and DNAm profiling. Stars represent significance, where p-value thresholds were Bonferroni corrected for multiple testing (22 independent tests) *=p<0.002, **=p<0.0005 and ***=p<4.52e-05.*

Cell proportions were also significantly correlated with PC1 (NeuN+: r=0.17, p= 2.1e-09; SOX10+: r= -0.292, p= 2.2e-25; and double negative r= 0.393, p= 2.9e-46). DNAm differs dramatically between cell types (Mendizabal & Yi, 2016) and so cellular heterogeneity can significantly influence the quantification of DNAm in bulk tissue. In addition, neuropathology is associated with loss of specific cell types (e.g. neurons)(Gómez-Isla *at al.*, 1997). Therefore, estimations of cell-type composition is an important variable to consider when analysing bulk tissue in studies of neuropathology.

In order to visualise the effect specific confounders were having on the PCs I plotted them, colouring samples by the trait to better visualise any potential confounding (see **Figure 4.7**). Brain bank was most correlated with PC1 (r=-0.5, p=7.6e-79; see **Figure 4.6** and **Figure 4.7**) but also strongly correlated with PC3 (r=-0.43, p=1.6e-56; see **Figure 4.6**). Derived cell proportions were most correlated with PC2 (NeuN+: r=-0.9128, p=2.2e-308; SOX10+: r= 0.9564, p= p=2.2e-308; and double negative: r-0.4188, p= 5.0e-53; see **Figure 4.6** and **Figure 4.7**) and explained nearly all of the variation in this PC (see **Table 4.2**). Sex was most correlated with PC4 and explained nearly all of the variation in this PC (see **Table 4.2**).

Because PC1 explained the most variation in the data (96.8%) I was most interested in identifying the variation explained by known covariates in this PC. Plate explained nearly half of the variation in PC1 (44.94%, see **Table 4.2**) with cell proportions increasing this to ~60% (see **Table 4.3**). Further covariates had little influence on PC1 and the remaining variation is not explained by known covariates. The addition of Braak NFT stage in the model against PC1 explained little extra variation in the data (see **Table 4.6**). This suggests PC1 could be a useful covariate to include in neuropathology analysis to account for unknown variation, with other known variables being added as specific confounders.

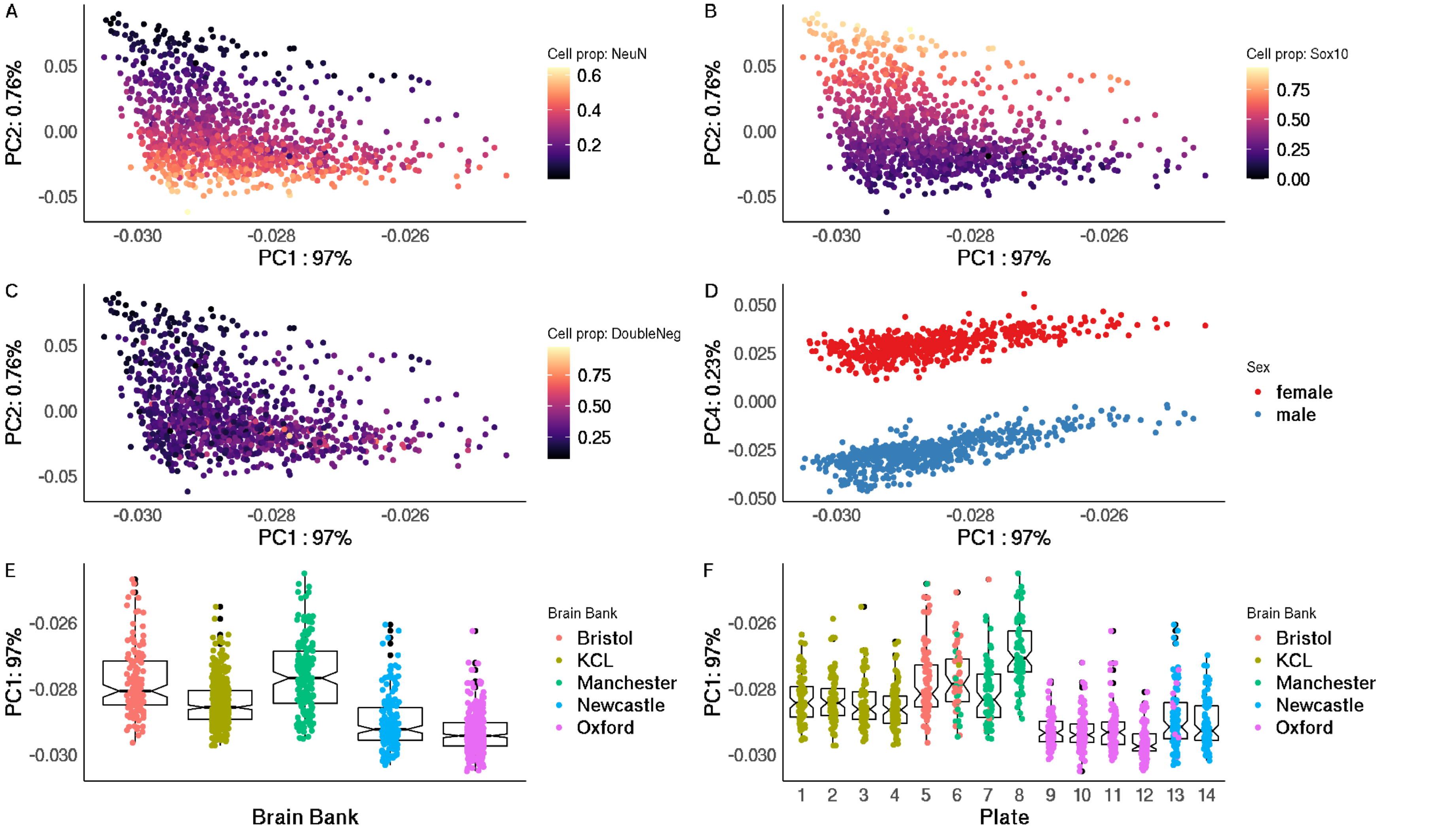


Figure 4.7: Plots of principal components (PCs) against significantly correlated traits. PC = Principal component. *The % after the PC is the amount of variance explained in the Brains for Dementia research DNA methylation datatset by that specific PC. Plate = 96 array on which the samples were run. Cell proportions were calculated based on an algorithm which used fluorescence activated nuclei sorting (FANS) data where cortical nuclei were stained with markers for neurons (NeuN+) and oligodendrocytes (Sox10+) and the remaining cells (double neg).*

Table 4.2: Variance explained by each individual known covariate in the first five principal components (PCs). *R^2 – a measure of the variance explained by the traits and shows how well terms fit the regression model.*

Covariates included in regression model	R^2 (%) PC1	R^2 (%) PC2	R^2 (%) PC3	R^2 (%) PC4	R^2 (%) PC5
Plate	45.55	5.07	34.93	3.89	23.82
Cell proportions	17.44	98.29	19.22	4.13	67.48
Sex	0.06	0.22	4.94	93.52	0.71
Age	0	0.01	0.2	1.72	1.14
BR	0.43	3.31	8.16	1.16	7.59
PMI	0.23	-0.07	0.18	-0.08	1.86

Table 4.3: Additive variance explained by known covariates in the first five principal components (PCs). *AdjR^2 = adjusted R^2 – a measure of the variance explained which shows how well terms fit the regression model where the statistic is adjusted based on the number of independent variables in the model*

Covariates included in regression model	AdjR^2 (%) PC1	AdjR^2 (%) PC2	AdjR^2 (%) PC3	AdjR^2 (%) PC4	AdjR^2 (%) PC5
Plate	44.94	4.01	34.2	2.82	22.97
Plate+NeuN	45.47	84.42	43.32	4.16	27.06
Plate+NeuN+DoubleNeg	59.13	98.4	57.37	7.22	84.33
Plate+NeuN+DoubleNeg+Age	59.23	98.43	57.34	9.52	84.41
Plate+NeuN+DoubleNeg+Age+Sex	59.21	98.47	63.36	98.21	85.33
Plate+NeuN+DoubleNeg+Age+Sex+BR	59.29	98.46	65.71	98.37	85.77
Plate+NeuN+DoubleNeg+Age+Sex+BR+PMI	59.38	98.48	65.78	98.37	85.76
Plate+NeuN+DoubleNeg+Age+Sex+BR+PMI+Braak	60.03	98.48	66.48	98.37	85.78
Unknown	39.97	1.52	33.52	1.63	14.22

I then repeated the analyses above incorporating SV's generated using *sva* to identify which method is optimal to use with the BDR dataset. I investigated how much variance was explained in PC1 by the SVs (see **Table 4.4** and **Table 4.5**). The majority of the variation in PC1 can be explained by SV2 (R^2 = 74.56%; see **Table 4.4**). However nearly all of the variation in PC1 was explained by SV1 and SV2 combined (R^2 = 95.52%; see **Table 4.5**).

To work out which traits explained the variation in each SV I calculated and plotted the correlations (see **Figure 4.8**) and the amount of variation explained in SV1-SV3 by known covariates (see **Table 4.4** and **Table 4.5**) derived using regression models, where SV was the dependent variable and the covariates were iteratively included as independent variables. SV1 was predominantly explained by experimental batch (plate) and derived cell proportions, and in combination these variables explained 97.16% of the variation (see **Table 4.6**). SV2 was also predominantly explained by plate and derived cell proportions (R^2 = 67.29%), however around 30% of the variation in SV2 was not explained by known covariates.

My comparisons show that more SVs than PCs would need to be included in a regression model to explain the unknown variance in the data. Therefore, in order to ensure the model was parsimonious (a model which explains data with a minimum number of parameters) I included PC1 in my regression models. Of note, only a very small amount of the variance in PC1 is explained by the target phenotypes (neuropathology) of my EWAS analysis (neuropathology, LOAD PRS and *APOE* genotype) and it captures more of the unknown variance than SVs without hampering the degrees of freedom.

Table 4.4: Variance explained by surrogate variables (SVs) in principal component (PC) 1. *R^2 – a measure of the variance explained by the SVs and shows how well terms fit the regression model.*

Surrogate Variables	**R^2 (%) PC1**
SV1	20.88
SV2	74.56
SV3	0.95
SV4	0.71
SV5	1.10
SV6	-0.08
SV7	-0.07
SV8	-0.03
SV9	-0.08
SV10	-0.08

Table 4.5: Additive variance explained by surrogate variables (SVs) in principal component (PC) 1. *AdjR^2 = adjusted R^2 – a measure of the variance explained which shows how well terms fit the regression model where the statistic is adjusted based on the number of independent variables in the model.*

Surrogate Variables	AdjR^2 (%) PC1
SV1	20.88
SV1+SV2	95.52
SV1+SV2+SV3	96.56
SV1+SV2+SV3+SV4	97.36
SV1+SV2+SV3+SV4+SV5	98.54
SV1+SV2+SV3+SV4+SV5+SV6	98.54
SV1+SV2+SV3+SV4+SV5+SV6+SV7	98.56
SV1+SV2+SV3+SV4+SV5+SV6+SV7+SV8	98.62
SV1+SV2+SV3+SV4+SV5+SV6+SV7+SV8+SV9	98.62
SV1+SV2+SV3+SV4+SV5+SV6+SV7+SV8+SV10	98.62

Table 4.6: Additive variance explained by known covariates in the first three surrogate variables (SVs). *AdjR^2 = adjusted R^2 – a measure of the variance explained which shows how well terms fit the regression model where the statistic is adjusted based on the number of independent variables in the model.*

Covariates included in regression model	AdjR^2 (%) SV1	AdjR^2 (%) SV2	AdjR^2 (%) SV3
Plate	7.29	36.86	1.99
Plate+DoubleNeg	27.19	43.09	2.61
Plate+DoubleNeg+Sox10	96.23	62.01	2.85
Plate+DoubleNeg+Sox10+NeuN	97.16	67.29	11.12
Plate+DoubleNeg+Sox10+NeuN+Age	97.19	67.69	12.36
Plate+DoubleNeg+Sox10+NeuN+Age+Sex	97.21	67.87	99.31
Plate+DoubleNeg+Sox10+NeuN+Age+Sex+BR	97.22	68.89	99.32
Plate+DoubleNeg+Sox10+NeuN+Age+Sex+PMI	97.24	69.19	99.32
Plate+DoubleNeg+Sox10+NeuN+Age+Sex+BR+PMI+Braak	97.26	69.5	99.32

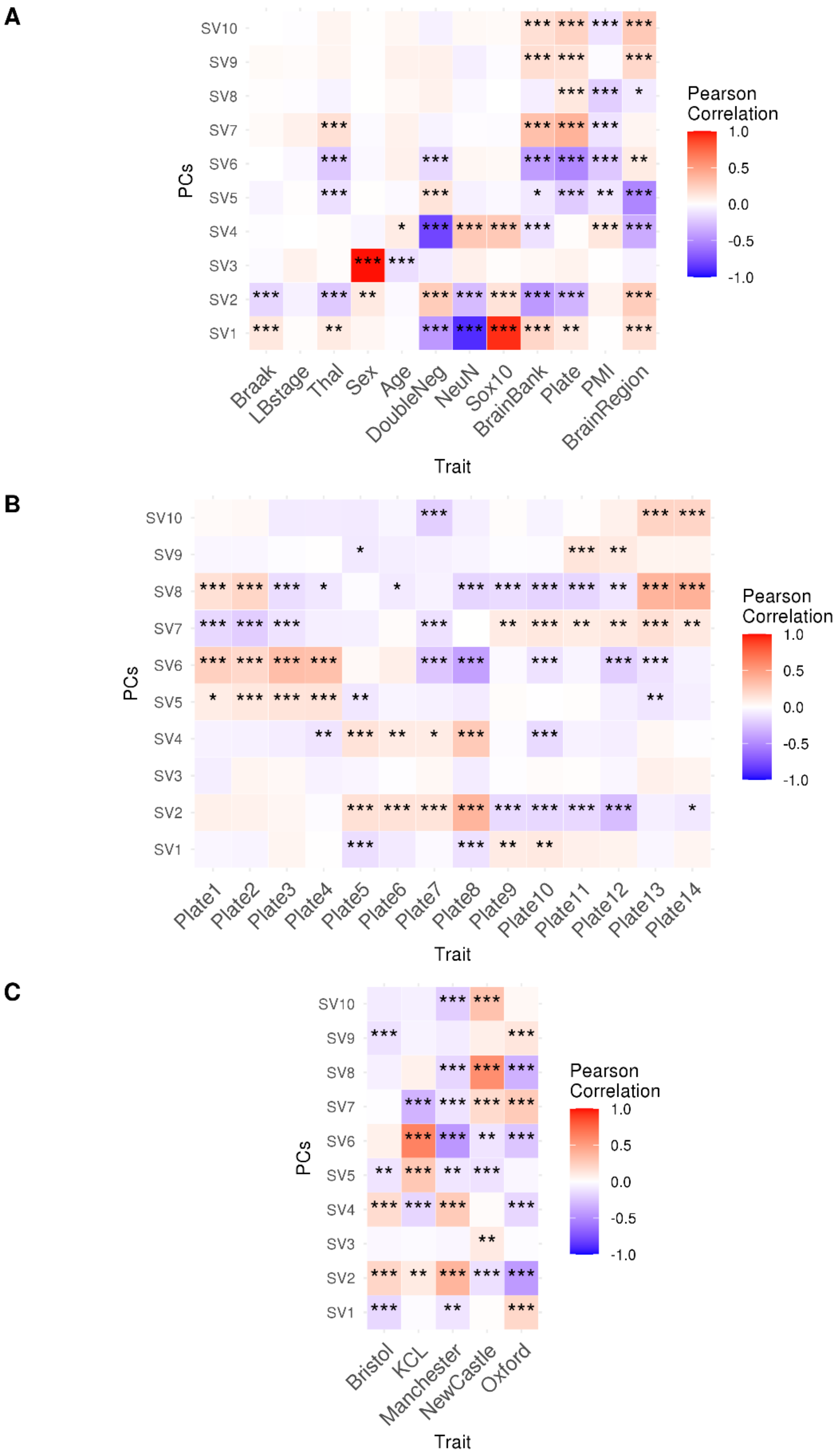


Figure 4.8: Heatmaps of the correlation between surrogate variables (SV) 1-10 and known covariates. *Shown are correlations **(A)** against all known covariates; **(B)** against plate – coded as a dummy variable and **(C)** against brain bank – coded as a dummy variable. NFT = neurofibrillary tangle stage; LB = Lewy body; PMI = post mortem interval; Thal = Thal phase; plate = 96 well array on which the samples were run. Cell proportions were estimated using a deconvolution algorithm incorporating cell-type-specific DNAm profiles generated using fluorescence activated nuclei sorting (FANS); cortical nuclei were stained with markers for neurons (NeuN+) and oligodendrocytes (Sox10+) and the remaining cells (double neg) before purification and DNAm profiling. Stars represent significance, where p-value thresholds were Bonferroni corrected for multiple testing (22 independent tests) *=p<0.002, **=p<0.0005 and ***=p<4.52e-05.*

4.4.2.1 Bacon reduces inflation in the data

I ran a preliminary EWAS using the PFC samples, associating variable DNAm against Braak NFT stage to identify if there was any evidence of inflation in the results. I ran two mixed effects linear regression models: 1) including age, sex, batch and derived cell proportions as covariates and 2) including the same covariates as (1) but additionally including PC1. The inclusion of PC1 reduced inflation (λ=1.34; see **Figure 4.9A**) in comparison to the model where it was not included (λ=2.68; see **Figure 4.9B**).

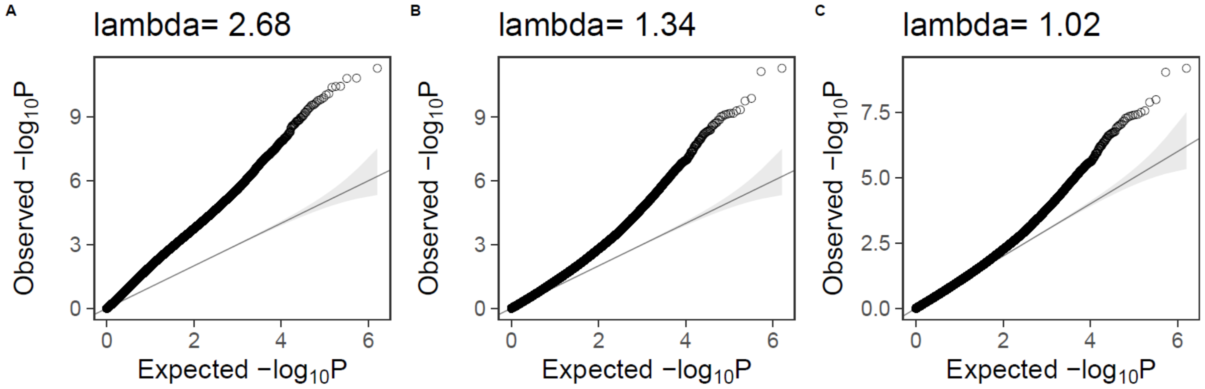


***Figure 4.9: Quantile-quantile plots of Braak Stage EWAS in the prefrontal cortex.** Shown are the expected (x-axis) against the observed (y-axis) quantiles in three EWAS against braak stage where in **(A)** Principal component (PC) 1 was not included as a covariate; in **(B)** PC1 was included as a covariate; and **(C)** PC1 was included as a covariate and bacon was applied to the results. Lambda = the genomic inflation factor which is a measure of inflation - a lambda > 1 suggests there is inflation in the data.*

When PC1 was not included in the model there was a slight negative skew in the t-statistics in comparison to when PC1 was included in the model where the t-statistics were normally distributed (see **Figure 4.10**). This further supports the inclusion of PC1 in the regression models as a measure to reduce bias.

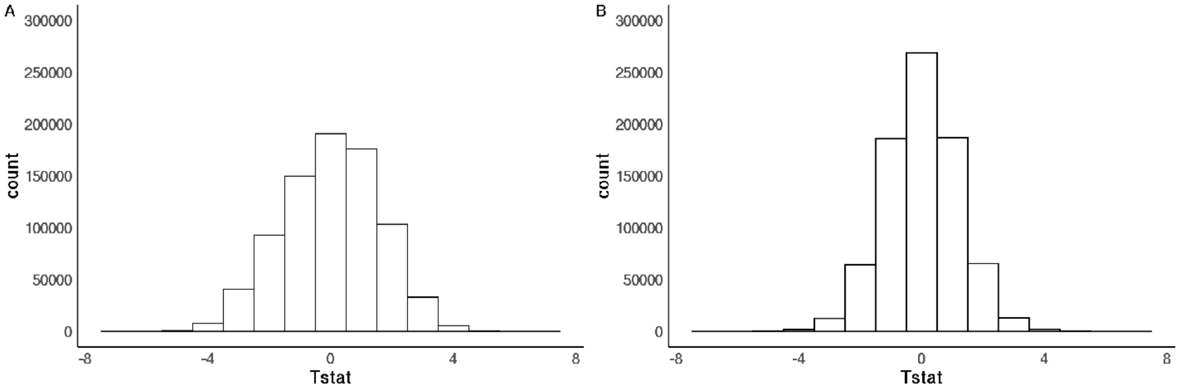


Figure 4.10: T-statistic distribution for two linear regression models against Braak stage, either including or excluding PC1. *Shown is the distribution when **(A)** PC1 was not included in the model and **(B)** PC1 was included in the model.*

Although the inclusion of PC1 reduced inflation, there was still some statistical inflation (see **Figure 4.9B**). Subsequently, I applied the *bacon* function to further reduce inflation (van Iterson *at al.*, 2017). After applying *bacon*, the inflation was greatly attenuated (λ=1.02; see **Figure 4.9C**) and there was no evidence of bias in the t-statistics (see **Figure 4.10**). Based on these results, if there was evidence of inflation in my EWAS analyses (λ > 1.2) I applied *bacon* correction.

4.4.3 Pathology-associated DNA methylation signatures across two cortical brain regions

4.4.3.1 Overview of analysis

Detailed neuropathological data in the BDR cohort was used to investigate the accumulation of Aβ plaques (measured by CERAD density and Thal Phase), tauopathy (measured by Braak NFT stage), synucleinopathy (measured by Braak LB stage), and TDP-43 proteinopathy (a binary variable of TDP-43 presence) influence epigenetic regulation. Briefly, to determine which DNAm sites were associated with dementia neuropathology I conducted an EWAS including all five measures (Braak NFT stage, CERAD density, Thal Phase, Braak LB stage and TDP43-status). I then used an ANOVA comparing the full model including the five neuropathology measures to a null model in which all five neuropathology measures were excluded. Since the model was looking at the effect of all neuropathology measures collectively there is no effect size estimate for this analysis. In each model the following covariates were included: age, sex, experimental batch (the specific 96-well plate on which the samples were processed), principal component 1 (PC1) and derived cell proportions.

4.4.3.2 Overview of results

My analysis identified 34 differentially methylation positions (DMPs) associated with dementia neuropathology at an experiment wide significance threshold (p<9e-08) (see **Figure 4.11** and **Table 4.7**). 16 (47%) of the DMPs were annotated to the EPIC array but are not present on the 450K, which demonstrates the utility of the newer platform and the advantage of the increased power in comparison to the 450K array. 24 of the DMPs were annotated to genes (using the UCSC gene annotation), with 18 of these genes being previously implicated in dementia and 6 being annotated to genes which have not previously been implicated in dementia.

4.4.3.3 Several neuropathology-associated DMPs are of relevance in the context of AD and associated neurobiological functions

Several of the DMPs identified in the neuropathology EWAS are relevant in the context of AD and associated neurobiological functions. These include:

- cg10208942 (p=1.56e-10) which is located on chromosome 17 and annotated to gene *SLC16A3* which has been identified as an AD risk gene in previous EWAS studies (De Jager *at al.*, 2014; Q. S. Li *at al.*, 2020).
- cg07061298 (p=2.37e-10) and cg22962123 (p=6.15e-09) which are located on chromosome 7 and are annotated to the *HOXA*3 gene region which has consistently been identified as an AD risk gene in EWAS, with studies reporting hypermethylation of this region (R. G. Smith *at al.*, 2018; R. Smith *at al.*, 2021).
- cg18032191 (p=1.07e-09) located on chromosome 12 and annotated to the gene *TNFRSF1A* which has been shown to be significantly up-regulated in AD in the several regions of the cortex including the hippocampus (Wang & Wang, 2020). Shang and colleagues (Shang *at al.*, 2015) found associations between *TNFRSF1A* and AD susceptibility in Caribbean Hispanic individuals through a genome-wide haplotype association study. Research by Steeland and colleagues supports a role for *TNFRSF1A* in AD pathogenesis by mediating neuronal cell death (Steeland *at al.*, 2018). The protein encoded by *TNFRSF1A* is one of the major receptors for tumour necrosis factor –alpha (TNF-alpha)

which is an important signalling protein in the immune system (Greco *at al.*, 2015).

- cg06913337 (p=2.64e-09) and cg09502865 (p=3.25e-08) located on chromosome 16 and annotated to the *ZFPM1* locus, which has previously been linked to psychosis in AD (Zheng *at al.*, 2015) and has shown to have suggestive significance with dementia with Lewy bodies (DLB) (Rongve *at al.*, 2019).
- cg20864214 (p=3.09e-08) and cg02776498 (p=3.35e-08) located on chromosome 11 and annotated to *ARHGEF17* were both previously associated with Braak NFT stage in the recent AD EWAS meta-analysis (R. Smith *at al.*, 2021).

4.4.3.4 Several neuropathology-associated DMPs have not been implicated in neurodegenerative disease

A number of the neuropathology-associated DMPs are annotated to genes that have not previously been implicated in neurodegenerative disease. These include:

- cg09221482 (p=3.93e-09), located on chromosome 6 and annotated to *AGPAT4* which is a mitochondrial lysophosphatidic acid acyltransferase involved in the regulation of brain phosphatidylcholine, phosphatidylethanolamine, and phosphatidylinositol levels (Bradley *at al.*, 2015). Mice deficient in *AGPAT4* have impaired spatial learning and memory compared to wild-type mice and this deficiency was associated with reduced α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) and *N*-methyl-D-asparate (NMDA) receptors (Bradley *at al.*, 2017). AMPA and NMDA receptors play key roles in studies of learning, memory and neurotoxicity and there is a strong association with dysfunction in these receptors and Aβ (Danysz & Parsons, 2012).
- cg07256503 (p=4.78e-09) located on chromosome 9 and annotated to *NUP214* which is a component of the nuclear pore complex (NPC) with a role in protein and mRNA nuclear export (Port *at al.*, 2016). Studies suggest changes in NPC could lead to dysfunction in nucleocytoplasmic transport and this may contribute to abnormal filamentous (e.g. paired helical filaments) aggregates in AD (Sheffield, Miskiewicz, Tannenbaum, & Mirra, 2006).

- cg05384277 (p=1.48e-08) located on chromosome 6 and annotated to *ANKRD6* - a ubiquitous protein that is associated with early development in mammals and is highly expressed in the brain, spinal cord, and heart of humans (Van Deveire *at al.*, 2012).
- cg11732190 (p=1.66e-08) located on chromosome 3 and annotated to *PRRT3* which belongs to the family of proline-rich proteins and previous EWAS have identified multiple DMPs residing in the *PRRT1* and *PRRT2* loci (R. Smith *at al.*, 2021).
- cg16021126 (p=5.41e-08) located on chromosome 13 and annotated to *SERP2* which protects unfolded target proteins against degradation and facilitates correct glycosylation. Of note, this site had suggestive significance (p<5e-05) with Braak NFT stage in the recent EWAS meta-analysis (R. Smith *at al.*, 2021).

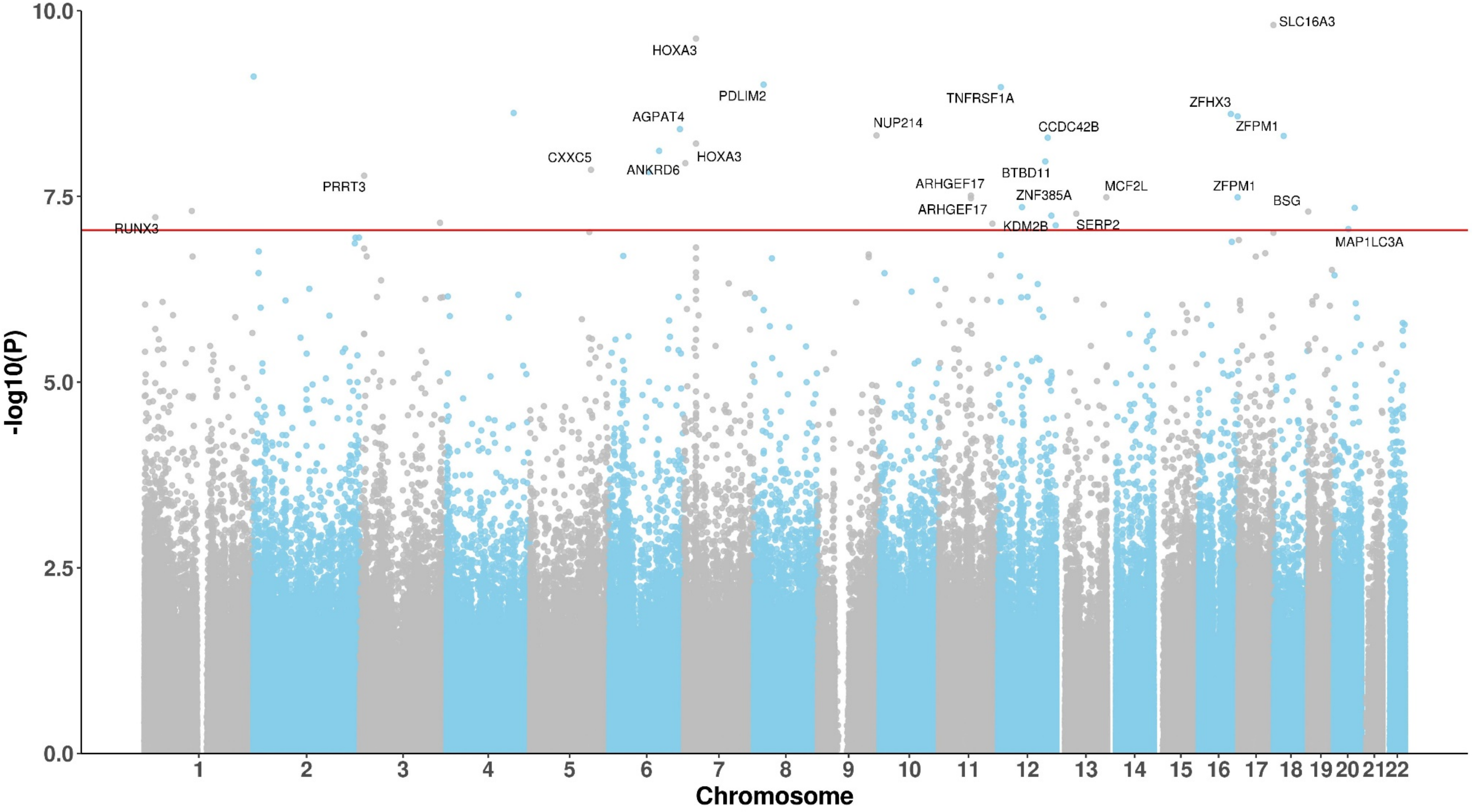


Figure 4.11: Cortex EWAS of neuropathology highlights experiment-wide significant differentially methylated positions. *A Manhattan plot showing results of a neuropathology EWAS conducted across two cortical regions (prefrontal cortex and occipital cortex) including five pathology measures (Braak NFT stage, CERAD density, Thal Phase, Braak LB stage and TDP43-status). The significant differentially methylated positions are annotated with their Illumina UCSC gene name, unless they are unannotated to a gene. The x-axis shows chromosomes 1-22 and the y-axis shows -log10(P), with the horizontal red line representing experiment wide significance (p < 9e-08).*

Table 4.7: Differentially methylated positions (DMPs) associated with neuropathology at an experiment wide significance threshold (p < 9e-08). *In total 34 DMPs were identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. The 'Array' column states if probes are also present on the Illumina 450K array. The 'Braak Meta P' relates to p-values for these probes from a recent meta-analysis of AD pathology (R. Smith at al., 2021). The 'Novel Gene' column indicates if the gene has previously been implicated in neurodegenerative disease and the 'Previously Identified in' column indicates how it was previously implicated in dementia.*

DNAm Site	P	Chr	BP	Nearest Gene	Genic location	Array	Braak meta P	Novel Gene	Previously Identified in
cg10208942	1.56e-10	17	80192754	*SLC16A3*	5'UTR	EPIC	-	previous	AD_EWAS
cg07061298	2.37e-10	7	27153847	*HOXA3*	5'UTR	450K & EPIC	4.57e-13	previous	AD_EWAS
cg08813888	7.69e-10	2	662396	-	-	450K & EPIC	8.85e-07	-	-
cg18100976	9.89e-10	8	22446737	*PDLIM2*	Body	EPIC	-	previous	AD_EWAS
cg18032191	1.07e-09	12	6443522	*TNFRSF1A*	Body	EPIC	-	previous	AD_GWAS
cg04459751	2.38e-09	4	153499517	-	-	EPIC	-	-	-
cg17877577	2.46e-09	16	73071822	*ZFHX3*	5'UTR	EPIC	-	previous	AD_EWAS
cg06913337	2.64e-09	16	88590404	*ZFPM1*	Body	450K & EPIC	2.79e-01	previous	AD_EWAS
cg09221482	3.93e-09	6	161557754	*AGPAT4*	Body	450K & EPIC	1.12e-04	novel	-
cg07256503	4.78e-09	9	134074137	*NUP214*	Body	EPIC	-	novel	-
cg23880946	4.85e-09	18	22539304	-	-	EPIC	-	-	-
cg15033653	5.13e-09	12	113587581	*CCDC42B*	TSS200	450K & EPIC	3.11e-06	previous	AD_EWAS
cg22962123	6.15e-09	7	27153605	*HOXA3*	5'UTR	450K & EPIC	2.21e-10	previous	AD_EWAS
cg07010192	7.71e-09	6	114056449	-	-	EPIC	-	-	-
cg13935577	1.07e-08	12	107974897	*BTBD11*	Body	450K & EPIC	8.17e-10	previous	AD_EWAS
cg19799702	1.13e-08	7	2757342	-	-	450K & EPIC	3.71e-03	-	-

cg14871225	1.39e-08	5	139040820	*CXXC5*	5'UTR	450K & EPIC	1.32e-07	previous	AD_EWAS
cg05384277	1.48e-08	6	90271986	*ANKRD6*	TSS200	EPIC	-	novel	-
cg11732190	1.66e-08	3	9989042	*PRRT3;PRRT3-AS1*	Body	EPIC	-	novel	-
cg20864214	3.09e-08	11	73054121	*ARHGEF17*	Body	450K & EPIC	1.68e-08	previous	AD_EWAS
cg07264904	3.25e-08	13	113633855	*MCF2L*	Body	450K & EPIC	7.76e-06	previous	AD_EWAS
cg09502865	3.25e-08	16	88600155	*ZFPM1*	Body	450K & EPIC	1.53e-08	previous	AD_EWAS
cg02776498	3.35e-08	11	73054185	*ARHGEF17*	Body	450K & EPIC	1.55e-07	previous	AD_EWAS
cg07332724	4.41e-08	12	54773114	*ZNF385A;LOC102724050*	Body	EPIC	-	previous	AD_EWAS
cg17247713	4.51e-08	20	47514319	-	-	EPIC	-	-	-
cg17571286	4.96e-08	1	108613442	-	-	EPIC	-	-	-
cg17640894	5.06e-08	19	574950	*BSG*	5'UTR	EPIC	-	novel	-
cg16021126	5.41e-08	13	44947611	*SERP2*	TSS1500	450K & EPIC	1.10e-05	novel	-
cg11724984	5.74e-08	12	121890864	*KDM2B*	Body	450K & EPIC	3.27e-09	previous	AD_EWAS
cg20695936	6.04e-08	1	25257624	*RUNX3*	TSS1500	450K & EPIC	1.04e-05	previous	AD_EWAS
cg25456014	7.17e-08	3	183278432	-	-	EPIC	-	-	-
cg10130088	7.37e-08	11	122138618	-	-	450K & EPIC	8.17e-04	-	-
cg16282686	7.77e-08	12	131707495	-	-	EPIC	-	-	-
cg27624319	8.70e-08	20	33147017	*MAP1LC3A*	Body	450K & EPIC	2.54e-06	previous	AD_EWAS

4.4.4 Individual neuropathology measures associated with DNA methylation

To identify DNAm sites at which differential DNAm is associated with each individual neuropathology measure, an EWAS was conducted sequentially for each measure (Braak NFT stage, CERAD density, Thal Phase, Braak LB stage and TDP43-status), controlling for age, sex, batch, derived cell proportions and PC1 as fixed effects and individual as a random effect. To identify if the effects were consistent across the two brain regions (OCC and PFC) I ran a further EWAS including an interaction between brain region and the measure of interest.

4.4.5 Braak NFT Stage-associated DNA methylation signatures across two cortical brain regions

I identified 26 experiment-wide significant DMPs associated with Braak NFT stage (see **Figure 4.12 and Table 4.8**). 15 (58%) of the DMPs were specific to the EPIC array, which demonstrates the utility of the newer platform and the advantage of the increased power in comparison to the 450K array; these 15 sites could not have been identified in the recent AD-EWAS of Braak NFT stage which meta-analysed 450K data from six cohorts (R. Smith *at al.*, 2021). The average magnitude of effect (i.e. the change in DNAm) for the significant DMPs per Braak stage was 0.44% (inter-quartile range [IQR] = 0.26-0.57%) and therefore there was a total mean DNAm change of 2.34% from Braak NFT stage 0-Braak NFT stage VI. Examples of DMP plots showing differences in DNAm across the different Braak stages is shown in **Figure 4.13**. The effects were highly consistent across the two regions we tested, as shown by the non-significant p-values (p > 9e-08) of the interaction term for these sites (see **Table 4.8**). 22 (83%) of the DMPs were significantly hypermethylated (i.e. increased methylation was associated with higher Braak NFT stage; see **Figure 4.14 and Table 4.8**) and the remaining 4 were hypomethylated (i.e. decreased methylation was associated with elevated Braak NFT stage). Of the 26 Braak NFT associated DMPs, 23 were annotated to genes (21 unique genes) and four of these genes have not previously implicated in dementia: *SERP2, PRRT3, GAST* and *RGS3*, although *SERP2* had been reported to have suggestive significance (p < 5e-05) in the latest AD-EWAS (R. Smith *at al.*, 2021; see **Table 4.8**). The DMPs annotated to *SERP2* (cg16021126) and *PRRT3* (cg11732190) were also identified in the EWAS against all neuropathology measures (see section **4.4.3.3** for a brief description of these genes), and both sites were significantly hypermethylated with increased Braak NFT stage (p=7.48e-10, p=3.48e-09, respectively).

4.4.5.1 Several Braak NFT-associated DMPs are of relevance in the context of AD and associated neurobiological functions

Several of the genes annotated to the significant DMPs have been previously implicated in AD from EWAS, GWAS or animal model studies. For example, cg08952306 was significantly hypermethylated (p=1.36e-08) with elevated Braak NFT score and is annotated to *SH2B2* – an adaptor signalling protein involved in obesity, insulin resistance, and glucose intolerance (Jamshidi, Snieder, Ge, Spector, & O'Dell, 2007). Recent research by Yijun and colleagues found that partial knockout of neuronal *SH2B* in Aβ expressing drosophila has an impact on their mobility and neurotransmission and in addition there was increased accumulation of Aβ in these drosophila (Shen *at al.*, 2017). They concluded that *SH2B1* is likely an upstream modulator of Aβ metabolism and suggests it has a potential role in AD pathogenesis. cg18032191 was significantly hypermethylated with elevated Braak NFT stage (p=7.13e-08) as well as general neuropathology (see **4.4.5.1**). It is annotated to *TNFRSF1A* (located on chromosome 12), for more details on this gene see **4.4.5.1.**

4.4.5.2 Several Braak NFT-associated DMPs have not previously been implicated in neurodegenerative disease

cg02553166 is annotated to a *GAST* which has not previously been implicated in dementia and was significantly hypermethylated with elevated Braak NFT stage (p=1.80e-08). *GAST* (located on chromosome 17) encodes for gastrin - a hormone whose function is to stimulate secretion of hydrochloric acid and the migration of gastric epithelial cells (Czinn & Blanchard, 2011). Previous studies have identified that the gastrin-releasing peptide (GRP) plays a role in neurological disorders (Roesler, Henriques, & Schwartsmann, 2006; Yang *at al.*, 2017). GRP elicits gastrin release and has been implicated in memory formation (Roesler *at al.*, 2006; Yang *at al.*, 2017). cg16107559 was significantly hypermethylated with elevated Braak NFT stage (p=6.85e-08) and is annotated to *RGS3* (located on chromosome 9) – which encodes a member of the regulator of G-protein signalling (RGS) family and specifically this protein is a GTPase-activating protein that inhibits G-protein-mediated signal transduction and has predominantly been associated with cardiac functioning (Yazdani, Yazdani, Méndez Giráldez, Aguilar, & Sartore, 2019).

Several genes were annotated to multiple DMPs, including *SERP2* (cg16021126 and cg01226614), *ZFPM1* (cg06913337 and cg09502865) and *ARHGEF17* (cg02776498 and cg20864214), strengthening the evidence for a role of these genes in AD pathology as a consequence of epigenetic dysregulation.

Table 4.8: Differentially methylated positions (DMPs) associated with Braak neurofibrillary tangle stage at an experiment wide significance threshold (p<9e-08). *In total 26 DMPs were identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. The Beta columns refers to %methylation change per unit increase in Braak NFT Stage. The 'Array' column states if probes are also present on the Illumina 450K array. The 'Braak Meta P' relates to p-values for these probes from a recent meta-analysis of AD pathology (R. Smith at al., 2021). The 'Novel Gene' column indicates if the gene has previously been implicated in neurodegenerative disease and the 'Previously Identified in' column indicates how it was previously implicated in dementia.*

DNAm site	Beta (%)	SE (%)	P	Chr	BP	Gene	Gene Region	Array	Braak meta P	Interaction BR P	Novel Gene	Previously Identified in
cg16021126	0.29	3.72e-02	7.48e-10	13	44947611	*SERP2*	TSS1500	450K & EPIC	1.10e-05	5.24e-01	novel	-
cg07061298	0.38	5.06e-02	2.16e-09	7	27153847	*HOXA3*	5'UTR; TSS1500	450K & EPIC	4.57e-13	5.70e-01	previous	AD_EWAS
cg06913337	-0.66	8.81e-02	2.68e-09	16	88590404	*ZFPM1*	Body	450K & EPIC	2.79e-01	2.49e-01	previous	AD_EWAS
cg11732190	0.41	5.53e-02	3.48e-09	3	9989042	*PRRT3;PRRT3-AS1*	Body; TSS200	EPIC	-	5.69e-01	novel	-
cg08813888	-0.36	4.92e-02	4.62e-09	2	662396	-	-	450K & EPIC	8.85e-07	6.03e-02		-
cg18100976	0.41	5.55e-02	4.82e-09	8	22446737	*PDLIM2*	Body	EPIC	-	6.23e-01	previous	AD_EWAS
cg14871225	-0.57	7.98e-02	9.29e-09	5	139040820	*CXXC5*	5'UTR	450K & EPIC	1.32e-07	3.51e-01	previous	AD_EWAS
cg11724984	0.43	6.11e-02	1.31e-08	12	121890864	*KDM2B*	Body	450K & EPIC	3.27e-09	7.27e-02	previous	AD_EWAS
cg08952306	0.59	8.28e-02	1.36e-08	7	101962123	*SH2B2*	3'UTR	450K & EPIC	1.27e-02	9.50e-01	previous	Animal model
cg07332724	0.33	4.73e-02	1.59e-08	12	54773114	*ZNF385A;* *LOC102724050*	Body	EPIC	-	9.63e-01	previous	AD_EWAS
cg02553166	-0.35	4.94e-02	1.80e-08	17	39867080	*GAST*	TSS1500	EPIC	-	5.10e-01	novel	-

cg01141438	0.40	5.72e-02	2.43e-08	3	189836207	*LEPREL1*	5'UTR; Body	450K & EPIC	9.60e-06	5.00e-01	previous	AD_EWAS
cg27624319	0.55	7.87e-02	2.69e-08	20	33147017	*MAP1LC3A*	Body	450K & EPIC	2.54e-06	9.50e-01	previous	AD_EWAS
cg11658414	0.26	3.71e-02	2.80e-08	20	1450828	-	-	EPIC	-	1.29e-01	-	-
cg02776498	0.32	4.68e-02	4.59e-08	11	73054185	*ARHGEF17*	Body	450K & EPIC	1.55e-07	8.04e-02	previous	AD_EWAS
cg06329036	0.31	4.61e-02	4.65e-08	4	7669886	*SORCS2*	Body	EPIC	-	4.11e-01	previous	AD_GWAS
cg12077223	0.35	5.13e-02	4.88e-08	11	74879354	*SLCO2B1*	Body	EPIC	-	2.75e-03	previous	AD_EWAS
cg20864214	0.60	8.81e-02	5.27e-08	11	73054121	*ARHGEF17*	Body	450K & EPIC	1.68e-08	7.51e-01	previous	AD_EWAS
cg17149093	0.41	6.04e-02	5.60e-08	10	6379445	-	-	EPIC	-	1.01e-01	-	-
cg13935577	1.00	1.49e-01	6.34e-08	12	107974897	*BTBD11*	Body	450K & EPIC	8.17e-10	7.89e-01	previous	AD_EWAS
cg09490371	0.66	9.85e-02	6.81e-08	2	233253024	*ECEL1P2*	TSS1500	450K & EPIC	8.23e-09	3.55e-01	previous	AD_EWAS
cg16107559	0.58	8.64e-02	6.85e-08	9	116225834	*RGS3*	TSS200; Body	EPIC	-	1.35e-01	novel	-
cg18032191	0.32	4.78e-02	7.13e-08	12	6443522	*TNFRSF1A*	Body	EPIC	-	8.53e-01	previous	AD_GWAS
cg01226614	0.24	3.60e-02	7.37e-08	13	44947593	*SERP2*	TSS1500	450K & EPIC	9.97e-05	4.94e-01	novel	-
cg09502865	0.38	5.61e-02	7.97e-08	16	88600155	*ZFPM1*	Body	450K & EPIC	1.53e-08	3.00e-01	previous	AD_EWAS
cg15033653	0.23	3.39e-02	8.07e-08	12	113587581	*CCDC42B*	TSS200	450K & EPIC	3.11e-06	7.70e-01	previous	AD_EWAS

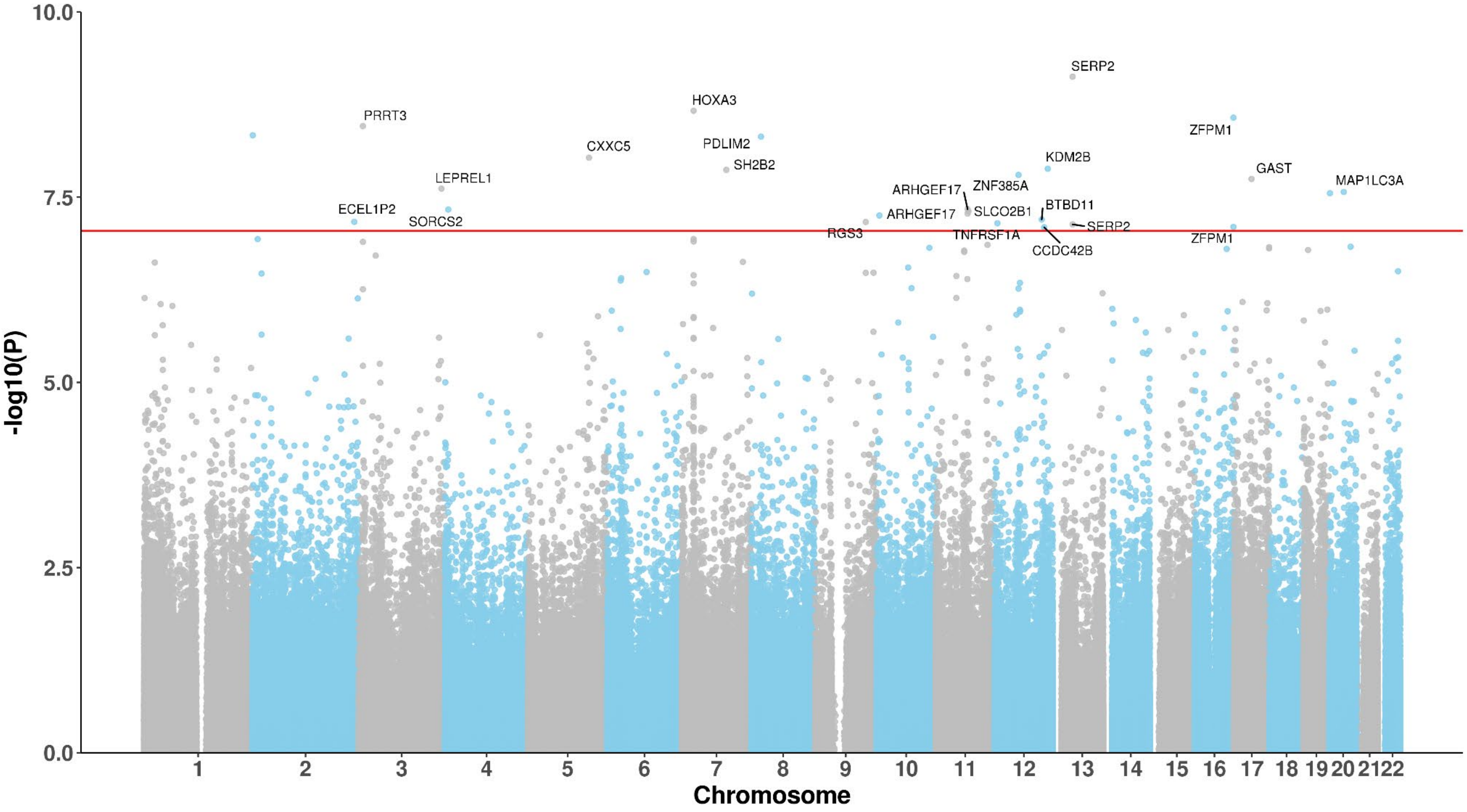


***Figure 4.12: Cortex EWAS of Braak neurofibrillary tangle stage highlights experiment-wide significant differentially methylated positions.** A Manhattan plot showing results of a Braak NFT stage EWAS across two cortical regions (prefrontal cortex and occipital cortex). The significant differentially methylated positions are annotated with their Illumina UCSC gene name, unless they are unannotated to a gene. The x-axis shows chromosomes 1-22 and the y-axis shows -log10(P), with the horizontal red line representing experiment wide significance (p< 9e-8).*

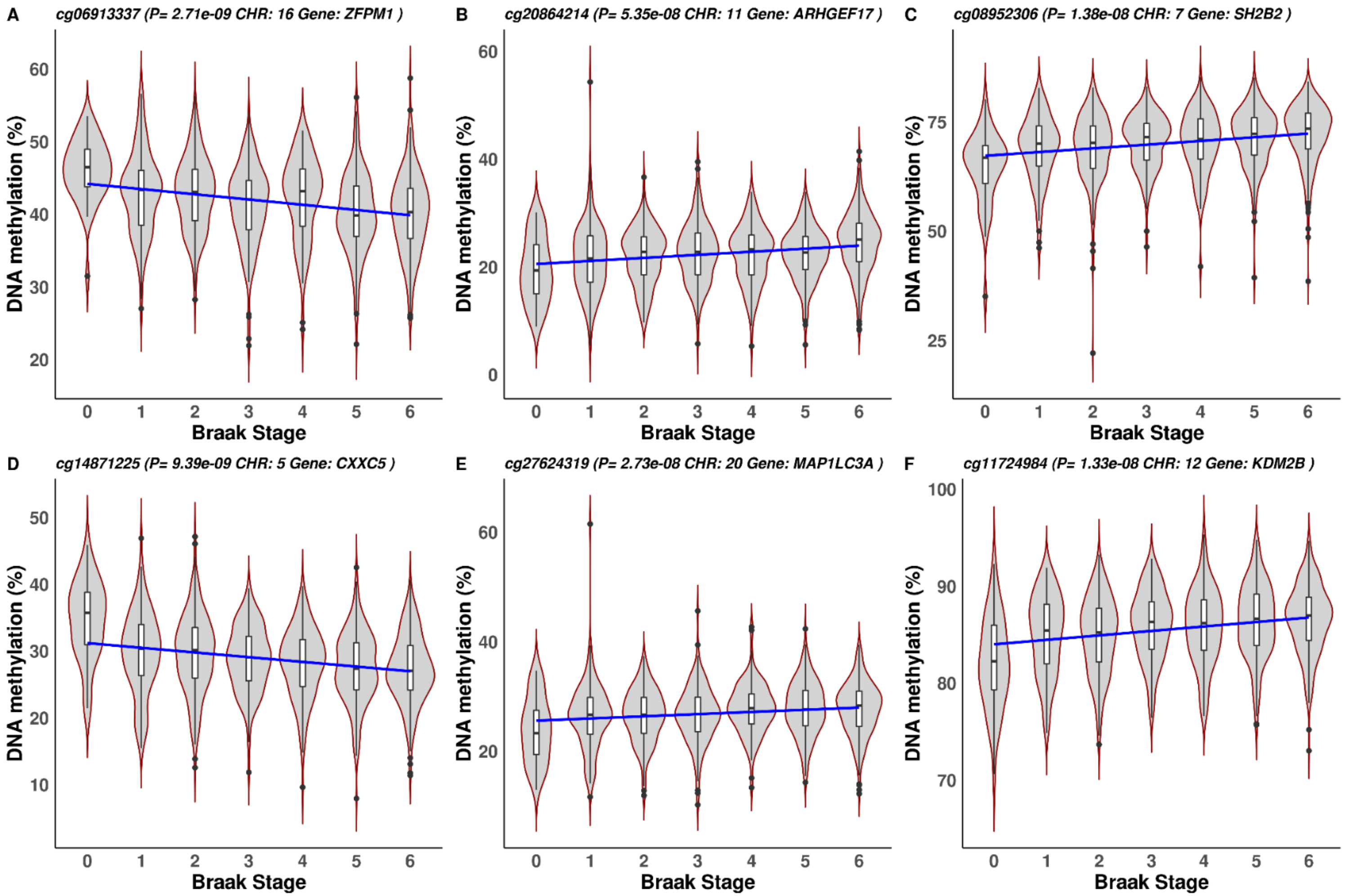


Figure 4.13: Differential methylation across the Braak neurofibrillary tangle stages for six of the top differentially methylated positions identified by EWAS. *The plots are for the six top differentially methylated positions with the strongest beta values. Violin plots for the DNA methylation values (adjusted for batch and covariates) across the 7 stages are shown, where the box in the middle represents the interquartile range (IQR), whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). The grey on the outside of the box plot is a density plot of the distribution of data, where the width represents frequency. The x-axis is Braak neurofibrillary tangle (BFT) stage. The y-axis is the methylation value as a percentage.*

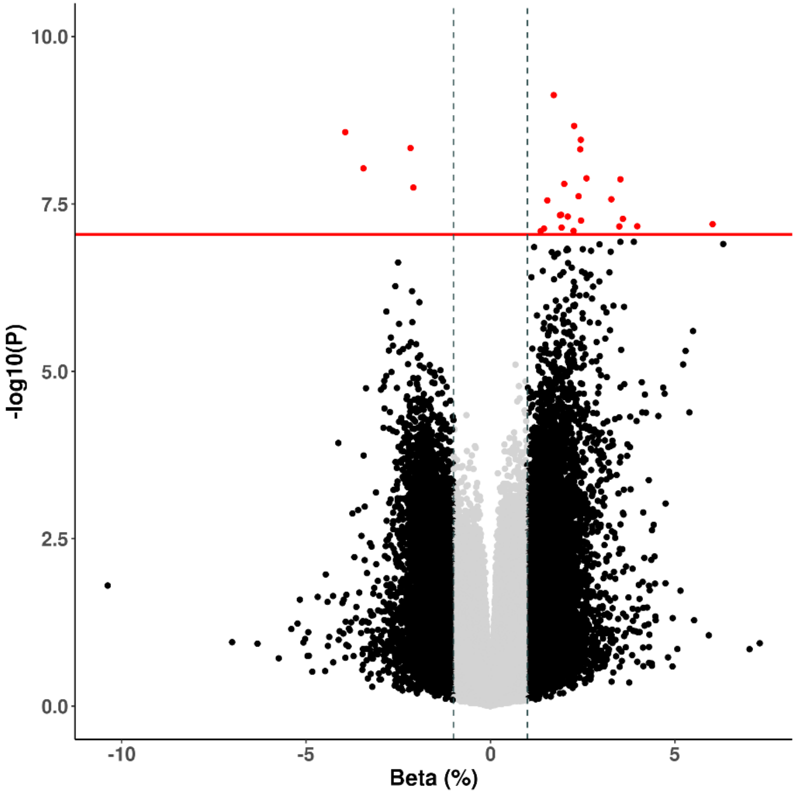


Figure 4.14: Volcano plot of differentially methylated positions (DMPs) identified in the Braak neurofibrillary tangle stage EWAS. *The x-axis shows beta effect size (ES) and the Y-axis shows -log10(p). Black probes indicate an ES difference ≥ 0.01, whilst red probes indicate an ES difference ≥ 0.01 and a p-value that reaches experiment-wide significance (EWS) (p < 9e-08).*

4.4.5.3 The results validate in the recent Braak NFT stage meta-analysis

I compared the results with the summary statistics from the cross-cortex Braak NFT meta-analysis recently published by our group (R. Smith *at al.*, 2021). I performed a binomial sign test of effect sizes to statistically evaluate consistency across the studies. 16 of the 26 significant sites were tested in the meta-analysis, which was primarily undertaken on samples profiled using the Illumina 450K array. Of these 16, 6 (38%) reached Bonferroni significance (p<9e-08) in the meta-analysis and 7 additional DMPs (44%) reached suggestive significance (p<5e-05). The direction of effect for the 16 significant DMP probes identified in the BDR Braak NFT EWAS was consistent across these studies (sign test p =1.5e-05, see **Figure 4.15**). In addition, of the 220 significant DMPs identified in the cross-cortex meta-analysis, 208 were tested in the BDR analysis; again the direction of effect was consistent across studies (sign test p =5.1e-61, see **Figure 4.15**). These results show how robust the EWAS results for AD pathology are. In addition, the strength of the binomial sign test suggests there may have been a number of false negatives in the recent meta-analysis, where DMPs did not reach the stringent p-value threshold but likely represent true associations for Braak NFT Stage.

4.4.5.4 Multiple DMRs associated with Braak NFT stage

I next conducted analyses to identify differentially methylated regions (DMRs) using the R package *dmrff* (Suderman *at al.*, 2018). I identified 61 DMRs, annotated to 54 genes (p adjusted < 0.05). The DMRs are shown in **Table 4.9** and example plots of DMRs spanning ≥6 probes are shown in **Figure 4.16.** Multiple DMRs were identified in the *HOXA* region, a region which has consistently been identified to have an association with AD pathology in EWAS studies (R. G. Smith *at al.*, 2018; R. Smith *at al.*, 2021). One of the largest and most significant DMRs was located in the gene *KLHL33*, which encompassed 11 DNAm sites (p adjusted = 1.60e-10; see **Figure 4.17**). *KLHL33* is a protein coding gene and is predominantly expressed in the brain in primary hippocampal neurons, astrocytes and oligodendrocytes and studies suggest it plays a role in the functioning and development of the nervous system and in the differentiation of oligodendrocytes (Jiang *at al.*, 2005; Soltysik-Espanola *at al.*, 1999).

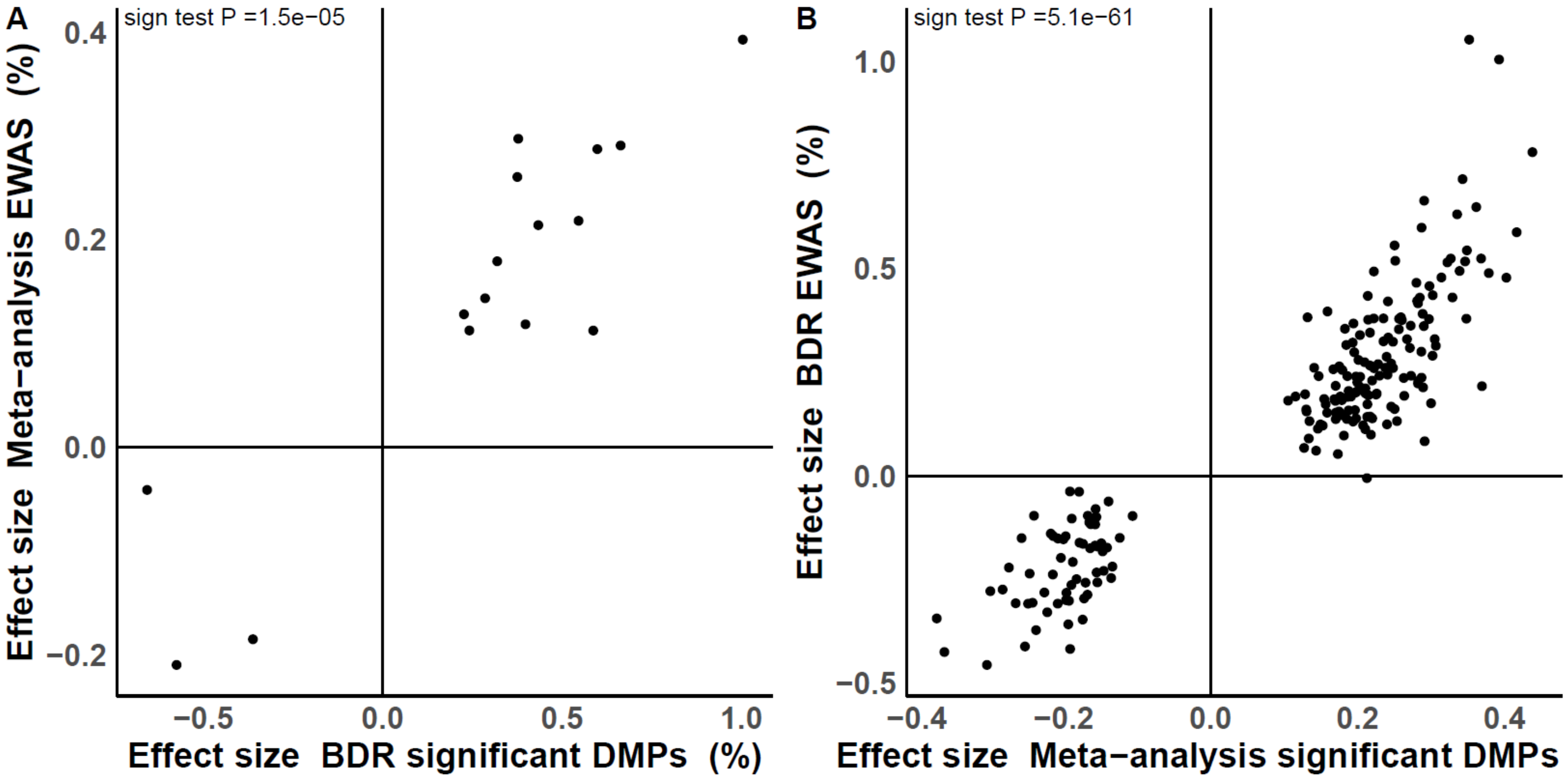


Figure 4.15: The effects sizes of the Braak neurofibrillary tangle stage (NFT) EWAS are consistent between the BDR EWAS and the recent meta-analysis of Braak NFT Stage conducted by Smith and colleagues (R. Smith et al., 2021). ***(A)*** *Compares the effects sizes from the Braak NFT EWAS differentially methylated probes conducted in BDR against the effect sizes of the probes which were run in the recent meta-analysis (16 probes).* ***(B)*** *Compares the effects sizes of the meta-analysis significant cross-cortex differentially methylated probes to the Braak NFT EWAS results (n=208).*

Table 4.9: Differentially methylated regions (DMRs) associated with Braak neurofibrillary tangle stage. *In total 61 DMRs were identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected p-value adjusted for the number of independent tests.*

Chr	**BP start**	**BP end**	**n**	**Gene**	**Beta (%)**	**SE (%)**	**P**	**P adjusted**
14	73520468	73520759	11	*KLHL33*	5.51	0.67	1.87e-16	1.60e-10
7	20903410	20904169	6	*HOXA3*	9.17	1.16	2.25e-15	1.92e-09
12	73054121	73054185	4	*CCDC42B*	6.71	0.89	5.07e-14	4.34e-08
12	27153580	27153847	7	*AGAP2*	8.37	1.13	1.32e-13	1.13e-07
11	101962112	101962123	4	*ATG16L2*	9.00	1.22	1.79e-13	1.53e-07
3	113587513	113587690	4	*PRRT3;PRRT3-AS1*	9.24	1.26	2.35e-13	2.02e-07
17	58131345	58132114	7	*SLC16A3*	6.27	0.87	5.07e-13	4.34e-07
1	72533202	72533664	5	*HEYL*	5.48	0.81	1.46e-11	1.25e-05
15	9988662	9989042	6	*STRA6*	5.26	0.79	2.58e-11	2.20e-05
15	80192161	80192655	4	*PSTPIP1*	5.02	0.76	3.16e-11	2.70e-05
7	54772804	54773114	3	*HOXA6*	6.06	0.92	3.56e-11	3.04e-05
14	80192754	80192794	7	-	-4.51	0.68	4.56e-11	3.91e-05
6	107974897	107975299	12	*ZBTB12*	3.61	0.55	5.69e-11	4.87e-05
9	116225793	116225834	3	*RUSC2*	7.24	1.11	6.40e-11	5.48e-05
11	40098781	40099015	3	-	-4.09	0.63	6.77e-11	5.79e-05
22	6443522	6443533	3	*LOC150381;C22orf26*	7.53	1.15	6.97e-11	5.97e-05
1	74494909	74495384	6	*ACAP3*	3.51	0.54	7.83e-11	6.71e-05
19	20648194	20648580	7	*CCER2*	3.31	0.51	9.03e-11	7.73e-05
2	77324526	77324758	7	*SPEG*	4.54	0.71	2.01e-10	1.72e-04

17	27185136	27185512	4	*ATP2A3*	7.32	1.16	2.54e-10	2.18e-04
19	574950	575412	3	*MAP4K1*	6.54	1.04	3.46e-10	2.96e-04
16	106330224	106331073	3	*RLTPR*	8.90	1.44	5.84e-10	5.00e-04
6	31868949	31869565	7	*SYNJ2*	5.69	0.92	6.13e-10	5.25e-04
6	35538890	35539007	3	*FYN*	-3.21	0.52	7.24e-10	6.20e-04
2	10920810	10921186	3	-	6.93	1.14	1.07e-09	9.19e-04
6	46451281	46451518	5	*COL11A2*	4.71	0.78	1.50e-09	1.28e-03
3	1228990	1229534	3	*PLCH1*	-2.59	0.43	1.80e-09	1.54e-03
1	39402823	39403373	3	*DUSP27*	-8.31	1.39	2.07e-09	1.77e-03
14	46483916	46484046	3	*C14orf115*	6.16	1.03	2.37e-09	2.03e-03
11	220298547	220299584	4	*RASSF7;C11orf35*	6.33	1.07	3.27e-09	2.80e-03
11	3848156	3848506	3	-	7.47	1.26	3.40e-09	2.91e-03
19	54764265	54764371	6	*PNMAL2*	5.70	0.97	4.01e-09	3.43e-03
7	39086923	39087186	6	*ACTB*	4.58	0.78	4.02e-09	3.44e-03
9	74475270	74475294	3	-	4.13	0.70	4.19e-09	3.58e-03
3	43814764	43814983	3	*TPRA1;MIR6825*	-5.70	0.98	5.19e-09	4.44e-03
11	23836012	23836047	3	*ARHGEF17*	6.31	1.08	5.64e-09	4.83e-03
3	67686832	67687119	5	-	-5.10	0.88	5.75e-09	4.93e-03
5	158438072	158438419	3	*SIL1*	-2.81	0.48	5.81e-09	4.97e-03
6	2005132	2005180	3	*PPT2;PRRT1*	6.00	1.04	6.66e-09	5.70e-03
6	58129855	58130154	4	*AGPAT4*	5.52	0.96	1.00e-08	8.59e-03
6	112042339	112042632	4	-	-4.42	0.77	1.01e-08	8.69e-03

16	44278551	44278628	5	*EMP2*	3.15	0.55	1.08e-08	9.24e-03
7	202814226	202814433	4	*HOXA5*	5.98	1.05	1.14e-08	9.78e-03
7	71868412	71868494	5	*HOXA3*	6.57	1.15	1.19e-08	1.01e-02
10	33132181	33132442	3	*C10orf54;CDH23*	5.39	0.95	1.37e-08	1.18e-02
7	155421970	155422159	3	*MYO1G*	5.71	1.01	1.50e-08	1.29e-02
17	56358318	56358504	5	*CAMTA2*	4.90	0.87	1.71e-08	1.46e-02
6	46471129	46471442	13	*HLA-DPB1*	5.54	0.99	2.17e-08	1.85e-02
1	167090618	167090757	5	*PBX1*	5.89	1.05	2.18e-08	1.87e-02
7	110349231	110349639	3	*RBM33*	5.67	1.01	2.20e-08	1.88e-02
9	74815131	74815187	4	*C9orf142*	6.42	1.15	2.39e-08	2.04e-02
7	27162780	27163095	4	*HOXA4*	6.53	1.17	2.65e-08	2.27e-02
11	560921	560951	3	*SLC22A8*	-5.32	0.96	2.78e-08	2.38e-02
12	44642868	44642932	4	*KRT86*	4.94	0.89	2.87e-08	2.46e-02
19	46999109	46999307	11	*C19orf36;MOBKL2A*	4.28	0.77	2.92e-08	2.50e-02
2	5569860	5570491	3	*CERKL*	5.95	1.07	3.03e-08	2.59e-02
1	43296491	43296526	4	*IL12RB2*	-4.46	0.81	3.11e-08	2.66e-02
1	138271837	138272020	5	*FLJ42875*	4.26	0.78	4.02e-08	3.44e-02
16	170585437	170585664	4	*PKD1*	4.66	0.85	4.68e-08	4.01e-02
1	59507038	59507393	3	*LOC101927851*	3.92	0.72	5.02e-08	4.29e-02
6	25257626	25257629	4	*PPT2;PRRT1*	5.04	0.93	5.35e-08	4.58e-02

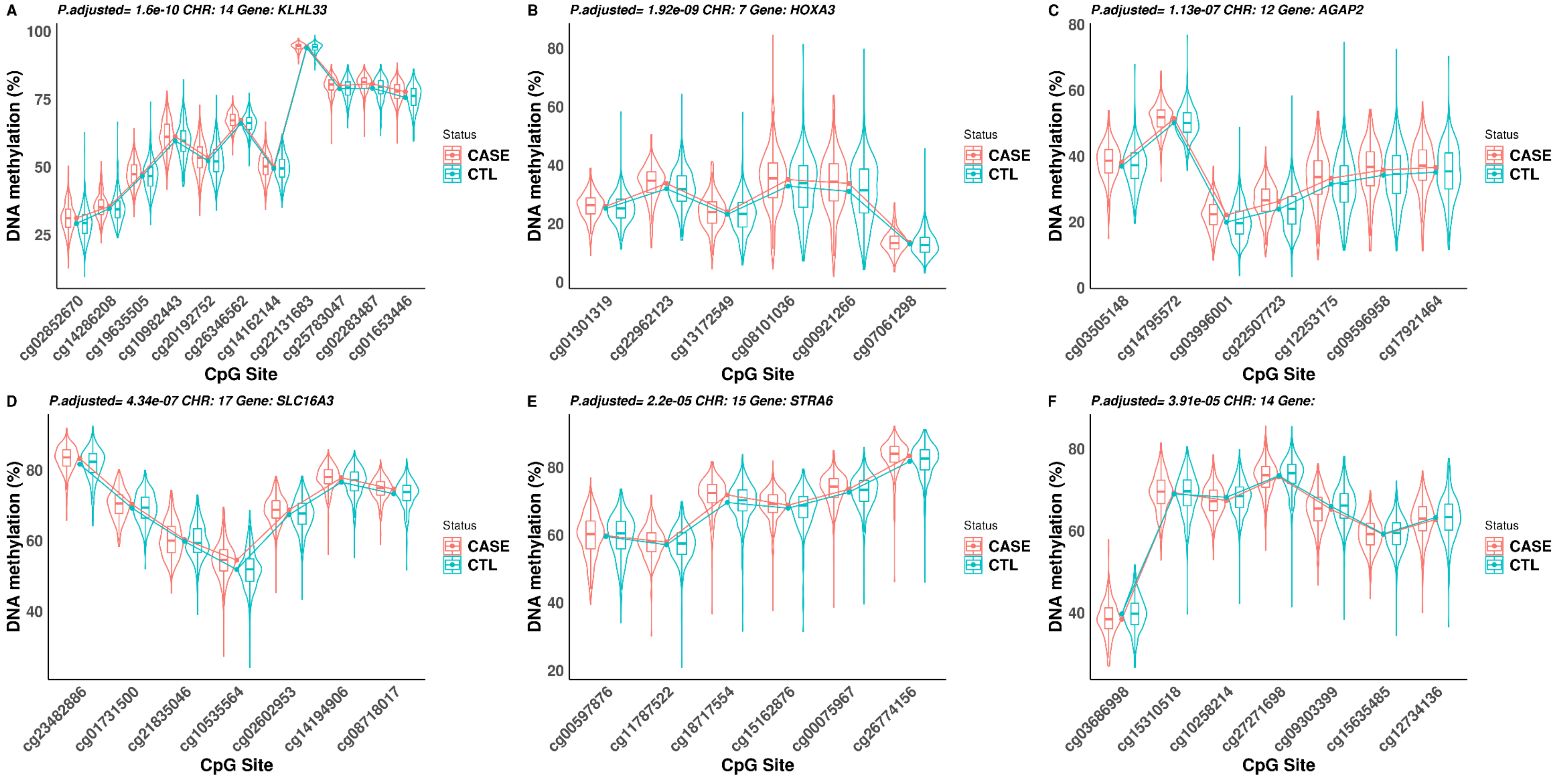


***Figure 4.16: Differentially methylated regions (DMRs) associated with Braak neurofibrillary tangle (NFT) stage.** An example of six DMRs which were associated with Braak NFT stage are shown. The x-axis are the DNAm sites ordered by genomic location (not to scale). The y-axis is DNA methylation as a percentage. Violin plots for the DNA methylation values (adjusted for batch and covariates) across at each DNAm site in the DMR is shown, where the box in the middle represents the interquartile range (IQR), whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). Samples were split into cases (Braak NFT stage > 4) and controls (Braak NFT stage < 3) in order to visualise the differences between the groups. The grey on the outside of the box plot is a density plot of the distribution of data, where the width represents frequency.*

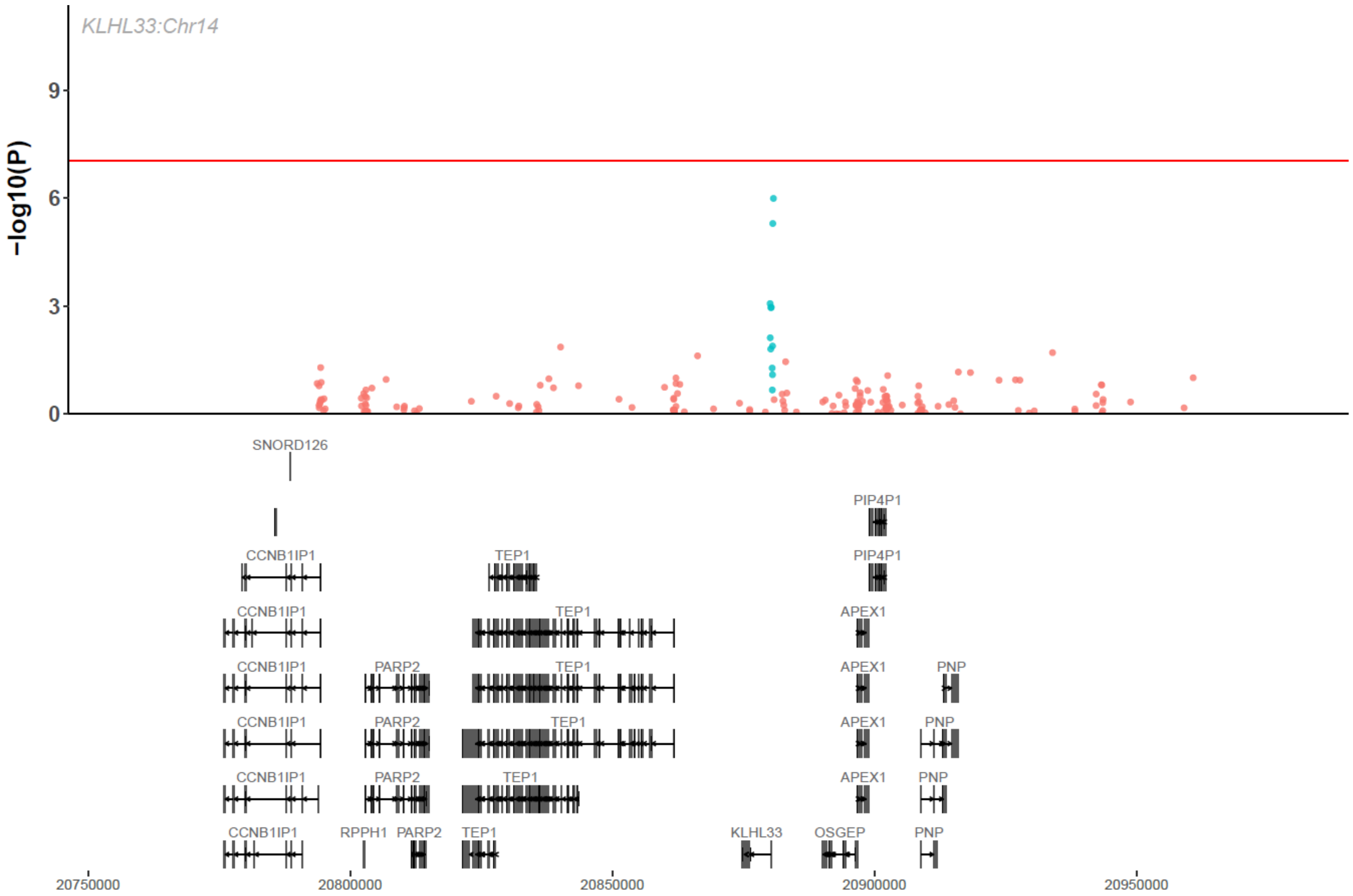


Figure 4.17: A differentially methylated region identified in the KLHL33 gene within chromosome 14 was associated with Braak NFT Stage. *Shown on the top is a zoomed in Manhattan plot around the KLHL33 region, where the x-axis represents the base position and the y-axis represents the –log10 p-value and each point on the plot represents a DNA methylation site. The red horizontal line represents experiment wide significance (P< 9e-08). The gene track shows the locations of the genes within in this region.*

4.4.5.5 Interaction effects between brain region and Braak NFT Stage

I investigated interactions between brain region and Braak NFT stage to identify if DNAm differences associated with AD pathology differed between the PFC and OCC. A single DMP reached experiment wide significance (where the interaction p-value <9e-08) - cg15984835 (annotated to chromosome 2; p=5.70e-08), suggesting there are differences between the two brain regions where DNAm decreases at a steeper gradient in the OCC than the PFC as Braak NFT stage increases (see **Figure 4.18**). This probe has no gene annotation. The lack of significant interaction effects suggest the majority of associations between Braak NFT stage and DNAm are consistent across the cortex. However, at a more relaxed threshold of p<5e-5 there were 79 differences (see **Table 4.10** for the top 20). Plots of top DMPs with the most significant interaction terms are shown in **Figure 4.18.** Some of the genes are of interest in relation to AD. For example, cg14721213 (annotated to chromosome 1) had a suggestive significant interaction term (p=6.70e-07) and is annotated to the gene *FMO2*. *FMO2* has been shown to be upregulated in the PFC, hippocampus and caudate in mice injected with silver-25 nm (Ag-25) – a nanoparticle found in everyday products such as toothpaste, shampoos, fabrics, deodorants, and kitchen utensils (Rahman *at al.*, 2009). They concluded that that Ag-25 nanoparticles may cause neurodegeneration via radical-induced oxidative stress altering gene expression with the consequence of apoptosis and neurotoxicity (Rahman *at al.*, 2009).

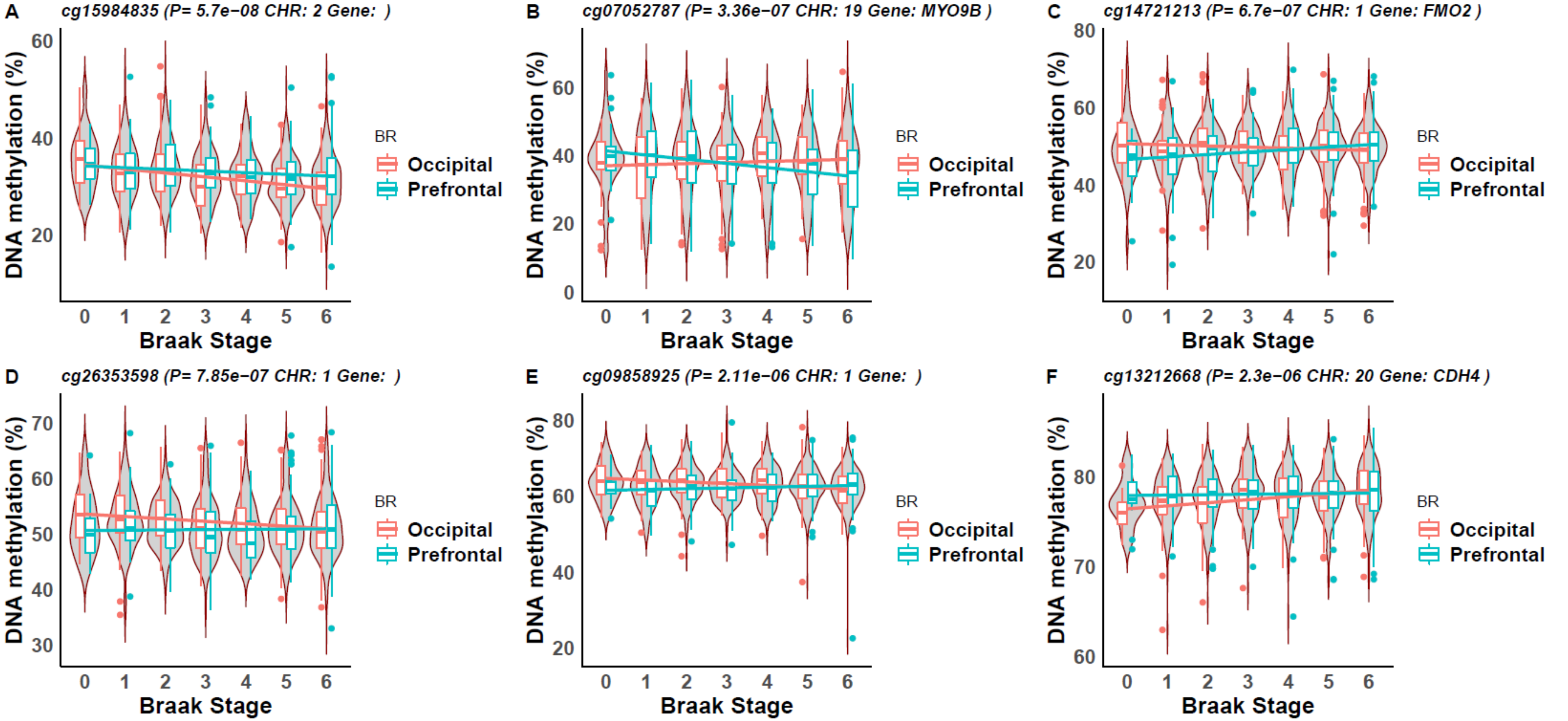


Figure 4.18: Interaction between Braak neurofibrillary tangle (NFT) stage and brain region. *Shown are the interaction results for the top 6 DNAm sites from the interaction EWAS, where an interaction term between Braak NFT stage and brain region was included. The plots are for the six most associated differentially methylated positions. Violin plots for the DNA methylation values (adjusted for batch and covariates) across the 7 stages are shown, where the boxes in the middle represents the interquartile range (IQR) for each brain region, whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). The grey on the outside of the box plot is a density plot of the distribution of data (not split by brain region), where the width represents frequency. The x-axis is Braak NFT stage. The y-axis is the methylation value as a percentage.*

Table 4.10: The top 20 differentially methylated positions identified in the interaction EWAS between brain region and Braak neurofibrillary tangle (NFT) stage. *Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. Shown for each DMP is the effect size (Beta), standard error (SE) and the p-value (P) for the effect of Braak NFT Stage on DNAm in addition to the interaction effect (Beta Int, SE Int and P Int) between Braak NFT stage and brain region (occipital cortex [reference] and prefrontal cortex).*

DNAm site	Beta (%)	SE (%)	P	Beta Int (%)	SE Int (%)	P Int	CHR	MAPINFO	Gene	Ref Group
cg15984835	-0.50	8.82e-02	2.98e-06	0.64	1.17e-01	5.70e-08	2	113444352	-	-
cg07052787	0.51	1.28e-01	9.58e-04	-0.79	1.54e-01	3.36e-07	19	17237120	*MYO9B*	Body
cg14721213	-0.44	1.32e-01	5.77e-03	0.89	1.78e-01	6.70e-07	1	171154612	*FMO2*	5'UTR
cg26353598	-0.32	1.10e-01	1.43e-02	0.57	1.14e-01	7.85e-07	1	14824776	-	-
cg09858925	-0.48	1.05e-01	1.25e-04	0.67	1.40e-01	2.11e-06	1	209537469	-	-
cg13212668	0.20	5.27e-02	1.32e-03	-0.34	7.15e-02	2.30e-06	20	59928453	*CDH4*	Body
cg14722315	-0.55	1.43e-01	1.54e-03	0.56	1.19e-01	2.88e-06	1	19240707	*IFFO2*	Body
cg10095305	-0.42	1.36e-01	1.12e-02	0.88	1.89e-01	3.43e-06	5	36158217	*SKP2*	5'UTR;Body
cg00739139	-0.32	8.26e-02	1.44e-03	0.51	1.10e-01	4.24e-06	7	65233991	-	-
cg14680131	0.24	8.58e-02	2.21e-02	-0.52	1.13e-01	4.27e-06	1	156131379	*SEMA4A*	Body
cg02715788	0.26	9.01e-02	1.47e-02	-0.57	1.23e-01	4.60e-06	8	119974400	-	-
cg00097088	-0.56	1.01e-01	4.04e-06	0.62	1.36e-01	5.17e-06	4	2627246	*FAM193A*	1stExon;5'UTR
cg03546163	0.24	9.55e-02	3.34e-02	-0.61	1.33e-01	5.53e-06	6	35654363	*FKBP5*	5'UTR
cg17734698	-0.31	1.01e-01	9.82e-03	0.57	1.26e-01	6.84e-06	12	129127968	*TMEM132C*	Body
cg23536138	-0.28	6.75e-02	6.71e-04	0.39	8.48e-02	6.94e-06	8	22084861	*PHYHIP*	Body
cg26135325	0.21	8.39e-02	3.97e-02	-0.47	1.04e-01	7.14e-06	1	152595322	*LCE3A*	1stExon
cg23707289	-0.14	5.37e-02	3.11e-02	0.30	6.53e-02	7.71e-06	18	60988099	*BCL2*	TSS1500
cg24924930	-0.61	1.39e-01	2.42e-04	0.79	1.75e-01	8.16e-06	2	241717598	*KIF1A*	Body
cg20030796	-0.43	1.01e-01	4.17e-04	0.62	1.40e-01	1.01e-05	17	72356254	*BTBD17*	Body
cg07305369	-0.34	8.13e-02	5.95e-04	0.48	1.09e-01	1.05e-05	11	64270338	-	-

4.4.6 Thal Phase-associated DNA methylation signatures across two cortical brain regions

I identified two experiment-wide significant DMPs (p<9e-08) associated with Thal phase, both hypermethylated with increasing pathology (see **Figure 4.19** and **Table 4.11**). The average magnitude of effect for the significant DMPs per Thal Phase was 0.30% (inter-quartile range [IQR] = 0.30-0.31%) and therefore there was a total mean DNAm change of 1.5% from Thal Phase 0 to Thal Phase 5. Changes in DNAm at these DMPs across Thal Phases is shown in **Figure 4.13**. The effects were consistent across the regions as shown by the non-significant interaction term between Thal Phase and brain region (p>9e-08; see **Table 4.11**). cg13515047 (located on chromosome 16) was significantly hypermethylated with elevated Thal phase (p=6.04e-08) and annotated to the gene *BCAR1*, a Cas protein involved in a range of cellular processes such as cell migration, apoptosis and cell cycle control and progenitor cell functioning (Tikhmyanova, Little, & Golemis, 2010). Although research into cas proteins has predominantly highlighted their role in cancer, they have also been linked to AD and PD (Y. Li *at al.*, 2008). The other DMP (cg11658414) is located on chromosome 20 and is not annotated to any gene.

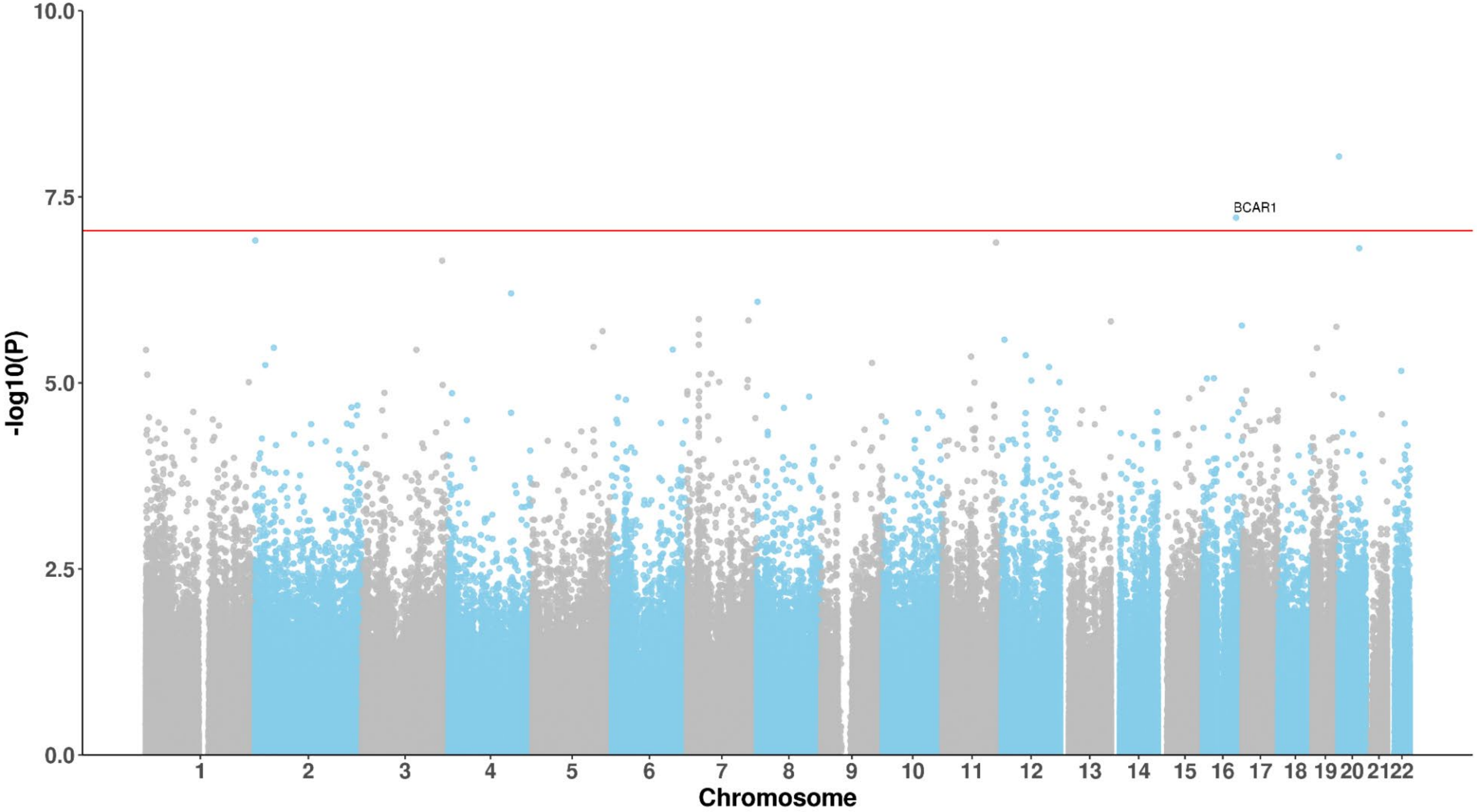


Figure 4.19: Cortex EWAS of Thal Phase stage highlights experiment-wide significant differentially methylated positions. *A Manhattan plot showing results of a Thal Phase EWAS across two cortical regions (prefrontal cortex and occipital cortex). The significant differentially methylated positions are annotated with their Illumina UCSC gene name unless they are unannotated to a gene. The x-axis shows chromosomes 1-22 and the Y-axis shows -log10(P), with the horizontal red line representing experiment wide significance (p < 9e-08).*

Table 4.11: Differentially methylated positions (DMPs) associated with Thal phase at an experiment wide significance threshold (p < 9e-08). *In total two DMPs were identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. The Beta columns refers to %methylation change per unit increase in Thal phase Stage. The 'Array' column states if probes are also present on the Illumina 450K array. The 'Braak Meta P' relates to p-values for these probes from a recent meta-analysis of AD pathology (R. Smith at al., 2021). The 'Novel Gene' column indicates if the gene has previously been implicated in neurodegenerative disease and the 'Previously Identified in' column indicates how it was previously implicated in dementia.*

DNAm site	Beta (%)	SE (%)	P	Chr	BP	Gene	Gene Region	Array	Braak meta P	Interaction BR P	Novel Gene	Previously Identified in
cg11658414	0.30	4.47e-02	9.11e-09	20	1450828	-		EPIC	-	1.58e-01	-	-
cg13515047	0.31	4.88e-02	6.04e-08	16	75298429	*BCAR1*	5'UTR	EPIC	-	1.73e-02	novel	-

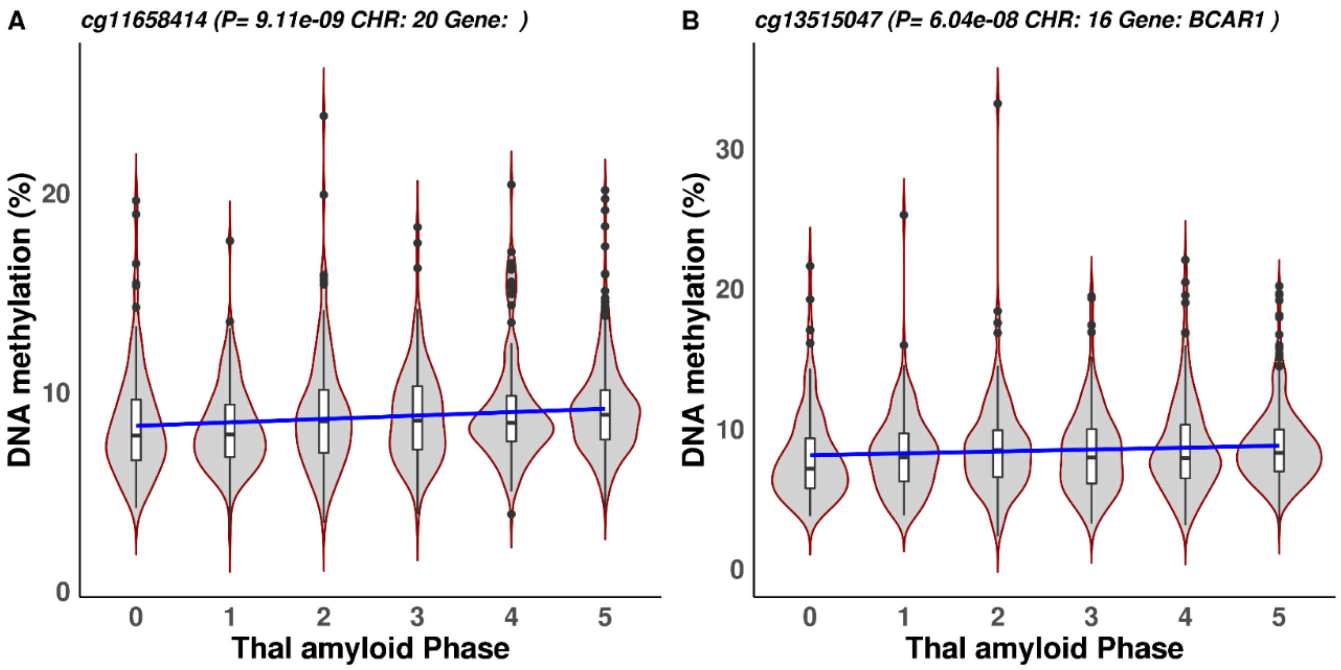


Figure 4.20: Differential methylation across the Thal Phases. *Violin plots for the DNA methylation values (adjusted for batch and covariates) across the 6 stages are shown, where the box in the middle represents the interquartile range (IQR), whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). The grey on the outside of the box plot is a density plot of the distribution of data, where the width represents frequency. The x axis is Thal phase. The y-axis is the methylation value as a percentage.*

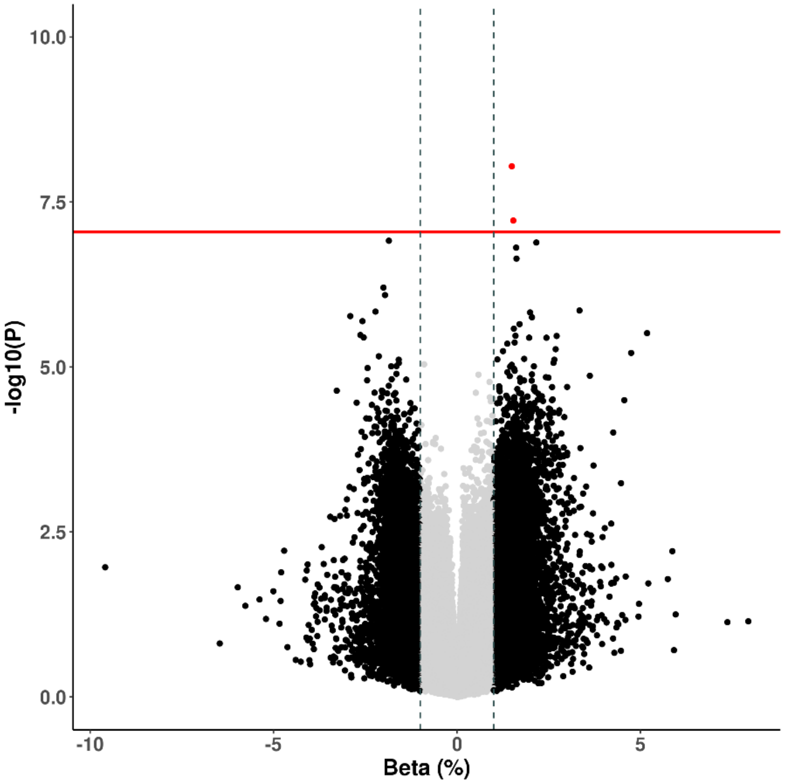


Figure 4.21: Volcano plot of differentially methylated positions (DMPs) identified in the Thal Phase EWAS. *The x-axis shows beta effect size (ES) and the Y-axis shows -log10(p). Black probes indicate an ES difference ≥ 0.01, whilst red probes indicate an ES difference ≥ 0.01 and a p-value that reaches experiment-wide significance (EWS) (p < 9e-08).*

4.4.6.1 Multiple DMRs associated with Thal Phase

I identified ten DMRs associated with Thal phase, with eight of these being annotated to genes (see **Table 4.12**). Example plots of DMRs comprising of >5 DNAm sites are shown in **Figure 4.28**. Interestingly several of DMRs overlapped with regions/sites identified in the Braak NFT EWAS and DMR analysis including sites annotated to *KLHL33*, *ARHGEF17*, *CCDC42B*, *HOXA*3, *AGAP2* and *SH2B2***.** *AGAP2* (ArfGAP with GTPase domain, ankyrin repeat and PH domain 2) is enriched in microglia, neurons and astrocytes and there is evidence for promoter DNA hypermethylation of this gene in AD (Liu, Wang, Marcora, Zhang, & Goate, 2019). An example DMR plot around *AGAP2* is shown in **Figure 4.23.**

4.4.6.2 Interaction effects between brain region and Thal Phase

I looked at interactions between brain region and Thal phase to identify if there were any associations of differing magnitude or direction of effect in either the PFC or OCC. three DMPs reached experiment wide significance for an interaction effect (p<9e-08) and two of these were annotated to genes. The most significant DMP was cg15086994 (chromosome 3, annotated to *CHCHD6*; p=2.04e-09). This suggests there are differences between the two brain regions where DNAm decreases at a steeper gradient in the OCC than the PFC as Thal Phase increases (see **Figure 4.18**)**.** *CHCHD6* is a member of the coiled-coil-helix-coiled-coil-helix domain (CHCHD)-containing proteins which encode small mitochondrial proteins (Zhou, Saw, & Tan, 2017). Several different CHCHD proteins have been linked to PD, ALS and FTD (Bannwarth *at al.*, 2014; Funayama *at al.*, 2015). The second most significant DMP with an interaction effect was cg19643390 (located on chromosome 17; p=5.26e-08) and is annotated to *ABCC3*. There is evidence for hypermethylation in the OCC with elevated braak but not in the PFC at this site (see **Figure 4.24**). *ABCC3* is an ABC transporter which is expressed in astrocytes and microglia and has been implicated in Alzheimer's disease (Pereira, Martins, Wiltfang, da Cruz E Silva, & Rebelo, 2018). At a more relaxed threshold of p<5e-5 there were 149 hits (see **Table 4.13** for the top 20 interaction associations).

Table 4.12: Differentially methylated regions (DMRs) associated with Thal Phase. *In total ten DMRs were identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected p-value adjusted for the number of independent tests.*

Chr	BP start	BP end	n	Gene	Beta (%)	SE (%)	P	P adjusted
14	20903445	20904169	10	*KLHL33*	5.27	0.80	5.20e-11	4.37e-05
2	220298547	220299584	7	*SPEG*	5.36	0.86	4.93e-10	4.14e-04
5	1385632	1386647	7	-	2.87	0.47	6.80e-10	5.71e-04
11	73053830	73054185	4	*ARHGEF17*	8.07	1.32	1.04e-09	8.72e-04
7	27153580	27153847	6	*HOXA3*	8.59	1.41	1.13e-09	9.50e-04
15	74494909	74495384	6	*STRA6*	5.44	0.94	7.49e-09	6.29e-03
12	58131345	58132105	6	*AGAP2*	7.89	1.38	1.10e-08	9.21e-03
12	113587240	113587690	7	*CCDC42B*	5.49	0.96	1.23e-08	1.03e-02
7	101961796	101962123	3	*SH2B2*	5.74	1.04	3.30e-08	2.77e-02
11	10920810	10921186	3	-	-4.13	0.75	3.46e-08	2.91e-02

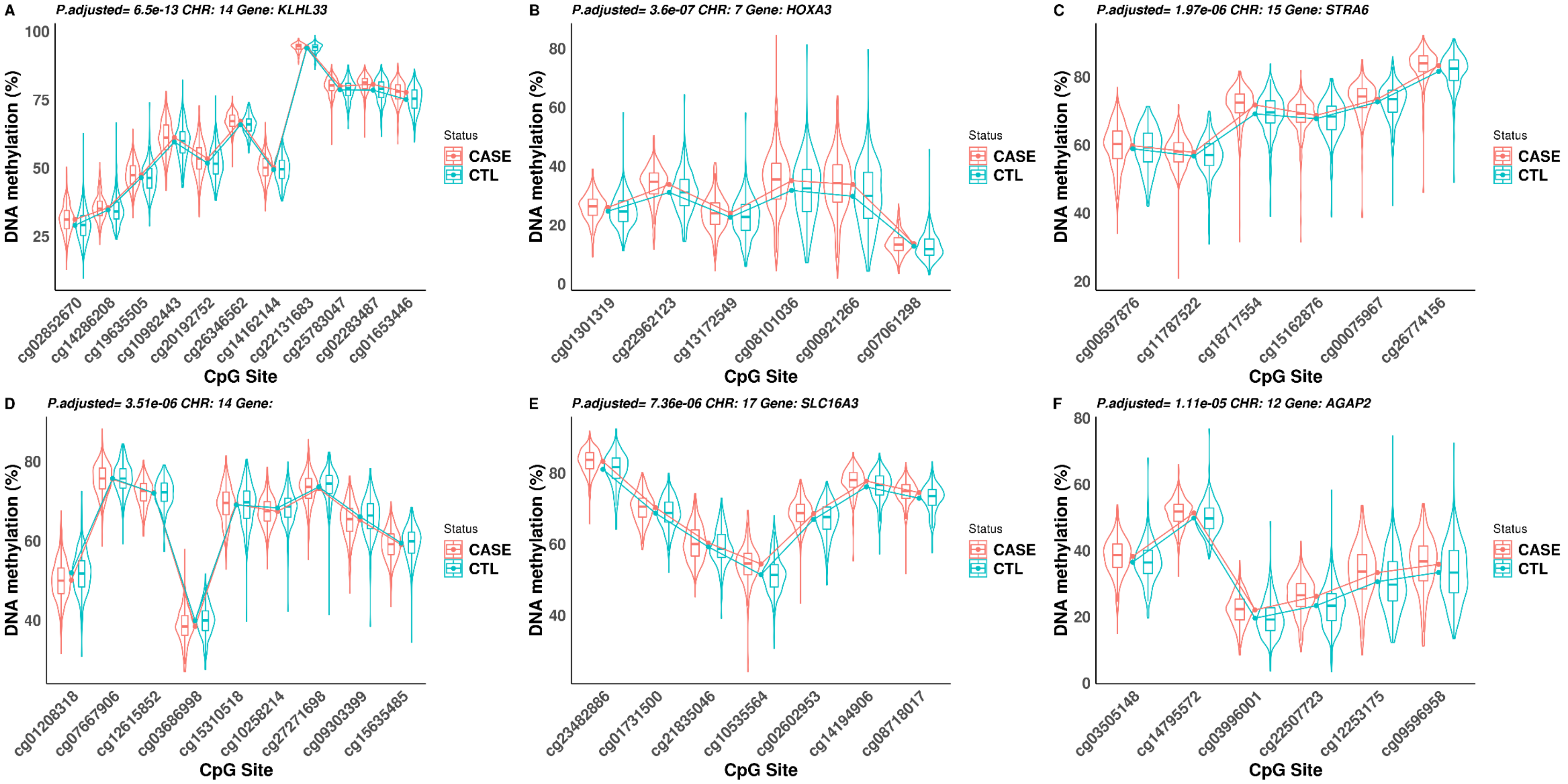


Figure 4.22: Differentially methylated regions (DMRs) associated with Thal phase. *An example of six DMRs which were associated with Thal phase are shown. The x-axis are the DNAm sites ordered by genomic location (not to scale). The y-axis is DNA methylation as a percentage. Violin plots for the DNA methylation values (adjusted for batch and covariates) across at each DNAm site in the DMR is shown, where the box in the middle represents the interquartile range (IQR), whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). Samples were split into cases (Thal phase > 3) and controls (Thal phase < 2) in order to visualise the differences between the groups. The grey on the outside of the box plot is a density plot of the distribution of data, where the width represents frequency.*

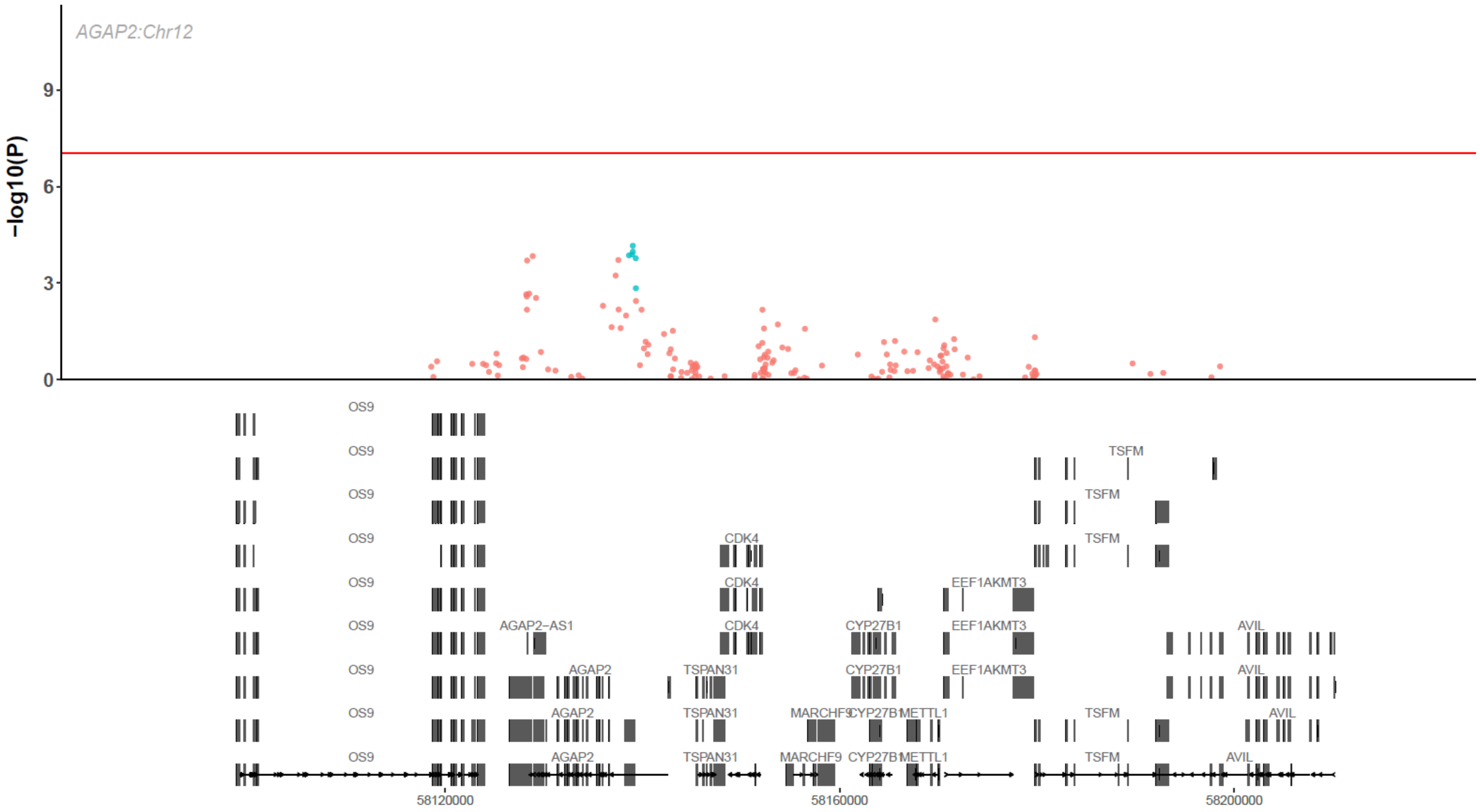


Figure 4.23: A differentially methylated region identified in the AGAP2 gene within chromosome 12 was associated with Thal Phase. *Shown on the top is a zoomed in Manhattan plot around the AGAP2 region, where the x-axis represents the base position and the y-axis represents the –log10 p-value and each point on the plot represents a DNA methylation site. The red horizontal line represents experiment wide significance (P< 9e-08). The gene track shows the locations of the genes within in this region*

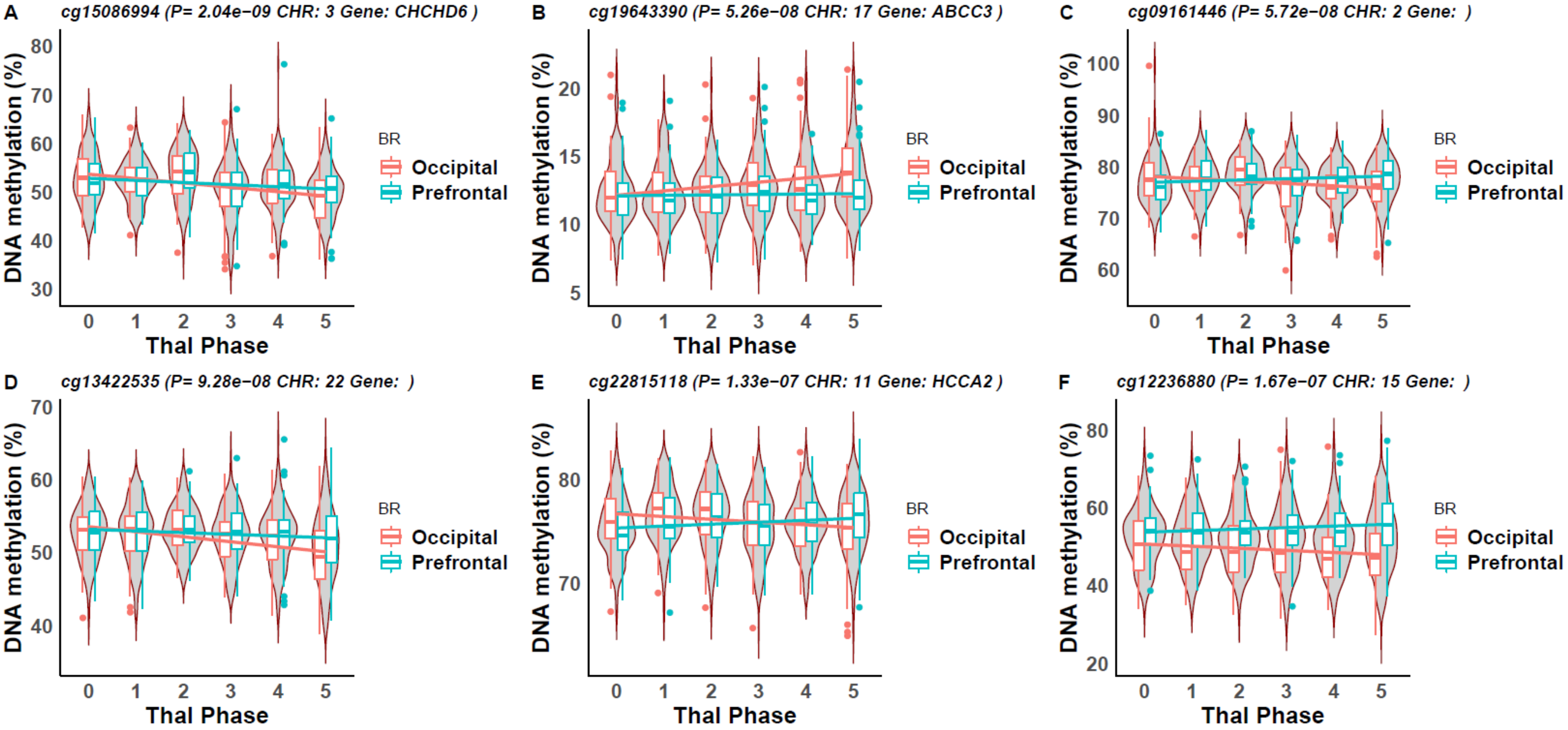


Figure 4.24: Interaction between Thal phase and brain region. Shown are the interaction results for the top 6 DNAm sites from the interaction EWAS, where an interaction term between Thal phase and brain region was included. *The plots are for the six top differentially methylated positions with the strongest beta values. Violin plots for the DNA methylation values (adjusted for batch and covariates) across the 6 stages are shown, where the boxes in the middle represents the interquartile range (IQR) for each brain region, whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). The grey on the outside of the box plot is a density plot of the distribution of data (not split by brain region), where the width represents frequency. The x-axis is Thal phase. The y-axis is the methylation value as a percentage.*

Table 4.13: The top 20 differentially methylated positions identified in the interaction EWAS between brain region and Thal Phase. *Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. Shown for each DMP is the effect size (Beta), standard error (SE) and the p-value (P) for the effect of Braak NFT Stage on DNAm in addition to the interaction effect (Beta Int, SE Int and P Int) between Thal phase and brain region (occipital cortex [reference] and prefrontal cortex).*

DNAm site	**Beta (%)**	**SE (%)**	**P**	**Beta Int (%)**	**SE Int (%)**	**P Int**	**CHR**	**MAPINFO**	**Gene**	**Ref Group**
cg15086994	-0.45	9.79e-02	4.67e-05	0.49	7.98e-02	2.04e-09	3	126440724	*CHCHD6*	Body
cg19643390	0.21	5.43e-02	6.58e-04	-0.29	5.18e-02	5.26e-08	17	48712205	*ABCC3*	TSS200
cg09161446	-0.30	8.55e-02	1.89e-03	0.49	8.96e-02	5.72e-08	2	240697468	-	-
cg13422535	-0.30	8.24e-02	1.31e-03	0.38	7.04e-02	9.28e-08	22	47077682	-	-
cg22815118	-0.20	6.45e-02	6.49e-03	0.35	6.62e-02	1.33e-07	11	1529654	*HCCA2*	Body
cg12236880	-0.22	9.26e-02	3.26e-02	0.56	1.06e-01	1.67e-07	15	68564635	-	-
cg22829375	-0.20	7.72e-02	2.00e-02	0.38	7.46e-02	4.68e-07	2	233764831	*NGEF*	Body
cg09156067	0.10	2.52e-02	6.98e-04	-0.16	3.11e-02	5.30e-07	15	90645907	*IDH2*	TSS200
cg04754212	-0.32	8.54e-02	1.02e-03	0.42	8.35e-02	5.43e-07	10	127506329	*UROS*	5'UTR
cg01799862	0.32	7.68e-02	2.95e-04	-0.30	6.13e-02	1.06e-06	16	67143802	*C16orf70*	TSS200
cg09781115	0.14	5.67e-02	2.54e-02	-0.36	7.33e-02	1.12e-06	10	99127233	*RRP12*	Body
cg15604993	-0.27	9.27e-02	9.68e-03	0.45	9.35e-02	2.14e-06	15	52549248	*MYO5C*	Body
cg09796739	-0.28	6.88e-02	3.65e-04	0.43	8.95e-02	2.29e-06	1	227924055	*JMJD4;SNAP47*	TSS1500;Body
cg09016797	0.33	1.06e-01	5.84e-03	-0.66	1.39e-01	2.39e-06	7	155790956	-	-
cg02171258	-0.46	1.10e-01	1.97e-04	0.57	1.21e-01	2.65e-06	6	116442223	*COL10A1;NT5DC1*	Body
cg22175345	-0.32	9.44e-02	2.35e-03	0.35	7.43e-02	2.74e-06	4	17194180	-	-
cg17944001	-0.22	7.81e-02	1.19e-02	0.38	8.03e-02	2.75e-06	1	15610717	*FHAD1*	Body
cg26297203	-0.27	7.58e-02	1.49e-03	0.33	6.86e-02	2.88e-06	2	52796128	-	-
cg22192710	0.26	9.84e-02	1.94e-02	-0.59	1.26e-01	2.92e-06	11	19363946	-	-
cg18113453	-0.35	8.61e-02	2.95e-04	0.52	1.10e-01	3.23e-06	6	33289660	*DAXX*	Body

4.4.7 CERAD density-associated DNA methylation signatures across two cortical brain regions

14 DMPs were associated with CERAD density (see **Figure 4.25** and **Table 4.14**). The average magnitude of effect for the significant DMPs per unit of CERAD density was 0.57% (inter-quartile range [IQR] = 0.48-0.63%) and therefore there was a total mean DNAm change of 2.16% from low CERAD score to high CERAD score. Examples of DMP plots across the different CERAD densities are shown in **Figure 4.13**. The effects were consistent across the PFC and OCC as shown by the non-significant interaction term between CERAD density and brain region (p>9e-08; see **Table 4.14**).

Ten (71%) of the DMPs were significantly hypermethylated (i.e. increased methylation was associated with increasing CERAD density; see **Figure 4.26**). Of the 14, 12 were annotated to unique genes and three of these genes have not previously been implicated in neurodegenerative disease (*BCAR1*, *OSCAR* and *SERP2*), although *SERP2* had been reported to have suggestive significance (p< 5e-05) in the latest AD-EWAS (R. Smith *at al.*, 2021). The same DMP which is annotated to *BCAR1* - cg13515047 – was significantly hypermethylated in both the Thal EWAS (p=6.04e-08) and the CERAD density EWAS (p=4.96e-09). A brief description of this gene is found in the previous results section (see **4.4.6**). The DMP cg05606351 was significantly hypermethylated with elevated CERAD density (p=2.03e-08) and is annotated to a gene which has not previously been implicated in dementia: *OSCAR* (located on chromosome 19). *OSCAR* (Osteoclast-associated immunoglobulin-like receptor) is classified in a group of cells that reabsorb bone and control bone homeostasis and is a member of the leukocyte receptor complex protein family regulates both innate and adaptive immune responses (Kim *at al.*, 2005). In contrast, many of the DMPs are annotated to genes that have been previously implicated in AD, with the majority of sites having been identified in the recent AD EWAS meta-analysis performed by our group (see **Table 4.14**)(R. Smith *at al.*, 2021).

4.4.7.1 Multiple DMRs associated with CERAD density

I identified 53 DMRs associated with CERAD density, with 50 of these being annotated to genes (see **Table 4.14).** Several of the DMRs overlapped with those identified in the Braak NFT and Thal EWAS and DMR analysis, suggesting they are not unique to specific neuropathology measures. These include regions within *KLHL33*, *LLGL2, ARHGEF17*, *HOXA*3, *ATG16L2* and *SLC16A3* (the top 6 DMRs). Several genes were found to harbour multiple DMRs including *HOXA*3, where 2 DMRs were identified. *SLC16A3* has been identified as an AD risk gene in previous EWAS studies (De Jager *at al.*, 2014; Q. S. Li *at al.*, 2020). An example plot of the *SLC16A3* DMR is show in **Figure 4.29**.

4.4.7.2 Interaction effects between brain region and CERAD density

I looked at interactions between brain region and CERAD density to identify if there were any associations of differing magnitude or direction of effect in either the PFC or OCC. No genome-wide DMPs reached experiment wide significance for an interaction effect (p>9e-08). The lack of significant interaction effects suggest the majority of associations between CERAD and DNAm are consistent across the cortex. However, at a more relaxed threshold of p<5e-5 there was evidence of an interaction at 21 sites. The top six DMPs are illustrated in **Figure 4.30** (see **Table 4.16** for the top 20 interaction associations). Several of top the associations were also identified in the Braak NFT stage interaction model, including the same top DMP - cg15984835.

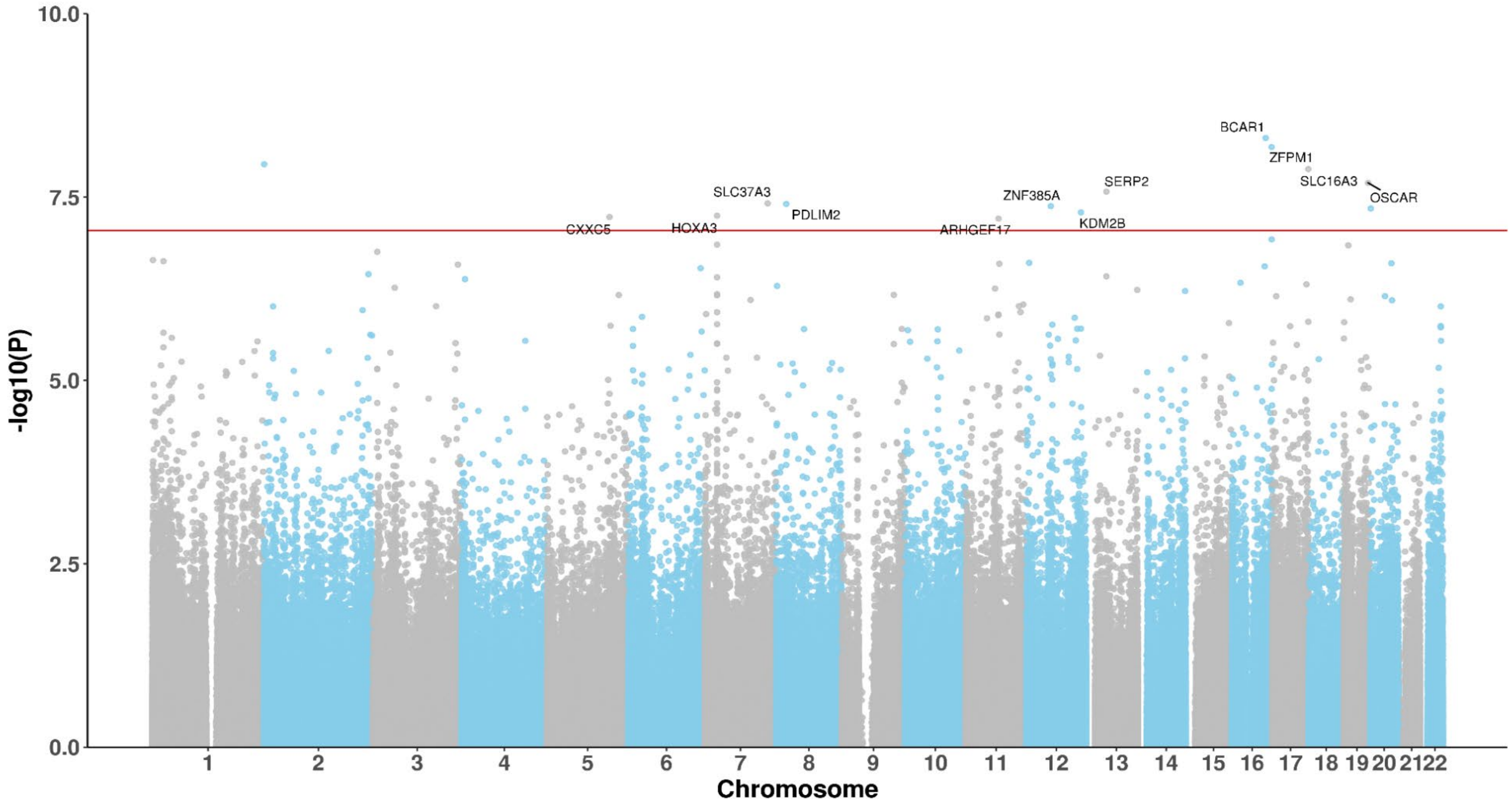


Figure 4.25: Cortex EWAS of CERAD density stage highlights experiment-wide significant differentially methylated positions. *A Manhattan plot showing results of a CERAD density EWAS across two cortical regions (prefrontal cortex and occipital cortex). The significant differentially methylated positions are annotated with their Illumina UCSC gene name, unless they are unannotated to a gene. The x-axis shows chromosomes 1-22 and the y-axis shows -log10(P), with the horizontal red line representing experiment wide significance (p< 9e-8).*

Table 4.14: Differentially methylated positions (DMPs) associated with CERAD density at an experiment wide significance threshold (p<9e-08). *In total 14 DMPs were identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. The Beta columns refers to %methylation change per unit increase in CERAD density. The 'Array' column states if probes are also present on the Illumina 450K array. The 'Braak Meta P' relates to p-values for these probes from a recent meta-analysis of AD pathology (R. Smith at al., 2021). The 'Novel Gene' column indicates if the gene has previously been implicated in neurodegenerative disease and the 'Previously Identified in' column indicates how it was previously implicated in dementia.*

DNAm site	**Beta (%)**	**SE (%)**	**P**	**Chr**	**BP**	**Gene**	**Gene Region**	**Array**	**Braak meta P**	**Interaction BR P**	**Novel Gene**	**Previously Identified in**
cg13515047	0.44	6.13e-02	4.96e-09	16	75298429	*BCAR1*	5'UTR	EPIC	-	2.54e-01	novel	-
cg06913337	-0.94	1.31e-01	6.58e-09	16	88590404	*ZFPM1*	Body	450K & EPIC	2.79e-01	2.49e-01	previous	AD_EWAS
cg08813888	-0.53	7.56e-02	1.13e-08	2	662396	-	-	450K & EPIC	8.85e-07	6.03e-02	-	-
cg10208942	0.56	7.91e-02	1.31e-08	17	80192754	*SLC16A3*	5'UTR	EPIC	-	4.02e-01	previous	AD_EWAS
cg05606351	0.64	9.19e-02	2.03e-08	19	54599285	*OSCAR*	Body	450K & EPIC	1.95e-03	4.46e-01	novel	-
cg16021126	0.39	5.70e-02	2.67e-08	13	44947611	*SERP2*	TSS1500	450K & EPIC	1.10e-05	5.24e-01	novel	-
cg15751131	-0.67	9.89e-02	3.86e-08	7	140090416	*SLC37A3*	5'UTR	450K & EPIC	3.12e-08	8.45e-01	previous	AD_EWAS
cg18100976	0.58	8.50e-02	3.93e-08	8	22446737	*PDLIM2*	Body	EPIC	-	6.23e-01	previous	AD_EWAS
cg07332724	0.50	7.32e-02	4.21e-08	12	54773114	*ZNF385A;LOC102724050*	Body	EPIC	-	9.63e-01	previous	AD_EWAS
cg11658414	0.38	5.56e-02	4.53e-08	20	1450828	-	-	EPIC	-	1.29e-01	-	-
cg11724984	0.63	9.29e-02	5.12e-08	12	121890864	*KDM2B*	Body	450K & EPIC	3.27e-09	7.27e-02	previous	AD_EWAS
cg07061298	0.52	7.77e-02	5.68e-08	7	27153847	*HOXA3*	5'UTR	450K & EPIC	4.57e-13	5.70e-01	previous	AD_EWAS
cg14871225	-0.82	1.22e-01	5.91e-08	5	139040820	*CXXC5*	5'UTR	450K & EPIC	1.32e-07	3.51e-01	previous	AD_EWAS
cg02776498	0.47	7.09e-02	6.22e-08	11	73054185	*ARHGEF17*	Body	450K & EPIC	1.55e-07	8.04e-02	previous	AD_EWAS

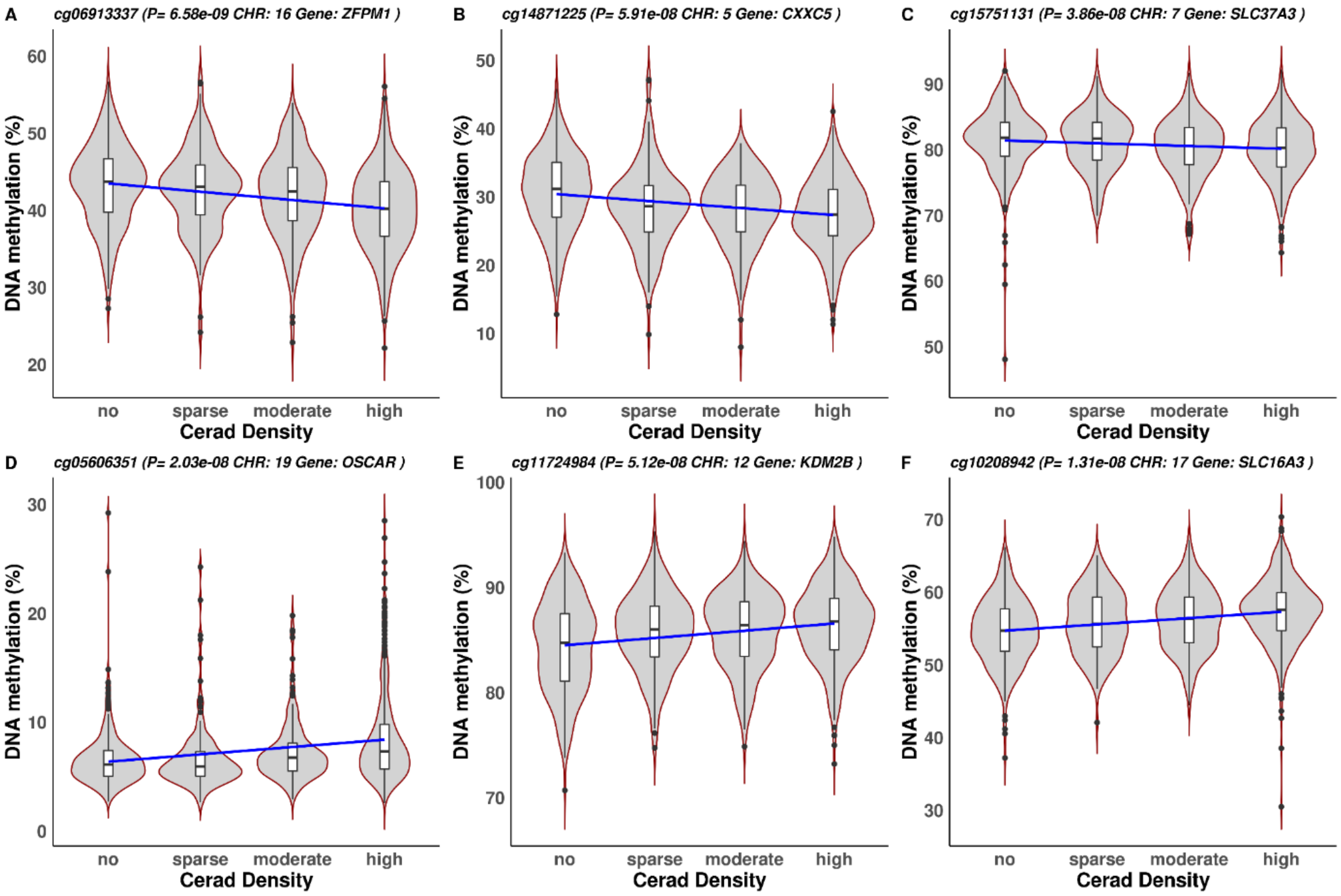

A
cg06913337 (P= 6.58e-09 CHR: 16 Gene: ZFPM1)
DNA methylation (%)
20
30
40
50
60
no
sparse
moderate
high
Cerad Density
B
cg14871225 (P= 5.91e-08 CHR: 5 Gene: CXXC5)
DNA methylation (%)
10
20
30
40
50
no
sparse
moderate
high
Cerad Density
C
cg15751131 (P= 3.86e-08 CHR: 7 Gene: SLC37A3)
DNA methylation (%)
50
60
70
80
90
no
sparse
moderate
high
Cerad Density
D
cg05606351 (P= 2.03e-08 CHR: 19 Gene: OSCAR)
DNA methylation (%)
0
10
20
30
no
sparse
moderate
high
Cerad Density
E
cg11724984 (P= 5.12e-08 CHR: 12 Gene: KDM2B)
DNA methylation (%)
70
80
90
100
no
sparse
moderate
high
Cerad Density
F
cg10208942 (P= 1.31e-08 CHR: 17 Gene: SLC16A3)
DNA methylation (%)
30
40
50
60
70
no
sparse
moderate
high
Cerad Density

Figure 4.26: Differential methylation for CERAD density for six of the top differentially methylated positions identified by EWAS. *The plots are for the six top differentially methylated positions with the strongest beta values. Violin plots for the DNA methylation values (adjusted for batch and covariates) across the 4 stages are shown, where the box in the middle represents the interquartile range (IQR), whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). The grey on the outside of the box plot is a density plot of the distribution of data, where the width represents frequency. The x axis is CERAD density. The y-axis is the methylation value as a percentage.*

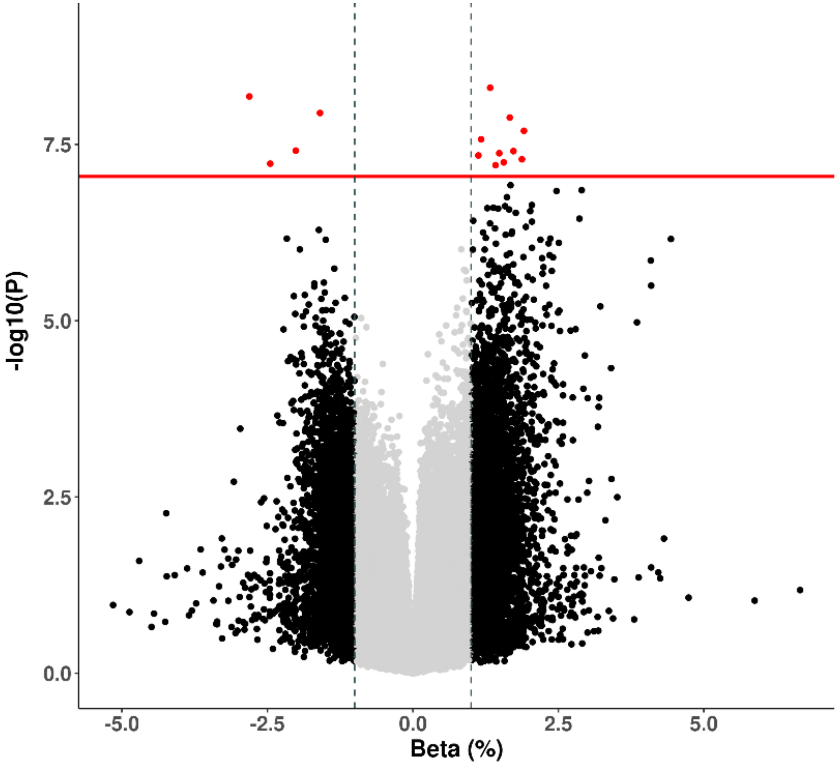


Figure 4.27: Volcano plot of differentially methylated positions (DMPs) identified in the CERAD density epigenome-wide association study. *The x-axis shows beta effect size (ES) and the y-axis shows -log10(p). Black probes indicate an ES difference ≥ 0.01, whilst red probes indicate an ES difference ≥ 0.01 and a p-value that reaches experiment-wide significance (EWS) (p < 9e-08).*

Table 4.15: Differentially methylated regions (DMRs) associated with CERAD density. *In total 51 DMRs were identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected p-value adjusted for the number of independent tests.*

Chr	BP start	BP end	n	Gene	Beta (%)	SE (%)	P	P adjusted
14	20903410	20904169	11	*KLHL33*	8.96	1.01	7.62e-19	6.50E-13
11	73053649	73054185	5	*ARHGEF17*	11.38	1.47	1.01e-14	8.62E-09
7	27153580	27153847	6	*HOXA3*	12.83	1.77	4.22e-13	3.60E-07
11	72533202	72533664	4	*ATG16L2*	13.27	1.86	1.12e-12	9.54E-07
12	6054616	6054811	5	*ANO2*	-9.04	1.28	1.83e-12	1.56E-06
15	74494909	74495384	6	*STRA6*	8.3	1.18	2.31e-12	1.97E-06
14	106329652	106331018	9	-	-6.71	0.97	4.11e-12	3.51E-06
12	113587513	113587690	4	*CCDC42B*	9.39	1.36	4.31e-12	3.68E-06
11	10920810	10921186	3	-	-6.54	0.95	4.72e-12	4.03E-06
17	80192161	80192655	7	*SLC16A3*	9.04	1.32	8.62e-12	7.36E-06
12	58131345	58132105	6	*AGAP2*	11.7	1.73	1.29e-11	1.11E-05
2	220298547	220299584	7	*SPEG*	7.35	1.09	1.43e-11	1.22E-05
15	77324526	77324758	4	*PSTPIP1*	7.8	1.16	1.51e-11	1.29E-05
7	27185136	27185512	3	*HOXA6*	9.36	1.39	1.55e-11	1.32E-05
11	74870268	74870640	3	*SLCO2B1*	10.26	1.59	1.13e-10	9.69E-05
1	43814306	43814983	4	*MPL*	10.96	1.71	1.48e-10	1.26E-04
3	9988590	9988889	3	*PRRT3;PRRT3-AS1*	11.32	1.77	1.80e-10	1.54E-04
6	168079212	168079818	4	-	-7.33	1.17	3.30e-10	2.82E-04
3	50359849	50360690	9	*HYAL2*	4.04	0.64	3.37e-10	2.88E-04

19	46999109	46999307	6	*PNMAL2*	9.12	1.47	4.98e-10	4.25E-04
9	35563884	35564160	5	*FAM166B*	6.38	1.03	5.81e-10	4.96E-04
22	46451281	46451518	3	*LOC150381;C22orf26*	10.9	1.77	6.72e-10	5.74E-04
16	10671842	10672543	5	*EMP2*	5.01	0.83	1.87e-09	1.60E-03
14	77542656	77543192	5	*LOC102724190*	6.43	1.08	2.32e-09	1.98E-03
10	46971347	46971908	7	*SYT15*	5.41	0.91	2.40e-09	2.05E-03
6	31868949	31869565	12	*ZBTB12*	4.97	0.84	3.17e-09	2.71E-03
19	1071020	1071208	3	*HMHA1*	11.07	1.92	7.80e-09	6.66E-03
19	8454867	8455567	11	*RAB11B*	4.78	0.83	8.24e-09	7.04E-03
19	2096331	2097096	12	*C19orf36;MOBKL2A*	6	1.04	9.15e-09	7.81E-03
17	74475270	74475402	4	*RHBDF2*	13.14	2.29	9.95e-09	8.50E-03
20	32319719	32320193	3	*ZNF341*	6.95	1.21	1.03e-08	8.84E-03
5	1523874	1524249	4	*LPCAT1*	5.61	0.98	1.11e-08	9.45E-03
15	75117658	75118427	4	*CPLX3*	3.85	0.67	1.11e-08	9.50E-03
6	112042339	112042632	3	*FYN*	-4.48	0.79	1.34e-08	0.01
13	113676829	113677188	4	*MCF2L*	5.68	1	1.46e-08	0.01
9	35538890	35539007	3	*RUSC2*	9.47	1.68	1.79e-08	0.02
12	123215248	123215684	8	*GPR81*	4.29	0.77	2.09e-08	0.02
21	33867372	33867551	3	*C21orf63*	7.64	1.36	2.12e-08	0.02
12	50497589	50498041	5	*GPD1*	6.28	1.13	2.80e-08	0.02
6	30853959	30854109	3	*DDR1*	8.19	1.48	2.91e-08	0.02
17	3848156	3848324	3	*ATP2A3*	9.9	1.79	2.91e-08	0.02

1	164545553	164545843	4	*PBX1*	9.01	1.63	3.43e-08	0.03
6	33047944	33048752	14	*HLA-DPB1*	7.96	1.45	3.77e-08	0.03
5	126205009	126205081	3	*Mar-03*	-8.52	1.55	3.86e-08	0.03
14	74815131	74815316	4	*C14orf115*	8	1.45	3.90e-08	0.03
11	67219631	67220043	5	*GPR152*	4.22	0.77	4.00e-08	0.03
1	167090618	167090757	3	*DUSP27*	-11.68	2.13	4.07e-08	0.03
6	158438072	158438419	7	*SYNJ2*	7.63	1.39	4.23e-08	0.04
6	32120773	32120826	3	*PPT2;PRRT1*	8.55	1.56	4.56e-08	0.04
19	37825309	37825679	8	*HKR1*	7.99	1.47	5.11e-08	0.04
7	27155036	27155358	5	*HOXA3*	9.66	1.78	5.33e-08	0.05
11	560921	560951	4	*RASSF7;C11orf35*	8.88	1.63	5.49e-08	0.05
6	33132181	33132442	5	*COL11A2*	6.44	1.19	5.72e-08	0.05

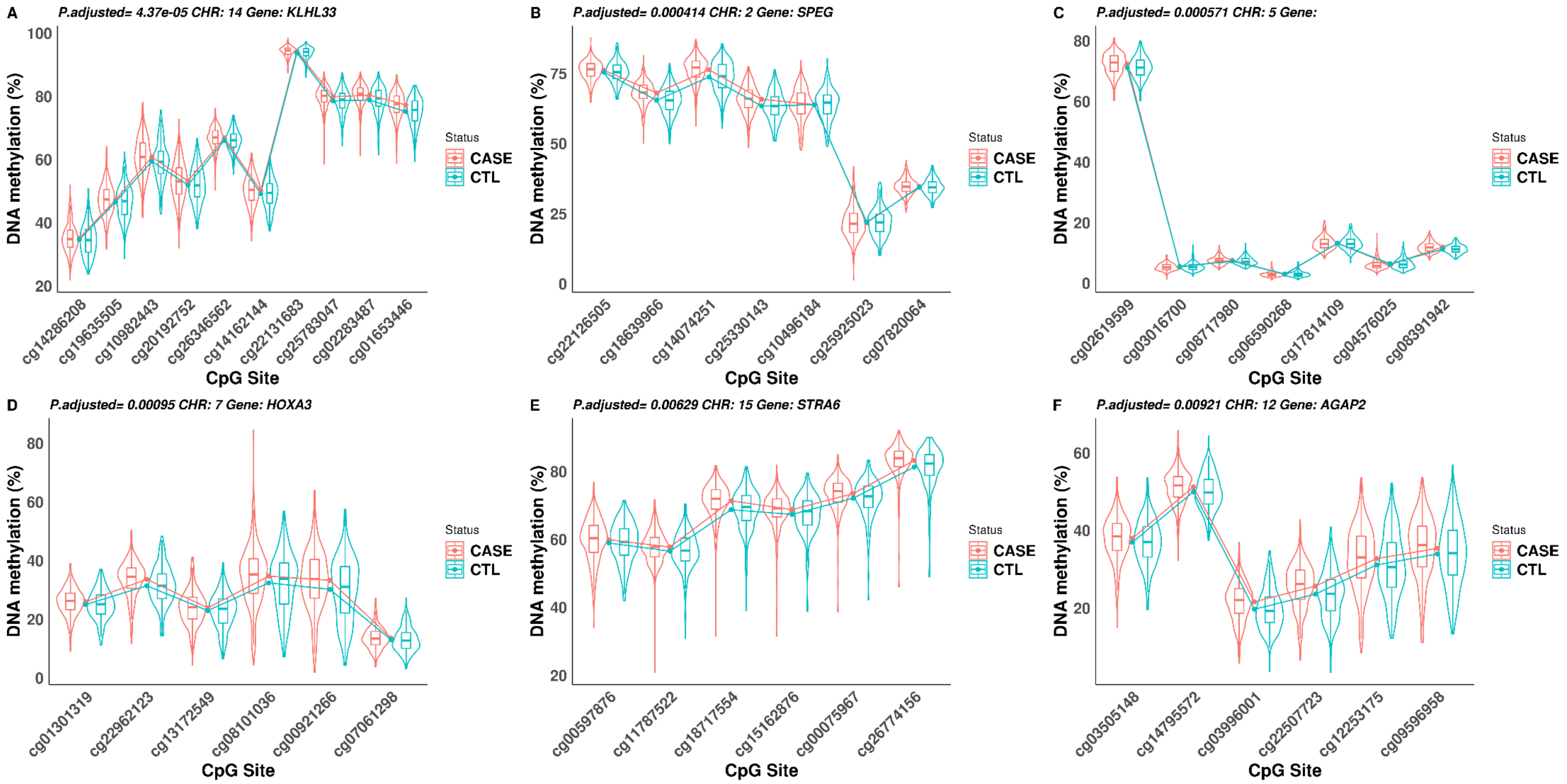


Figure 4.28: Differentially methylated regions (DMRs) associated with CERAD density. *The top six most significant DMRs associated CERAD density are shown. The x-axis are the DNAm sites ordered by genomic location. The y-axis is DNA methylation as a percentage. Violin plots for the DNA methylation values (adjusted for batch and covariates) across at each DNAm site in the DMR is shown, where the box in the middle represents the interquartile range (IQR), whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). Samples were split into cases (CERAD density > 2) and controls (CERAD density < 2) in order to visualise the differences between the groups. The grey on the outside of the box plot is a density plot of the distribution of data, where the width represents frequency.*

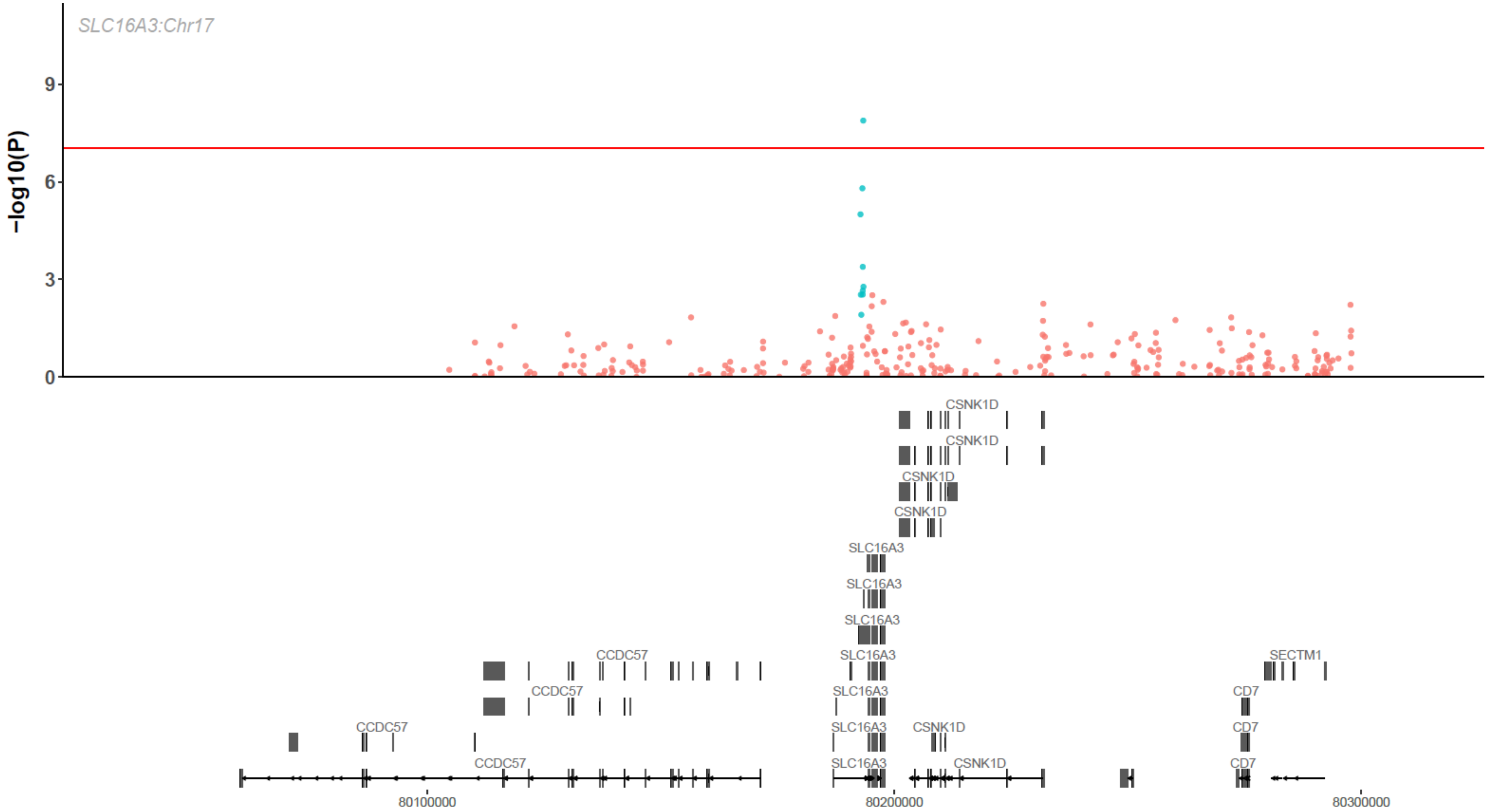


Figure 4.29: A differentially methylated region identified in the SLC16A3 gene within chromosome 17 was associated with CERAD density. *Shown on the top is a zoomed in Manhattan plot around the SLC16A3 region, where the x-axis represents the base position and the y-axis represents the –log10 p-value and each point on the plot represents a DNA methylation site. The red horizontal line represents experiment wide significance (P< 9e-08). The gene track shows the locations of the genes within in this region.*

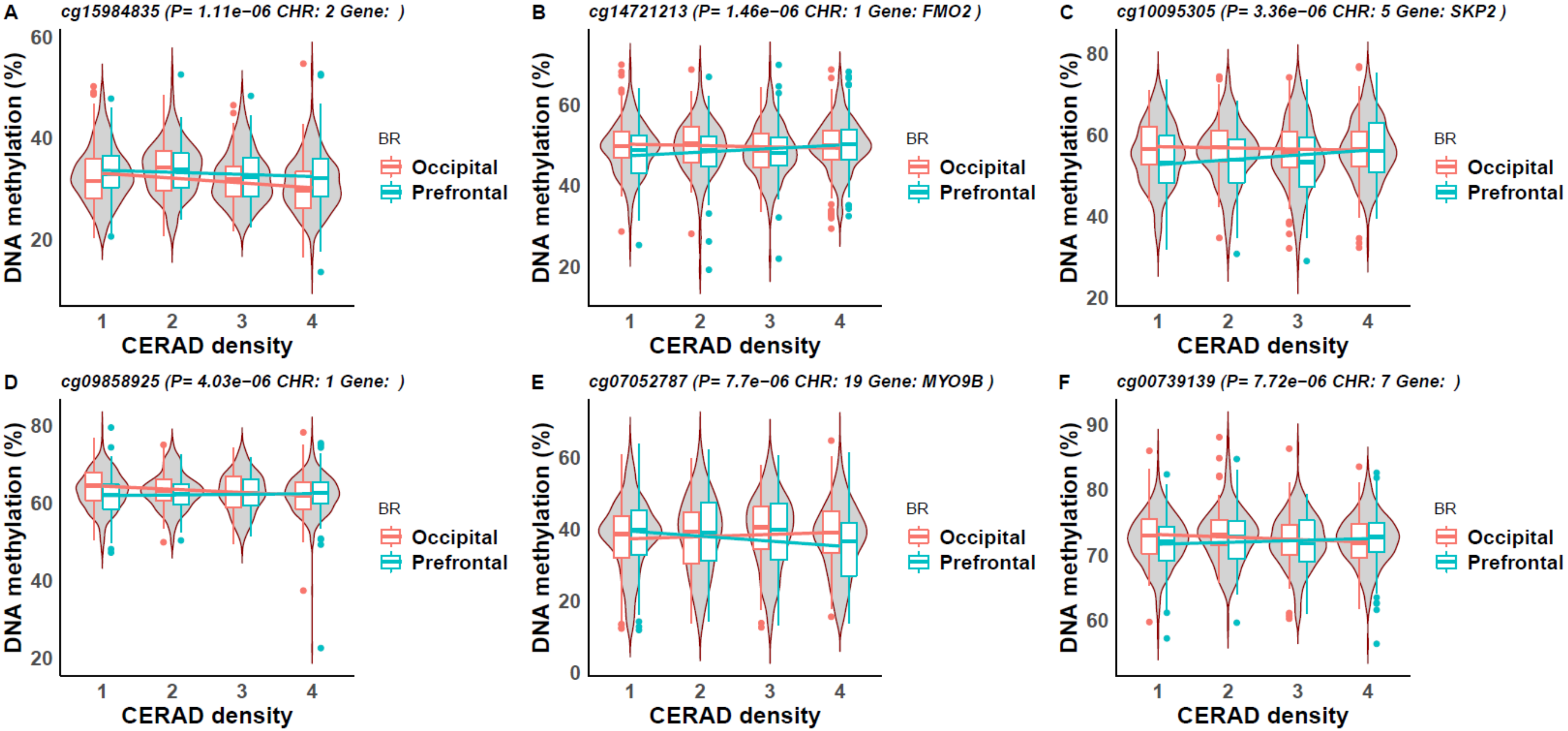


Figure 4.30: Interaction between CERAD density and brain region. *Shown are the interaction results for the top 6 DNAm sites from the interaction EWAS, where an interaction term between CERAD density and brain region was included. The plots are for the six top differentially methylated positions with the strongest beta values. Violin plots for the DNA methylation values (adjusted for batch and covariates) across the 4 stages are shown, where the boxes in the middle represents the interquartile range (IQR) for each brain region, whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). The grey on the outside of the box plot is a density plot of the distribution of data (not split by brain region), where the width represents frequency. The x-axis is cerad density (1= no/ none; 2= sparse; 3= moderate and 4= frequent/ high). The y-axis is the methylation value as a percentage.*

Table 4.16: The top 20 differentially methylated positions identified in the interaction EWAS between brain region and CERAD density. *Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. Shown for each DMP is the effect size (Beta), standard error (SE) and the p-value (P) for the effect of CERAD density on DNAm in addition to the interaction effect (Beta Int, SE Int and P Int) between CERAD density and brain region (occipital cortex [reference] and prefrontal cortex).*

DNAm site	Beta (%)	SE (%)	P	Beta Int (%)	SE Int (%)	P Int	CHR	MAPINFO	Gene	Ref Group
cg15984835	-0.59	0.09	1.75e-07	0.64	0.13	1.11e-06	2	113444352	-	-
cg14721213	-0.36	0.13	2.12e-02	0.90	0.19	1.46e-06	1	171154612	*FMO2*	5'UTR
cg10095305	-0.37	0.14	2.24e-02	0.89	0.19	3.36e-06	5	36158217	*SKP2*	5'UTR;Body
cg09858925	-0.45	0.10	2.95e-04	0.68	0.15	4.03e-06	1	209537469	-	-
cg07052787	0.43	0.13	9.28e-03	-0.81	0.18	7.70e-06	19	17237120	*MYO9B*	Body
cg00739139	-0.30	0.08	2.41e-03	0.52	0.12	7.72e-06	7	65233991	-	-
cg20030796	-0.40	0.10	9.40e-04	0.63	0.14	9.73e-06	17	72356254	*BTBD17*	Body
cg03546163	0.31	0.10	1.32e-02	-0.60	0.14	1.22e-05	6	35654363	*FKBP5*	5'UTR
cg13129591	0.45	0.11	8.85e-04	-0.66	0.15	1.47e-05	3	152939446	-	-
cg13212668	0.26	0.06	2.36e-04	-0.33	0.08	1.61e-05	20	59928453	*CDH4*	Body
cg07305369	-0.31	0.08	1.67e-03	0.49	0.11	1.88e-05	11	64270338	-	-
cg12148585	-0.27	0.08	4.06e-03	0.48	0.11	2.22e-05	2	208006420	*KLF7*	Body
cg00184799	0.26	0.09	3.29e-02	-0.56	0.13	2.31e-05	8	38722544	-	-
cg02320784	-0.31	0.13	3.51e-02	0.75	0.18	2.40e-05	12	26195988	*RASSF8*	5'UTR
cg05211470	0.17	0.06	3.30e-02	-0.36	0.08	2.74e-05	5	114909169	-	-
cg18341924	0.04	0.01	2.07e-02	-0.07	0.02	3.45e-05	7	16684906	*BZW2;ANKMY2*	TSS1500;Body
cg13994827	0.31	0.08	2.41e-03	-0.47	0.11	4.01e-05	8	33715673	-	-
cg27378972	-0.30	0.10	8.73e-03	0.55	0.13	4.10e-05	8	77690582	*ZFHX4*	Body
cg13678096	0.27	0.07	1.70e-03	-0.39	0.09	4.20e-05	1	234987477	-	-
cg02350425	0.21	0.07	1.46e-02	-0.39	0.09	4.71e-05	8	126944311	*LINC00861*	Body

4.4.8 Braak LB Stage-associated DNA methylation signatures across two cortical brain regions

No DMPs were identified to be associated with Braak LB stage. This likely reflects a lack of power as the vast majority of donors were Braak LB Stage 0 (n= 385, 72%) with only 69 donors (13%) being the highest Braak LB stage. One DMR was found to be associated with Braak LB stage and the sites which make up this region reside in the *RGS19* gene (see **Table 4.17**; **Figure 4.31**; and **Figure 4.32**). *RGS19* encodes a protein belonging to the regulators of G-protein signalling family interacts with the G protein GAI3. Previous work by our group identified *RGS19* as a blood gene expression marker of AD (Lunnon *at al.*, 2013). *RGS19* was not associated with any of the other neuropathology measures.

4.4.8.1 Interaction effects between brain region and Braak LB Stage

I looked at interactions between brain region and Braak LB stage to identify if there were any associations of differing magnitude or direction of effect in either the PFC or OCC. No DMPs reached experiment wide significance for an interaction effect (p>9e-08) which is unsurprising as no main effect associations were identified between DNAm and Braak LB stage. However, at a more relaxed threshold of 5e-5 DMPs were identified. The top six DMPs are illustrated in **Figure 4.33** (see **Table 4.18** for the top 20 interaction associations). The most significant DMP with a gene annotation (cg15030392, located on chromosome 21) was hypermethylated with elevated Braak LB stage in the OCC, but there was near to no effect in the PFC (see **Figure 4.33**). This site is annotated to *DSCAM*, which plays a role in neurite growth and synaptogenesis (Jia *at al.*, 2011). Overexpression of *DSCAM* in the cerebral cortex has been associated with learning and memory defects in *APP* transgenic mice (Jia *at al.*, 2011).

Table 4.17: Differentially methylated region associated with Braak Lewy Body stage. *One DMR was identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected p-value adjusted for the number of independent tests.*

Chr	**BP start**	**BP end**	**n**	**Gene**	**Beta (%)**	**SE (%)**	**P**	**P adjusted**
20	62710773	62710905	3	*RGS19;OPRL1*	2.44	0.44	3.84e-08	3.19e-02

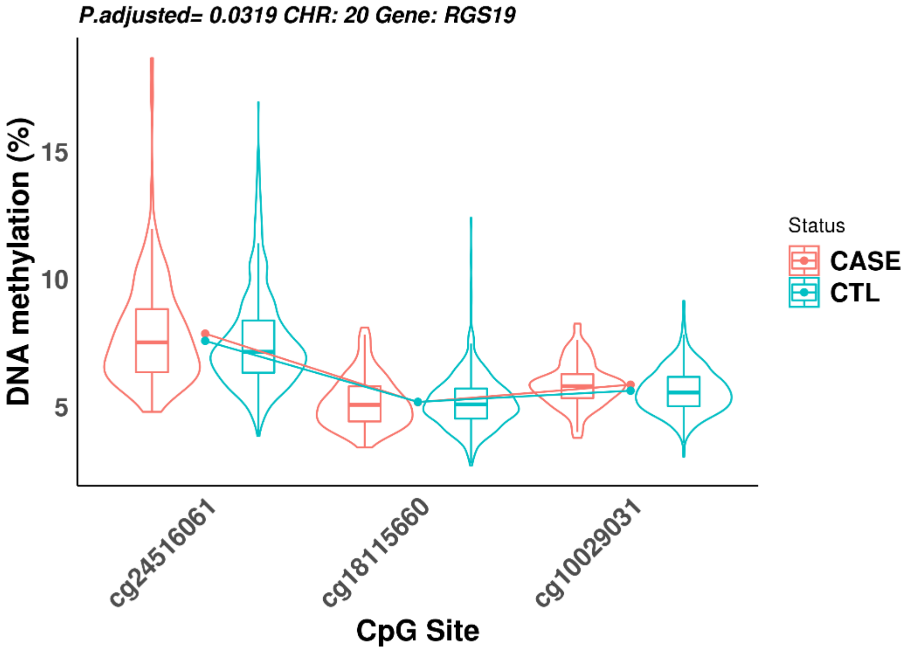


Figure 4.31: Differential methylation across a region in the gene RGS19 was associated with Braak Lewy body stage. *The x-axis are the DNAm sites ordered by genomic location. The y-axis is DNA methylation as a percentage. Violin plots for the DNA methylation values (adjusted for batch and covariates) across at each DNAm site in the DMR is shown, where the box in the middle represents the interquartile range (IQR), whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). Samples were split into cases (Braak LB stage > 3) and controls (Braak LB stage < 3) in order to visualise the differences between the groups. The grey on the outside of the box plot is a density plot of the distribution of data, where the width represents frequency.*

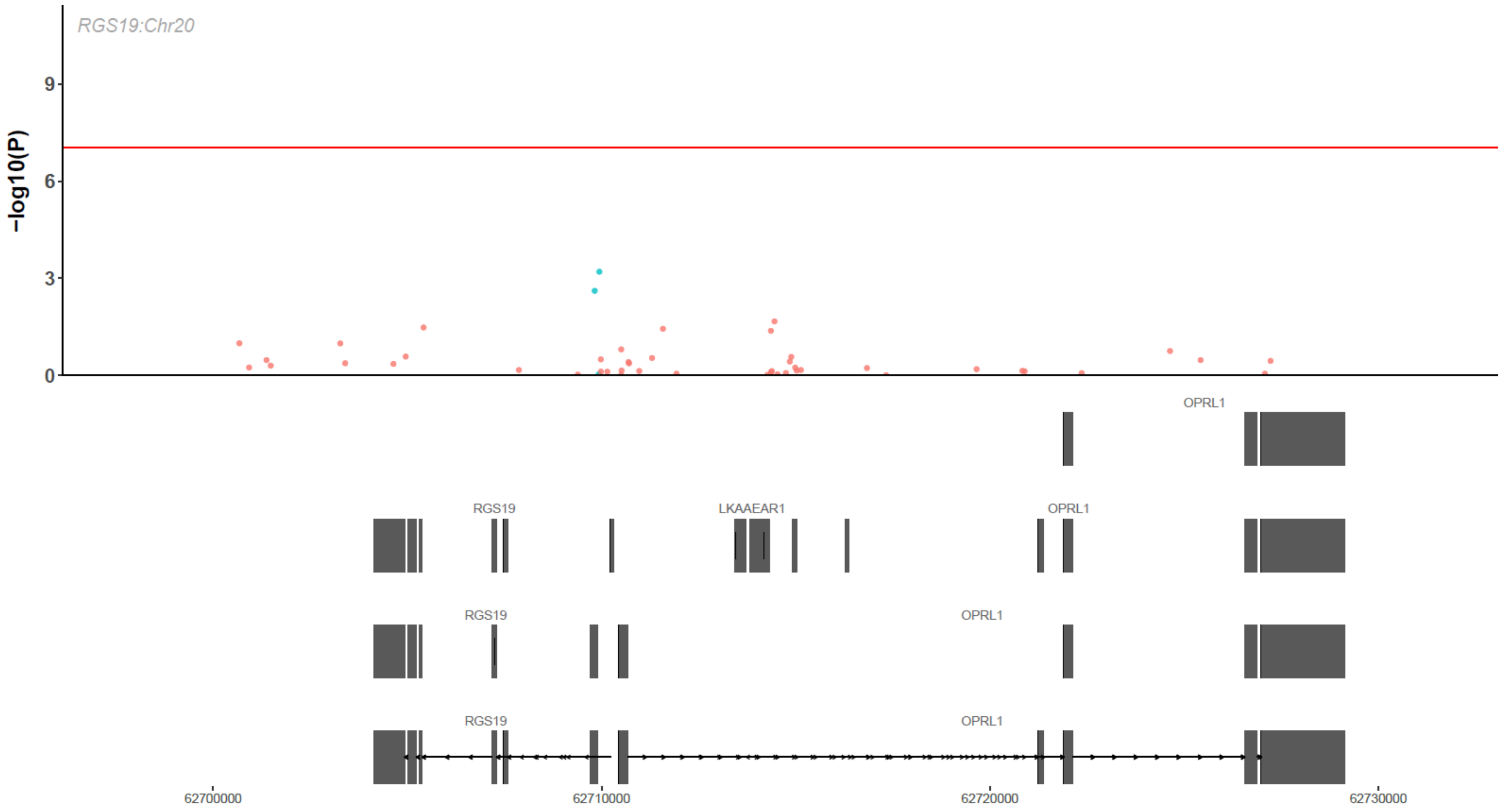


Figure 4.32: A differentially methylated region identified in the RGS19 gene within chromosome 20 was associated with Braak Lewy body Stage. *Shown on the top is a zoomed in Manhattan plot around the RGS19 region, where the x-axis represents the base position and the y-axis represents the –log10 p-value and each point on the plot represents a DNA methylation site. The red horizontal line represents experiment wide significance (P< 9e-08). The gene track shows the locations of the genes within in this region.*

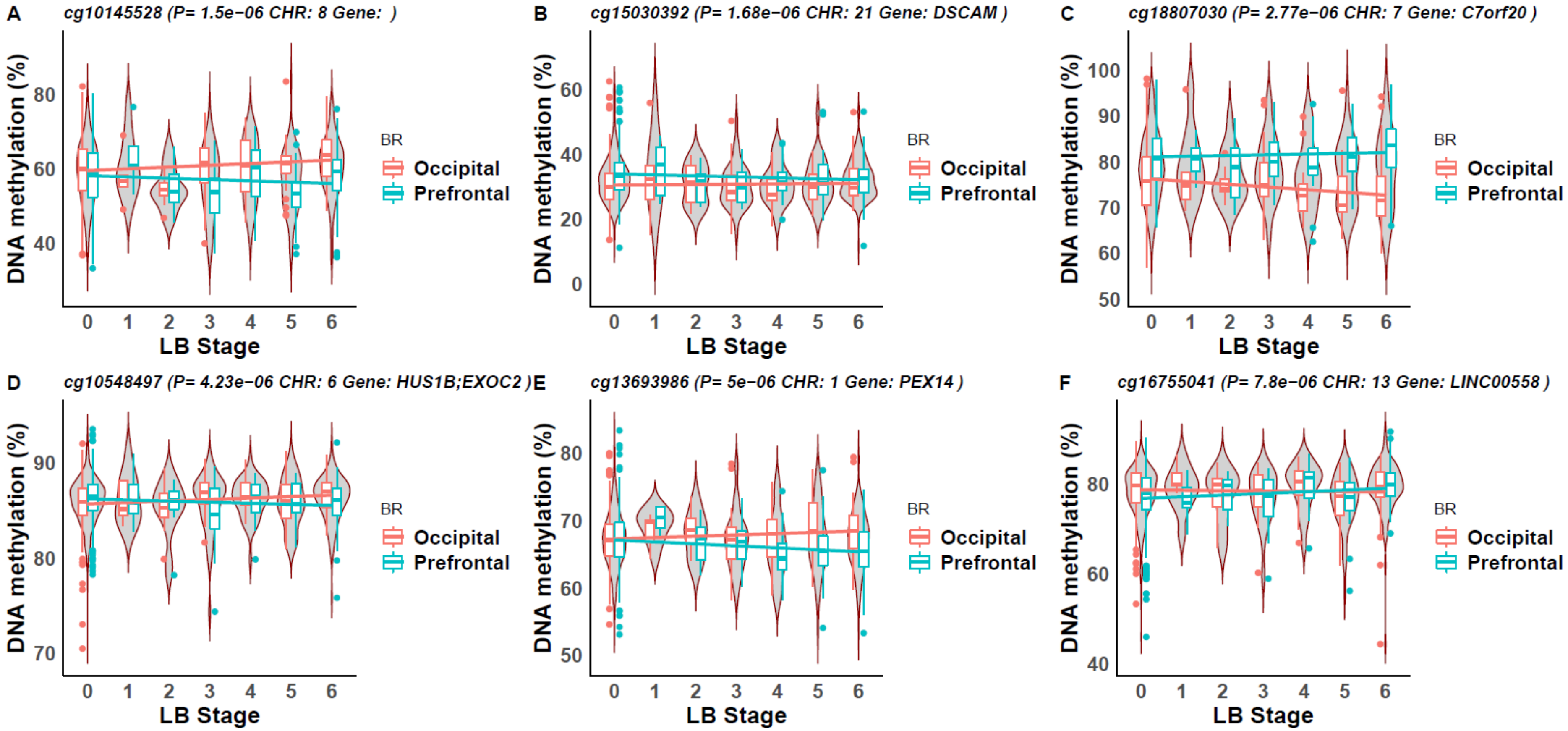


Figure 4.33: Interaction between Braak Lewy body (LB) stage and brain region. *Shown are the interaction results for the top 6 DNAm sites from the interaction EWAS, where an interaction term between Braak LB stage and brain region was included. The plots are for the six most associated differentially methylated positions. Violin plots for the DNA methylation values (adjusted for batch and covariates) across the 7 stages are shown, where the boxes in the middle represents the interquartile range (IQR) for each brain region, whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). The grey on the outside of the box plot is a density plot of the distribution of data (not split by brain region), where the width represents frequency. The x-axis is Braak LB stage. The y-axis is the methylation value as a percentage.*

Table 4.18: The top 20 differentially methylated positions identified in the interaction EWAS between brain region and Braak Lewy body (LB) stage. *Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. Shown for each DMP is the effect size (Beta), standard error (SE) and the p-value (P) for the effect of Braak LB Stage on DNAm in addition to the interaction effect (Beta Int, SE Int and P Int) between Braak LB stage and brain region (occipital cortex [reference] and prefrontal cortex).*

DNAm site	Beta (%)	SE (%)	P	Beta Int (%)	SE Int (%)	P Int	CHR	MAPINFO	Gene	Ref Group
cg10145528	0.36	0.13	4.02e-03	-0.84	0.17	1.50e-06	8	102234257	-	-
cg15030392	0.25	0.09	4.95e-03	-0.57	0.12	1.68e-06	21	41627176	*DSCAM*	Body
cg18807030	-0.41	0.08	3.77e-07	0.50	0.11	2.77e-06	7	931980	*C7orf20*	Body
cg10548497	0.16	0.05	7.58e-04	-0.29	0.06	4.23e-06	6	657338	*HUS1B;EXOC2*	TSS1500;Body;5'UTR
cg13693986	0.19	0.08	1.90e-02	-0.49	0.11	5.00e-06	1	10597737	*PEX14*	Body
cg16755041	-0.22	0.08	5.23e-03	0.51	0.11	7.80e-06	13	54388934	*LINC00558*	TSS1500
cg04456662	0.29	0.11	8.06e-03	-0.57	0.13	7.86e-06	6	30115407	*TRIM40*	Body
cg02077669	-0.52	0.13	3.55e-05	0.78	0.17	9.01e-06	1	66730524	*PDE4B*	Body
cg18151437	0.15	0.07	3.98e-02	-0.41	0.09	1.15e-05	19	56638601	-	-
cg04958157	-0.32	0.07	5.96e-06	0.40	0.09	1.23e-05	13	113748321	*MCF2L*	Body
cg23795559	-0.30	0.11	3.04e-03	0.65	0.15	1.24e-05	14	89642474	*FOXN3*	Body
cg13159060	-0.24	0.07	2.42e-04	0.39	0.09	1.26e-05	8	145701897	*FOXH1*	TSS200
cg14435390	-0.15	0.06	1.02e-02	0.33	0.07	1.40e-05	19	42913381	*LIPE;LIPe-AS1*	Body
cg23858195	0.14	0.06	2.28e-02	-0.35	0.08	1.62e-05	6	31936253	*SKIV2L*	Body
cg07667591	-0.19	0.08	1.13e-02	0.45	0.10	1.67e-05	17	1944208	*OVCA2;DPH1*	TSS1500;Body
cg07553378	0.18	0.05	6.81e-04	-0.30	0.07	1.99e-05	7	50726586	*GRB10*	Body
cg20943946	-0.41	0.11	1.09e-04	0.64	0.15	2.13e-05	14	29850567	-	-
cg14882481	0.30	0.09	8.41e-04	-0.54	0.13	2.23e-05	11	107437051	*ALKBH8*	TSS1500
cg08329137	0.23	0.07	1.88e-03	-0.42	0.10	2.56e-05	5	175615937	*LOC643201*	Body
cg17067226	0.21	0.08	5.09e-03	-0.45	0.11	2.96e-05	3	157211395	*VEPH1*	Body

4.4.9 TDP-43-associated DNA methylation signatures across two cortical brain regions

Elevated TDP-43 status was significantly hypomethylated at one site (cg06423355; p=5.47e-08) (see **Figure 4.34**; **Figure 4.35** and **Figure 4.36**). Although this DMP is not annotated to a gene based on the Illumina manifest (see **Table 4.19**), it is located ~50kb from *STK38L* (see **Figure 4.37**), which has been associated with *APP* processing in drosophila (Huichalaf *at al.*, 2019).

I identified five DMRs associated with TDP-43 status (see **Table 4.20** and **Figure 4.38**). Four were annotated to genes (*BAT1, ACADS, CACNA1G* and *EVA1C*) and all were unique to the analysis of TDP-43 status. BAT1 is part of the major histocompatibility complex (MHC) and a member of the DEAD-box family of RNA helicases; research suggests it is involved in the regulation and the production of inflammatory cytokines which are associated with AD pathology (Gnjec *at al.*, 2008). ACADS encodes for a tetrameric mitochondrial flavoprotein - a member of the acyl-CoA dehydrogenase family. Defects in *ACADS* can lead to neurological dysfunction and brain pathology such as demyelination in peripheral neurons (He *at al.*, 2007). *CACNA1G* is a low-voltage calcium channel which has been implicated in neurodegenerative diseases (Coutelier *at al.*, 2015) and downregulation of this gene has been correlated with altered *APP* processing and the occurrence of AD markers in human tissue, mice, and cellular models (Rice, Berchtold, Cotman, & Green, 2014). An example plot of the DMR residing in *CACNA1G* is shown in **Figure 4.39**. *EVA1C* is involved in carbohydrate binding and has been implicated in AD (Cacabelos, Cacabelos, & Torrellas, 2014).

4.4.9.1 Interaction effects between brain region and TDP-43 status

No DMPs reached experiment wide significance for an interaction effect (p>9e-08), which is unsurprising as there was only a single association with the main effect. However, at a more relaxed threshold of p<5e-5 there six DMPs were identified (see **Figure 4.40** and **Table 4.21**). The most significant DMP with a gene annotation (cg14914238; located on chromosome 15) was annotated to *HOMER2* and the direction of effect was different in the PFC (hypermethylation was associated with TDP-43 status) compared to the OCC (hypomethylation was association with TDP-43 status). *HOMER2* has been shown to interact with *APP* and its expression inhibits *APP* processing and in turn this reduces the production of Aβ (Parisiadou *at al.*, 2008).

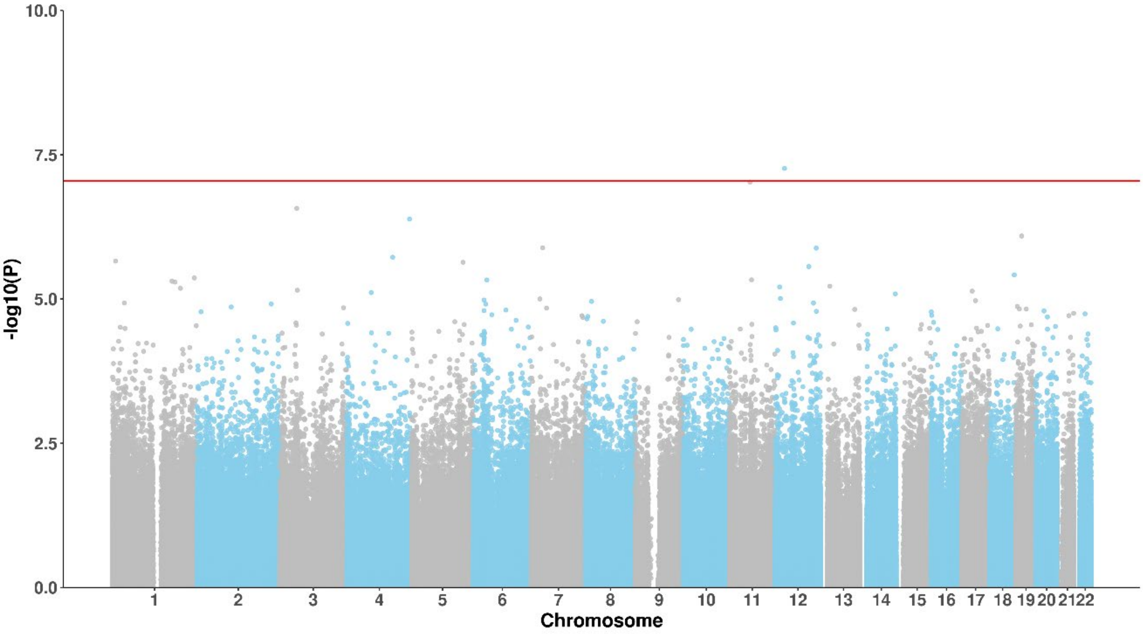


Figure 4.34: Cortex EWAS of TDP43 highlights an experiment-wide significant differentially methylated position. *A Manhattan plot showing results of TDP-43 EWAS across two cortical regions (prefrontal cortex and occipital cortex). The significant differentially methylated positions are annotated with their Illumina UCSC gene name, unless they are unannotated to a gene. The x-axis shows chromosomes 1-22 and the Y-axis shows -log10(P), with the horizontal red line representing experiment wide significance (p< 9e-08).*

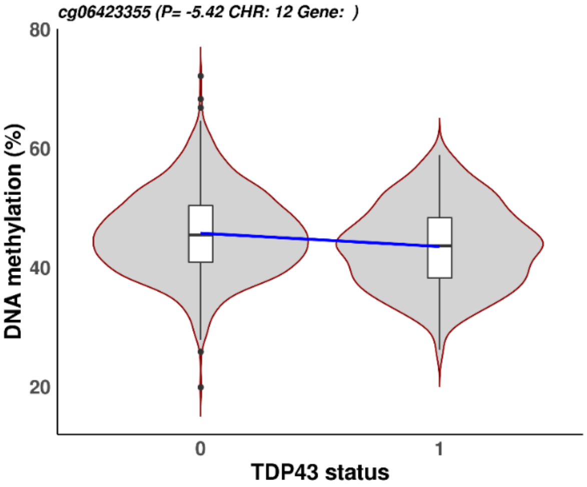


Figure 4.35 Differential methylation for TPD-43 status at cg06423355. *Violin plots for the DNA methylation values (adjusted for batch and covariates) for TDP-43 stats us shown, where the box in the middle represents the interquartile range (IQR), whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). The grey on the outside of the box plot is a density plot of the distribution of data, where the width represents frequency. The x-axis TDP-43 status (0= control, 1 = cases). The y-axis is the methylation value as a percentage.*

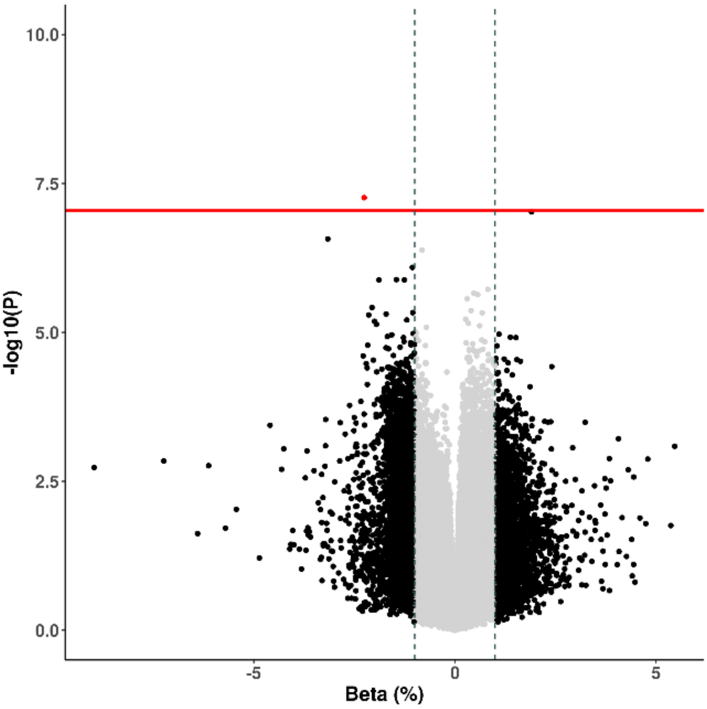


Figure 4.36: Volcano plot of differentially methylated positions (DMPs) identified in the TDP-43 status epigenome-wide association study. *The X-axis shows beta effect size (ES) and the Y-axis shows -log10(p). Black probes indicate an ES difference ≥ 0.01, whilst red probes indicate an ES difference ≥ 0.01 and a p-value that reaches experiment-wide significance (EWS) (p < 9e-08).*

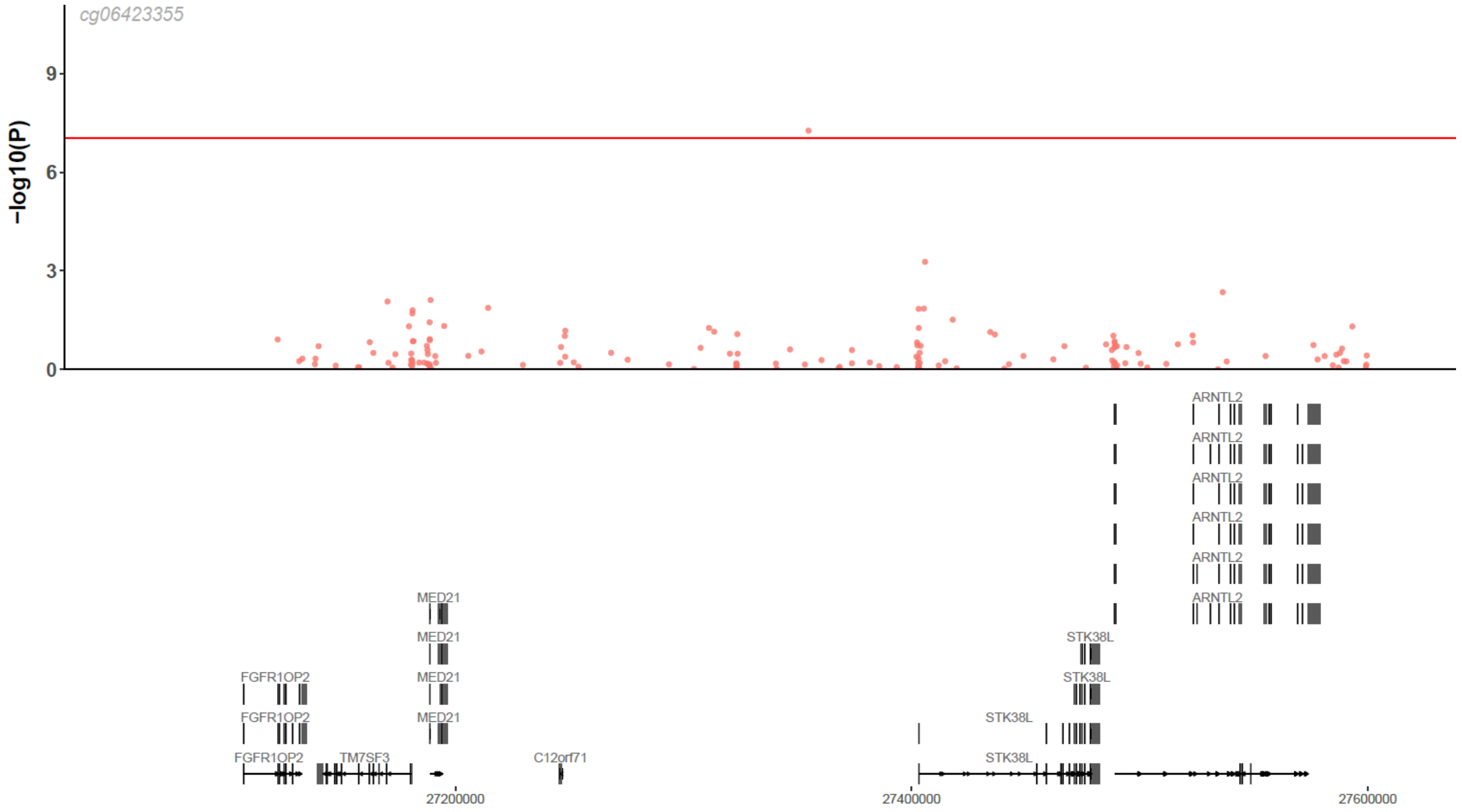


Figure 4.37: The closest gene to the DMP associated with TDP-43 status is STK38L. *Shown on the top is a zoomed in Manhattan plot around the top DMP identified in the TDP-43 EWAS, where the x-axis represents the base position and the y-axis represents the –log10 p-value and each point on the plot is a DNA methylation site. The red horizontal line represents experiment wide significance (P< 9e-08). The gene track shows the locations of the genes within in this region.*

Table 4.19: Differentially methylated position (DMP) associated with TDP-43 status at an experiment wide significance threshold (p<9e-08). *One DMP was identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. The Beta columns refers to %methylation change per unit increase in CERAD density. The 'Array' column states if probes are also present on the Illumina 450K array. The 'Braak Meta P' relates to p-values for these probes from a recent meta-analysis of AD pathology (R. Smith at al., 2021). The 'Novel Gene' column indicates if the gene has previously been implicated in neurodegenerative disease and the 'Previously Identified in' column indicates how it was previously implicated in dementia.*

DNAm site	Beta (%)	SE (%)	P	Chr	BP	Gene	Gene Region	Array	Braak meta P	Interaction BR P	Novel Gene	Previously Identified in
cg06423355	-2.256	0.42	5.47e-08	12	27347047	-	-	EPIC	-	4.70e-02	novel	-

Table 4.20: Differentially methylated regions (DMRs) associated with TDP-43 status. *In total five DMRs were identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected p-value adjusted for the number of independent tests.*

Chr	BP start	BP end	n	Gene	Beta (%)	SE (%)	P value	P adjusted
6	31507785	31508665	11	*BAT1*	-16.55	2.53	5.80e-11	5.10e-05
12	121163023	121163497	7	*ACADS*	-15.93	2.44	6.50e-11	5.72e-05
8	17767511	17768115	8	-	-15.8	2.82	2.16e-08	0.02
17	48637858	48638190	8	*CACNA1G*	11.07	1.99	2.64e-08	0.02
21	33784903	33785434	6	*C21orf63*	10.45	1.92	5.26e-08	0.05

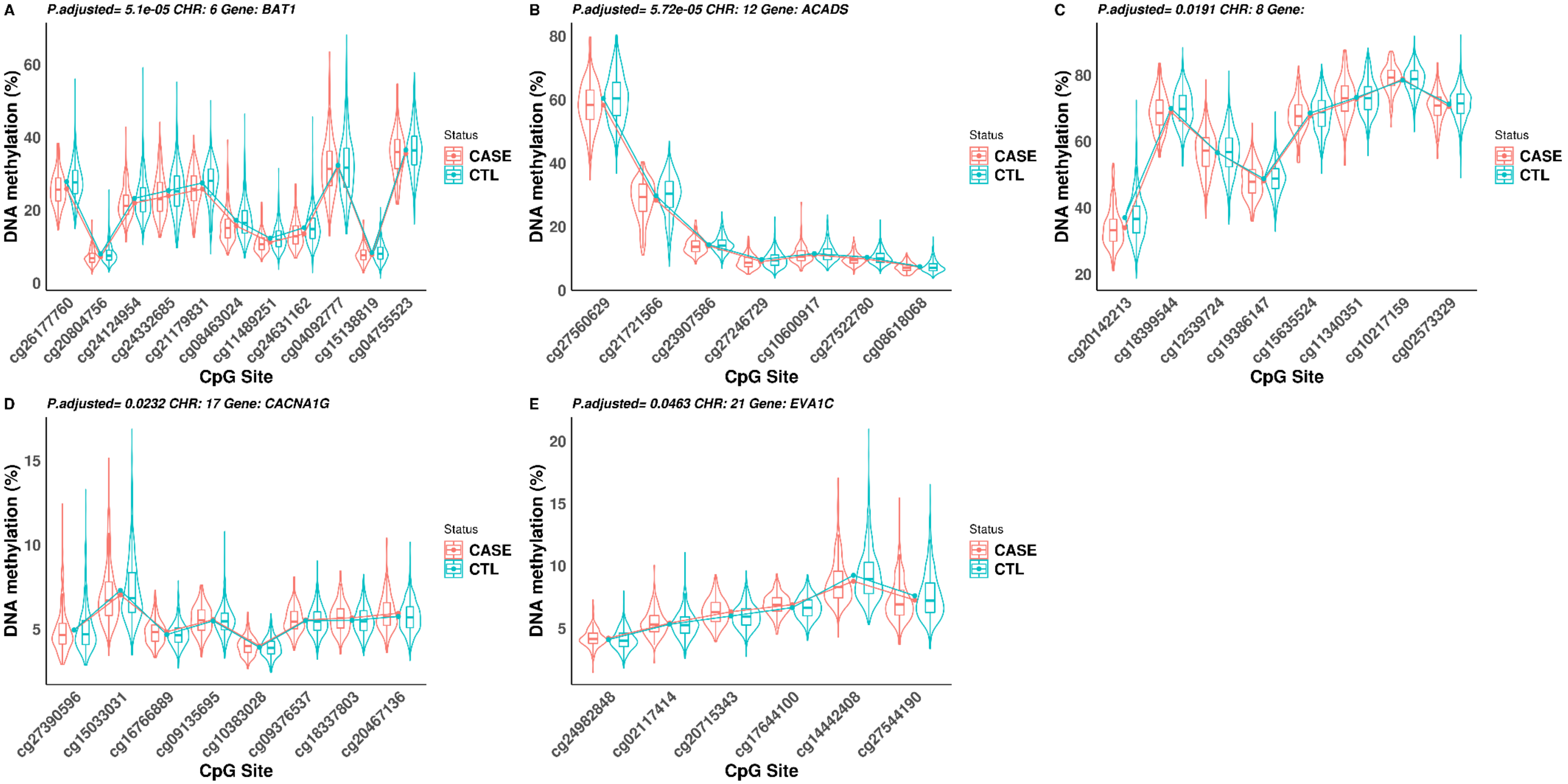


Figure 4.38: Differentially methylated regions (DMRs) associated with TDP-43 status. *The significant DMRs associated with TDP-43 status are shown. The x-axis are the DNAm sites ordered by genomic location. The y-axis is DNA methylation as a percentage. Violin plots for the DNA methylation values (adjusted for batch and covariates) across at each DNAm site in the DMR is shown, where the box in the middle represents the interquartile range (IQR), whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). DNA methylation values for cases (TPD-43 present) and controls were plotted separately to aid visualisation of differences between the groups. The grey on the outside of the box plot is a density plot of the distribution of data, where the width represents frequency.*

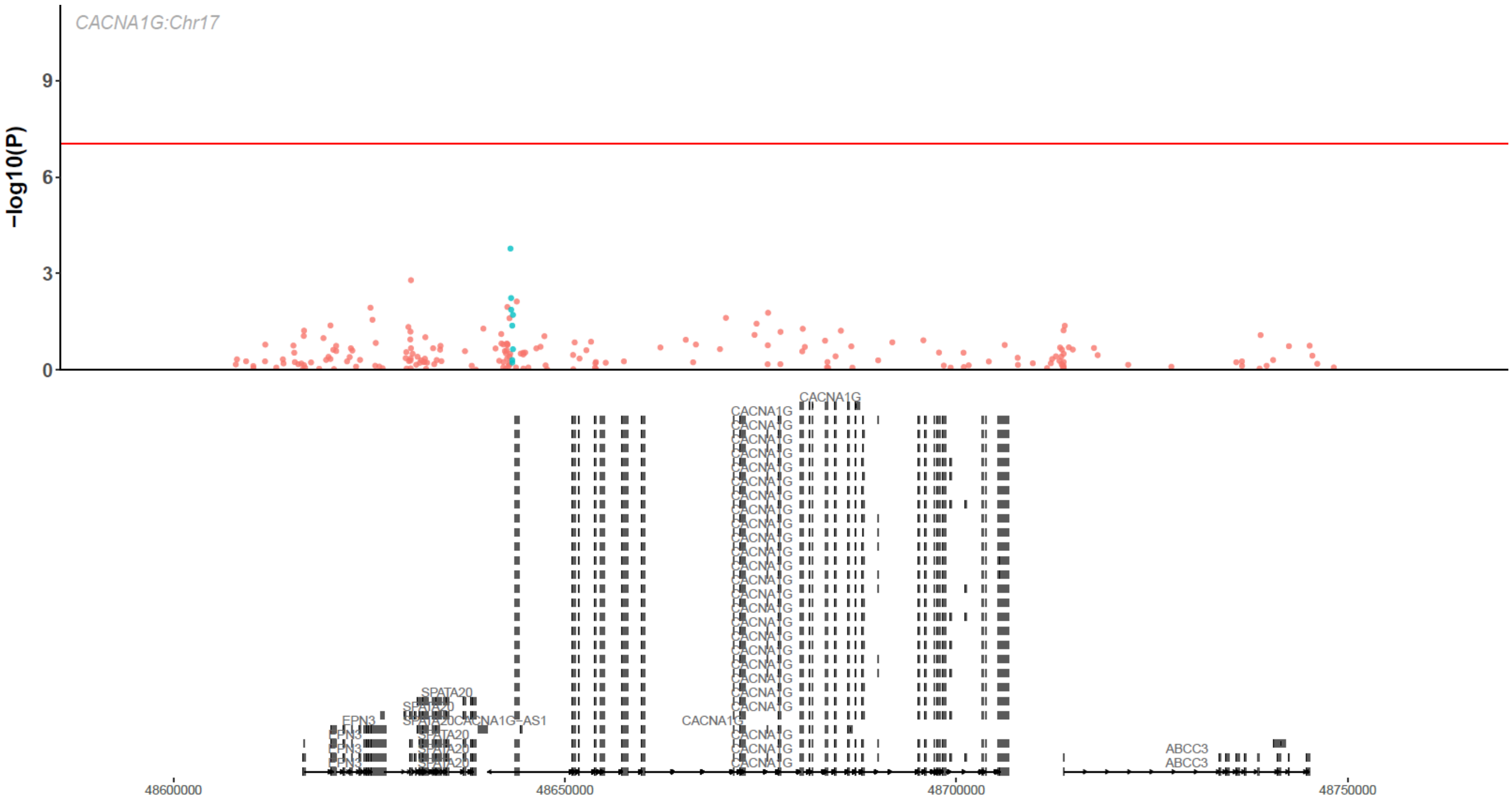


__Figure 4.39: A differentially methylated region identified in the CACNA1G gene within chromosome 17 was associated with TDP-43 status.__ Shown on the top is a zoomed in Manhattan plot around the CACNA1G region, where the x-axis represents the base position and the y-axis represents the –log10 p-value and each point on the plot represents a DNA methylation site. The red horizontal line represents experiment wide significance (P< 9e-08). The gene track shows the locations of the genes within in this region.

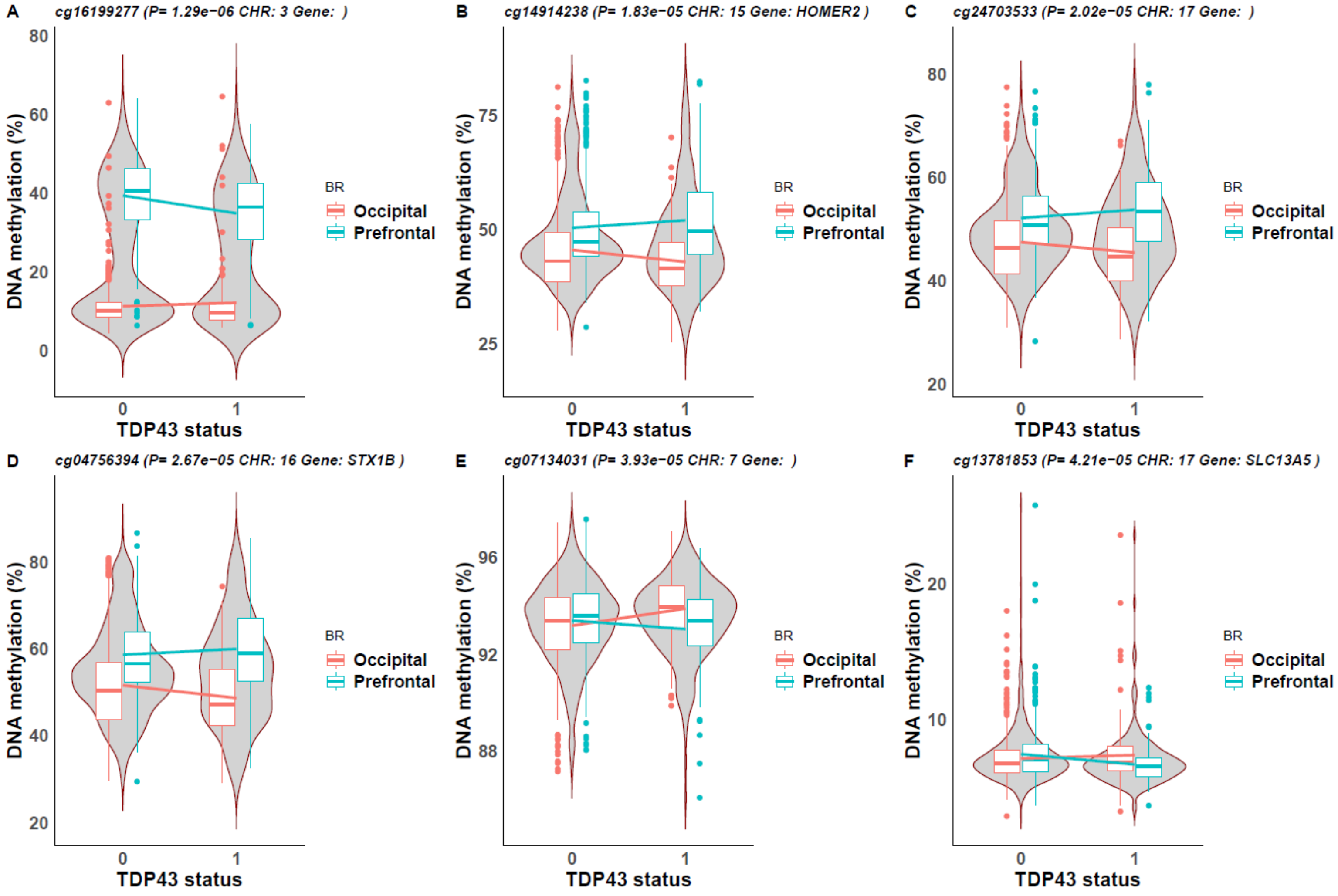


Figure 4.40: Interaction between TDP-43 status and brain region. *Shown are the interaction results for the top 6 DNAm sites from the interaction EWAS, where an interaction term between TDP-43 status and brain region was included. The plots are for the six most associated differentially methylated positions. Violin plots for the DNA methylation values (adjusted for batch and covariates) for TDP-43 status (0=control, 1=cases) are shown, where the boxes in the middle represents the interquartile range (IQR) for each brain region, whilst the whisker lines represent the minimum (quartile 1 – 1.5 x IQR) and the maximum (quartile 3 + 1.5 x IQR). The grey on the outside of the box plot is a density plot of the distribution of data (not split by brain region), where the width represents frequency. The x-axis is TDP-43 status. The y-axis is the methylation value as a percentage.*

Table 4.21: 6 differentially methylated positions were identified in the interaction EWAS between brain region and TDP43 status at the threshold p<5e-05. *Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. Shown for each DMP is the effect size (Beta), standard error (SE) and the p-value (P) for the effect of TDP-43 status on DNAm in addition to the interaction effect (Beta Int, SE Int and P Int) between TDP43 status and brain region (occipital cortex [reference] and prefrontal cortex).*

DNAm site	**Beta (%)**	**SE (%)**	**P**	**Beta Int (%)**	**SE Int (%)**	**P Int**	**CHR**	**MAPINFO**	**Gene**	**Ref Group**
cg16199277	2.14	0.88	4.48e-02	-5.73	1.18	1.29e-06	3	147089290	-	-
cg14914238	-1.20	0.35	1.00e-04	2.03	0.47	1.83e-05	15	83620321	*HOMER2*	Body
cg24703533	-0.92	0.33	9.46e-04	1.87	0.44	2.02e-05	17	45055318	-	-
cg04756394	-1.46	0.34	2.10e-06	1.94	0.46	2.67e-05	16	31017083	*STX1B*	Body
cg07134031	0.59	0.16	8.75e-04	-0.88	0.21	3.93e-05	7	157073436	-	-
cg13781853	0.54	0.19	1.45e-02	-1.03	0.25	4.21e-05	17	6616867	*SLC13A5*	TSS200

4.4.10 Global Clinical Dementia Rating is correlated with neuropathology

In addition to identifying methylomic variation associated with neuropathology, I investigated differences in DNAm associated with cognitive decline measured using the clinical dementia rating (CDR). Cognitive decline has previously been shown to be associated with brain atrophy and neuropathology (Thomas *at al.*, 2020). To identify DNAm sites in which differential methylation is associated with the CDR I ran an EWAS controlling for age, sex, batch, derived cell proportions and PC1 as fixed effects and individual as a random effect. To identify if the effects were consistent across the two brain regions (OCC and PFC) I ran a further EWAS including an interaction between brain region and the measure of interest.

I identified ten DMPs associated with CDR (see **Figure 4.41** and **Table 4.22**). The average magnitude of effect for the significant DMPs per additional CDR point was 0.20% (inter-quartile range [IQR] = 0.07-0.33%). The effects were consistent across the regions as shown by the non-significant interaction term between the CDR and brain region (P > 9e-08; see **Table 4.22**). All of the DMPs were significantly hypermethylated with increasing cognitive decline (i.e. increased DNA methylation was associated with elevated CDR score). Of the ten, eight were annotated to unique genes and four of these have not previously been implicated in neurodegenerative disease (ST3GAL1, ARHGAP22, CD5L and FHL3). The protein encoded by ST3GAL1 is a membrane protein involved in catalysing sialic acid and is predominantly found in the Golgi (Wu *at al.*, 2018) and ST3GAL3 mutations impair the development of higher cognitive functions (Hu *at al.*, 2011). ARHGAP22 (Rho GTPase Activating Protein 22) encodes a GTPase activating protein and Rho GTPases have been implicated in AD pathogenesis (Aguilar, Zhu, & Lu, 2017).

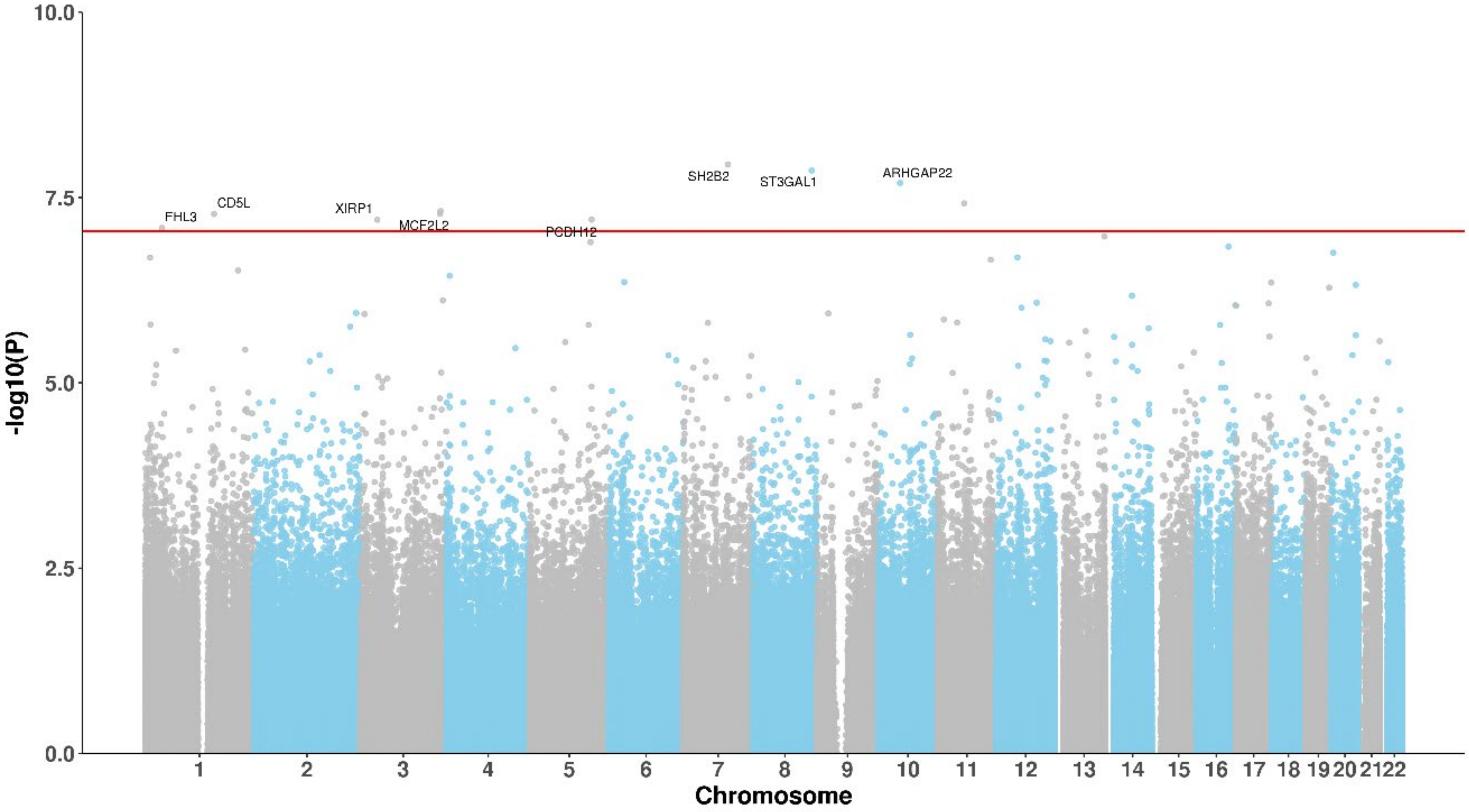


Figure 4.41: Cortex EWAS of the Clinical Dementia Rating (CDR) highlights experiment-wide significant differentially methylated positions. *A Manhattan plot showing results of a CDR EWAS across two cortical regions (prefrontal cortex and occipital cortex). The significant differentially methylated positions are annotated with their Illumina UCSC gene name, unless they are unannotated to a gene. The x-axis shows chromosomes 1-22 and the Y-axis shows -log10(P), with the horizontal red line representing experiment wide significance (p< 9e-08).*

Table 4.22: Differentially methylated positions (DMPs) associated with Clinical Dementia Rating at an experiment wide significance threshold (p<9e-08). *In total ten DMPs were identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. The Beta columns refers to %methylation change per unit increase in CDR. The 'Array' column states if probes are also present on the Illumina 450K array. The 'Braak Meta P' relates to p-values for these probes from a recent meta-analysis of AD pathology (R. Smith at al., 2021). The 'Novel Gene' column indicates if the gene has previously been implicated in neurodegenerative disease and the 'Previously Identified in' column indicates how it was previously implicated in dementia.*

DNAm site	Beta (%)	SE (%)	P	Chr	BP	Gene	Gene Region	Array	Braak meta P	Interaction BR P	Novel Gene
cg08952306	0.98	0.15	1.13e-08	7	101962123	*SH2B2*	3'UTR	450K & EPIC	1.27e-02	1.82e-01	AD GWAS
cg11715466	0.60	0.09	1.37e-08	8	134515903	*ST3GAL1*	5'UTR	EPIC	-	1.95e-01	-
cg14325837	0.49	0.08	2.02e-08	10	49731563	*ARHGAP22*	Body	450K & EPIC	2.19e-04	8.90e-04	-
cg07394019	-0.70	0.11	3.80e-08	11	61146265	-		EPIC	-	7.66e-02	-
cg08944839	0.56	0.09	4.83e-08	3	184414548	-		EPIC	-	4.25e-01	-
cg22909901	0.61	0.10	5.26e-08	3	182981707	*MCF2L2;B3GNT5*	Body	450K & EPIC	2.55e-05	5.38e-01	AD DNAm
cg11139878	0.70	0.11	5.29e-08	1	157811790	*CD5L*	TSS200	450K & EPIC	1.16e-02	1.42e-01	-
cg12845808	0.60	0.09	6.29e-08	5	141338604	*PCDH12*	1stExon	450K & EPIC	6.94e-07	4.23e-02	AD DNAm
cg02921257	0.48	0.08	6.30e-08	3	39233858	*XIRP1*	5'UTR	450K & EPIC	3.78e-04	6.25e-02	AD DNAm
cg16464569	0.53	0.09	8.16e-08	1	38472003	*FHL3*	TSS1500	EPIC	-	3.71e-01	-

4.4.10.1 Multiple DMRs associated with CDR

In addition to the 10 DMPs, I identified 15 DMRs associated with cognitive decline; of these 14 were annotated to genes (see **Table 4.23**). The most significant DMR which is annotated to *KLHL33* was also identified in the neuropathology EWAS and DMR analysis and has been described above (see **4.4.5.4**). However most of the genes which the DMRs were annotated to were unique to CDR including *GRASP, FAM110A, SMYD2*, *RPH3AL* and *HYAL2*. The largest DMR contained 12 CpG sites and was located on chromosome 12 and annotated to *GRASP. GRASP* is involved in neurite development and studies suggest it regulates cognitive function by modulation of neuronal plasticity (Yanpallewar, Barrick, Palko, Fulgenzi, & Tessarollo, 2012). A recent study found that *GRASP* variation is associated with memory function and cognitive ability in men which schizophrenia, although the findings can extend to other disorders where cognitive functioning is a core feature such as in dementia (Matosin, Green, Newell, & Fernandez-Enright, 2017). My results further support a role for *GRASP* in cognitive functioning.

Table 4.23: Differentially methylated regions (DMRs) associated with Clinical Dementia Rating. *In total 15 DMRs were identified. Probe information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected p-value adjusted for the number of independent tests. Gene = UCSC Gene name.*

CHR	BP start	BP end	n	Gene	Beta (%)	SE (%)	P value	P adjusted
14	20903445	20904169	10	*KLHL33*	8.6	1.17	2.04e-13	1.84e-07
12	52403511	52404853	12	*GRASP*	4.26	0.62	5.16e-12	4.64e-06
9	35563884	35564160	5	*FAM166B*	7.97	1.19	2.15e-11	1.93e-05
20	821350	822198	8	*FAM110A*	-3.98	0.69	8.59e-09	7.72e-03
1	214454282	214454506	5	*SMYD2*	7.3	1.27	8.83e-09	7.94e-03
17	202759	202988	3	*RPH3AL*	11.53	2	8.88e-09	7.98e-03
3	50359978	50360690	7	*HYAL2*	4.91	0.85	9.16e-09	8.23e-03
12	115123251	115124067	5		-9.43	1.65	1.09e-08	9.76e-03
7	1575621	1575824	3	*MAFK*	10.51	1.85	1.43e-08	1.29e-02
13	22033473	22033696	3	*ZDHHC20*	9.49	1.69	1.90e-08	1.71e-02
1	109849545	109849854	4	*MYBPHL*	4.47	0.8	1.96e-08	1.76e-02
2	220298547	220299116	3	*SPEG*	9.97	1.81	3.28e-08	2.95e-02
19	30302030	30302793	6	*CCNE1*	6.05	1.1	3.35e-08	3.01e-02
2	86115700	86116061	3	*ST3GAL5-AS1;ST3GAL5*	8.26	1.5	3.64e-08	3.27e-02
14	77542608	77542763	4	*LOC102724190*	7.17	1.31	4.14e-08	3.73e-02

4.4.10.2 DMPs associated with Alzheimer's disease across the cortex are predominantly a result of the disease itself as opposed to specific neuropathology measures

To identify DMPs which are unique to each neuropathology measure I conducted conditional analyses in which I ran an EWAS for each of the five neuropathology measures controlling for the other four measures in addition to age, sex, batch, derived neural cell proportions and PC1. No significant DMPs remained for any of the five pathology measures when controlling for the other four measures, suggesting that differential methylation is predominantly driven by some aspect of neurodegeneration rather than reflecting specific features of each neuropathology measure. This can be visualised when looking at the quantile-quantile (QQ) plots, where the plots lie along the null line when neuropathology measures are included as covariates in comparison to when they are not (see **Figure 4.43**). Although the QQ plots where neuropathology was not included as a covariate appear slightly inflated, it is worth noting that the lambda's were small; there was no evidence of bias in the t-statistics (see **Figure 4.43**) and including the neuropathology measures led to no significance of any DNAm site, suggesting that these results are driven by neuropathology/ disease and not an uncontrolled latent variable. Of note, the inclusion of just one other neuropathology measure as a covariate also reduced the number of significant hits to zero across all analyses.

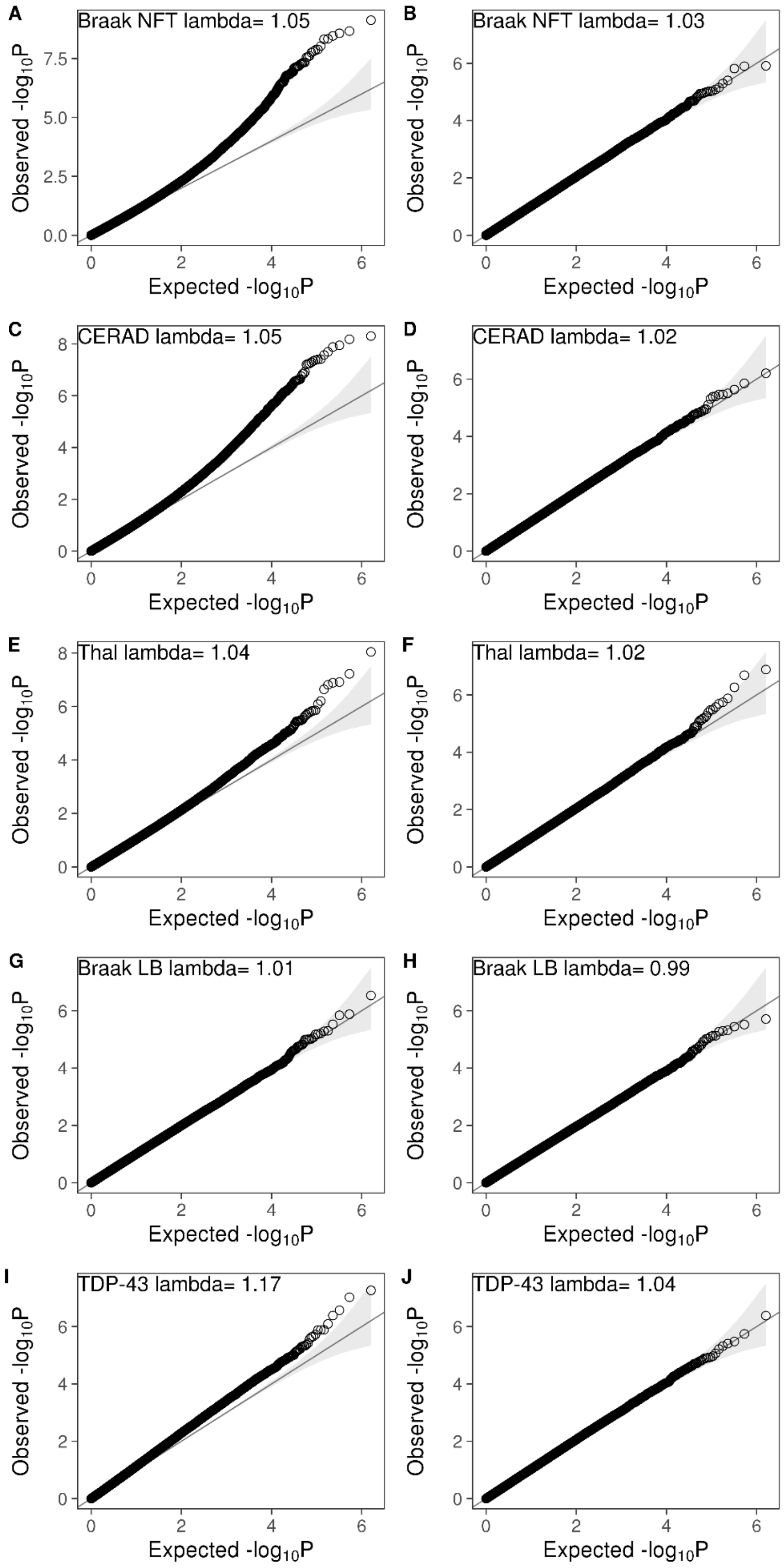


Figure 4.42: Quantile-quantile plots of the ten neuropathology EWAS. *Shown are the expected (x-axis) against the observed (y-axis) quantiles in each EWAS against neuropathology where in (A, C, E, G and I) the EWAS of each neuropathology measure (Braak neurofibrillary tangle stage, CERAD, Thal Phase, Braak Lewy Body Stage and TDP-43 status, respectively) controlled for age, sex, batch, cell proportions and PC1 as fixed effects and individual as a random effect; and (B, D, F, H and J) the EWAS of each neuropathology measure (Braak neurofibrillary tangle stage, CERAD, Thal Phase, Braak Lewy Body Stage and TDP-43 status, respectively) controlled for age, sex, batch, cell proportions, PC1and the other four neuropathology measures as fixed effects and individual as a random effect. Lambda = the genomic inflation factor which is a measure of inflation - a lambda > 1 suggests there is inflation in the data.*

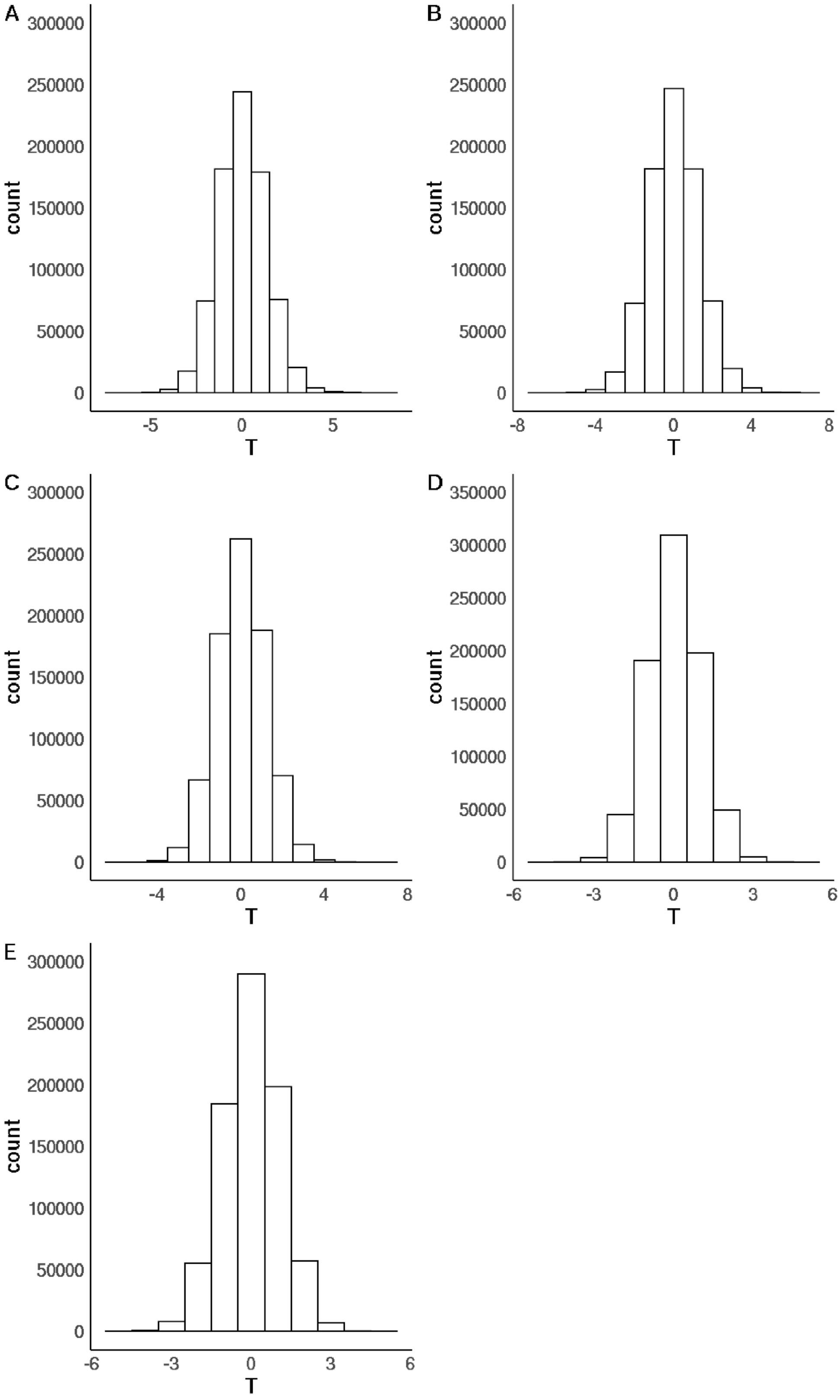


***Figure 4.43: There was no evidence of bias in the t-statistics for each of the five neuropathology EWAS.** T-statistics for the EWAS of each neuropathology measure controlling for age, sex, batch, cell proportions and PC1 as fixed effects and individual as a random effect where neuropathology is (A) Braak neurofibrillary tangle stage; (B) CERAD density; (C) Thal Phase; (D) Braak Lewy Body Stage; and (E) TDP-43 status.*

4.4.11 The direction of differential DNAm is generally consistent across the neuropathology measures

Many of the same DMPs and genes were identified across the different neuropathology EWAS. Additionally, the suggestive significant (p<5e-05) DMPs were characterised by the same direction of effect across the analyses, as demonstrated by the significant (Bonferroni p<0.05/25=0.002) binomial sign tests (see **Figure 4.44**-**Figure 4.49**). This indicates that differential methylation is predominantly driven by the disease itself rather than reflecting specific features of each neuropathology measure. The most consistent results, as statistically evaluated using the sign test, were between Braak NFT stage, Thal Phase and CERAD density (see **Figure 4.44**-**Figure 4.49**). Although there was still strong evidence of consistent effects, the neuropathology measure with the least concordance with the other measures was Braak LB Stage. Interestingly this was the only measure where the effects sizes did not correlate with CDR (sign test p=0.13; see **Figure 4.47**). Of note, Braak LB stage had the smallest number of suggestive DMPs due to lower power and therefore we cannot conclude that this neuropathology measure is presenting a different pattern to the other measures.

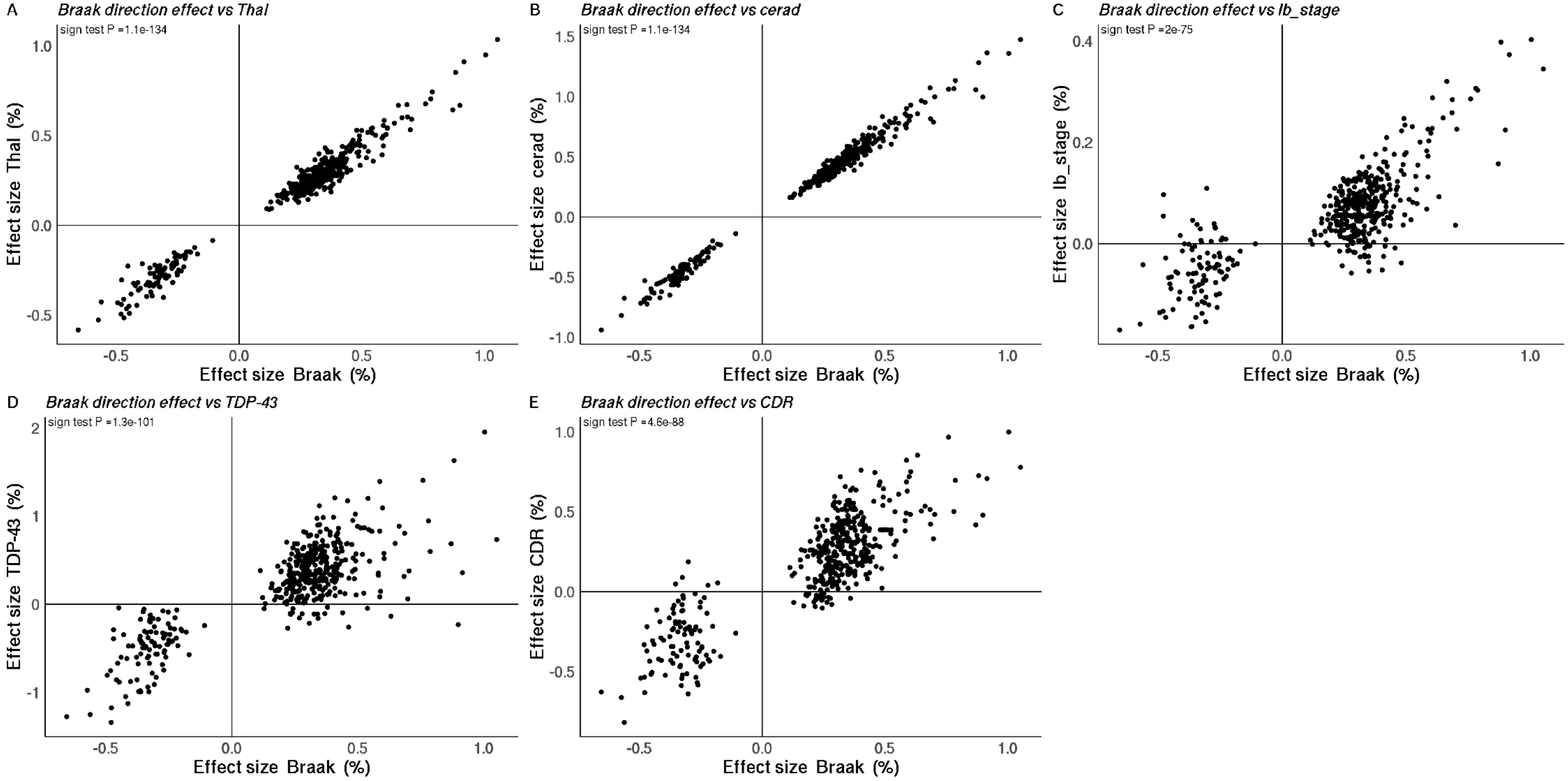


Figure 4.44: Direction of effect of the suggestive significant hits (p<5e-05) from the Braak neurofibrillary tangle (NFT) stage EWAS were generally consistent with the other neuropathology measures. *Comparing the consistency of effect sizes (% change per Braak NFT stage) between Braak NFT stage and the other neuropathology and clinical dementia rating (CDR) EWAS. Sign test P = p-value of binomial sign test of effect sizes which statistically evaluates consistency across the measures.*

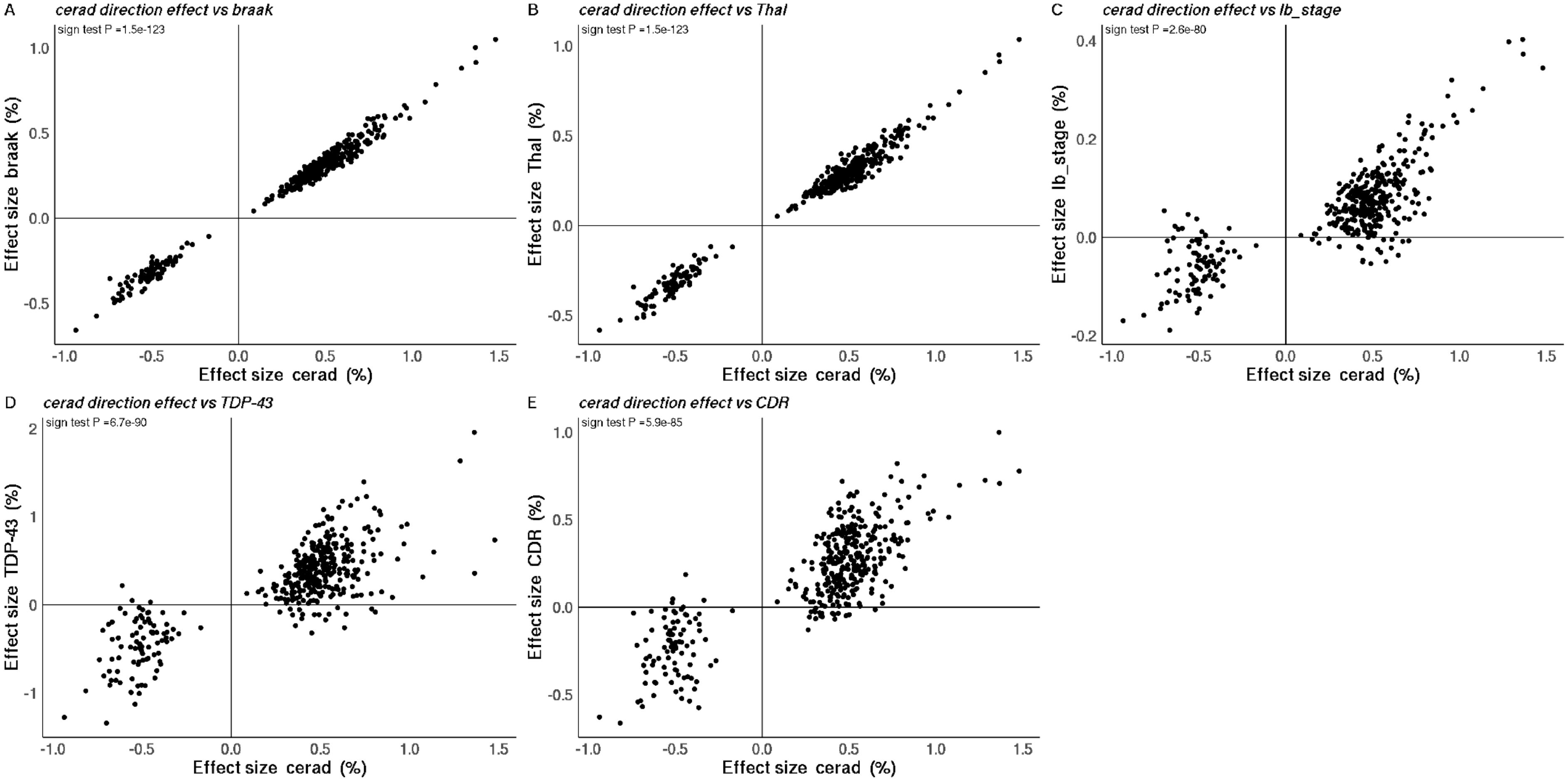


Figure 4.45: Direction of effect of the suggestive significant hits (p<5e-05) from the CERAD density EWAS were generally consistent with the other neuropathology measures. *Comparing the consistency of effect sizes (% change per CERAD density level) between CERAD density and the other neuropathology and clinical dementia rating (CDR) EWAS. Sign test P = p-value of binomial sign test of effect sizes which statistically evaluates consistency across the measures.*

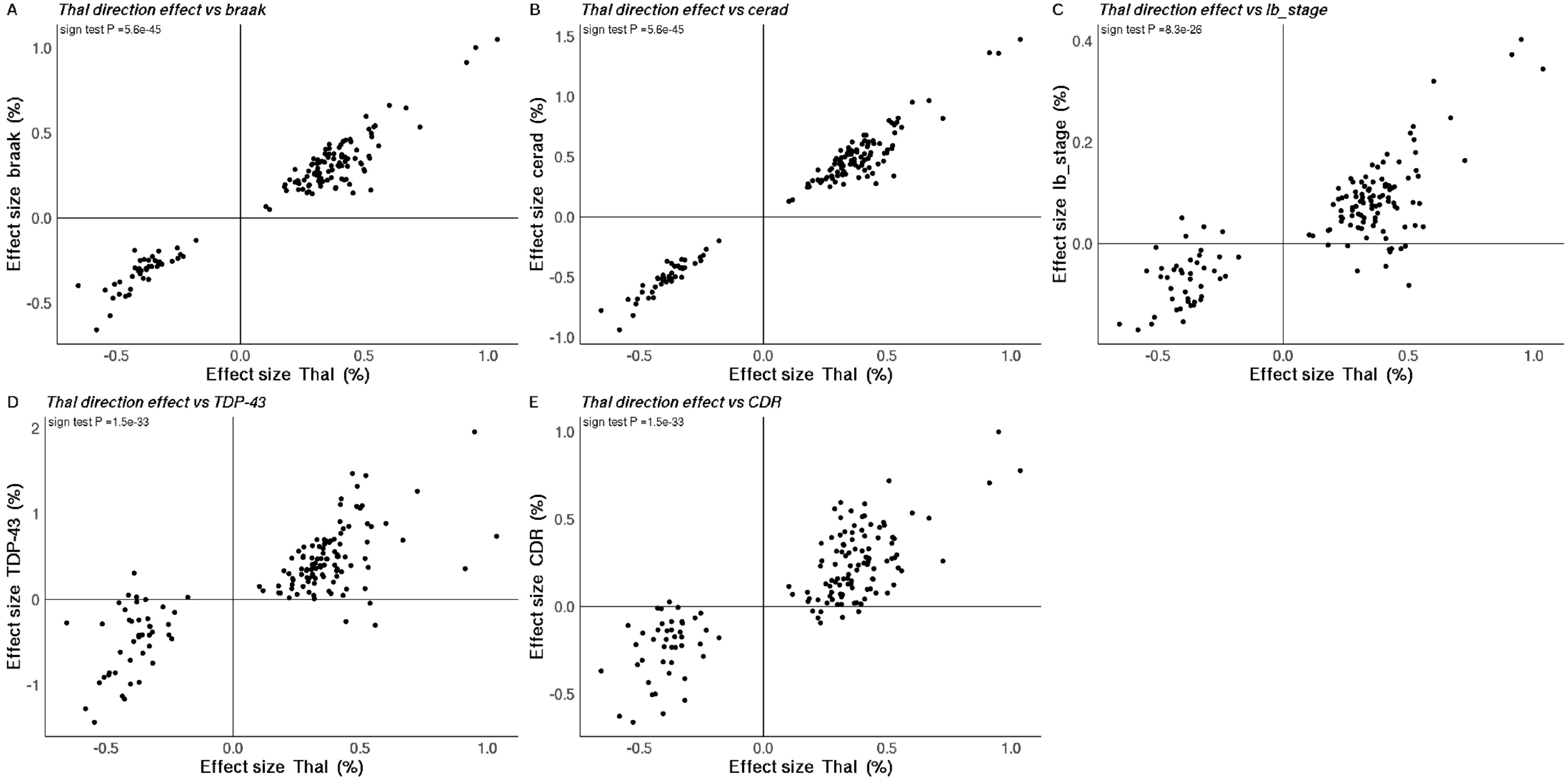


Figure 4.46: Direction of effect of the suggestive significant hits (p<5e-05) from the Thal phase EWAS were generally consistent with the other neuropathology measures. *Comparing the consistency of effect sizes (% change per Thal Phase) between Thal phase and the other neuropathology and clinical dementia rating (CDR) EWAS. Sign test P = p-value of binomial sign test of effect sizes which statistically evaluates consistency across the measures.*

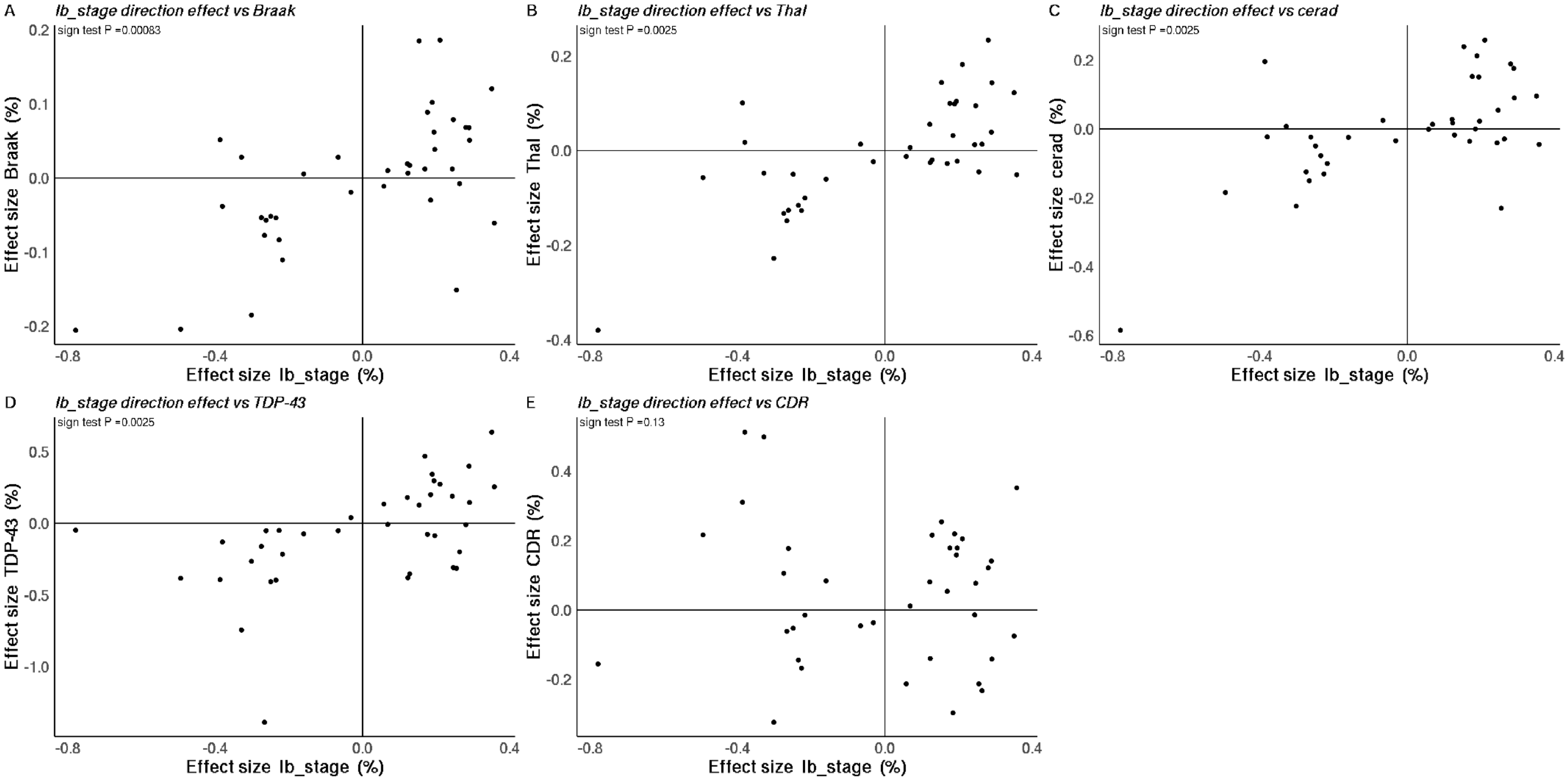


Figure 4.47: Direction of effect of the suggestive significant hits (p<5e-05) from the Braak Lewy Body (LB) stage EWAS had some consistence with the other neuropathology measures but not the clinical dementia rating (CDR). *Comparing the consistency of effect sizes (% change per Braak LB stage) between the Braak LB stage EWAS and the other neuropathology and clinical dementia rating (CDR) EWAS. Sign test P = p-value of binomial sign test of effect sizes which statistically evaluates consistency across the measures.*

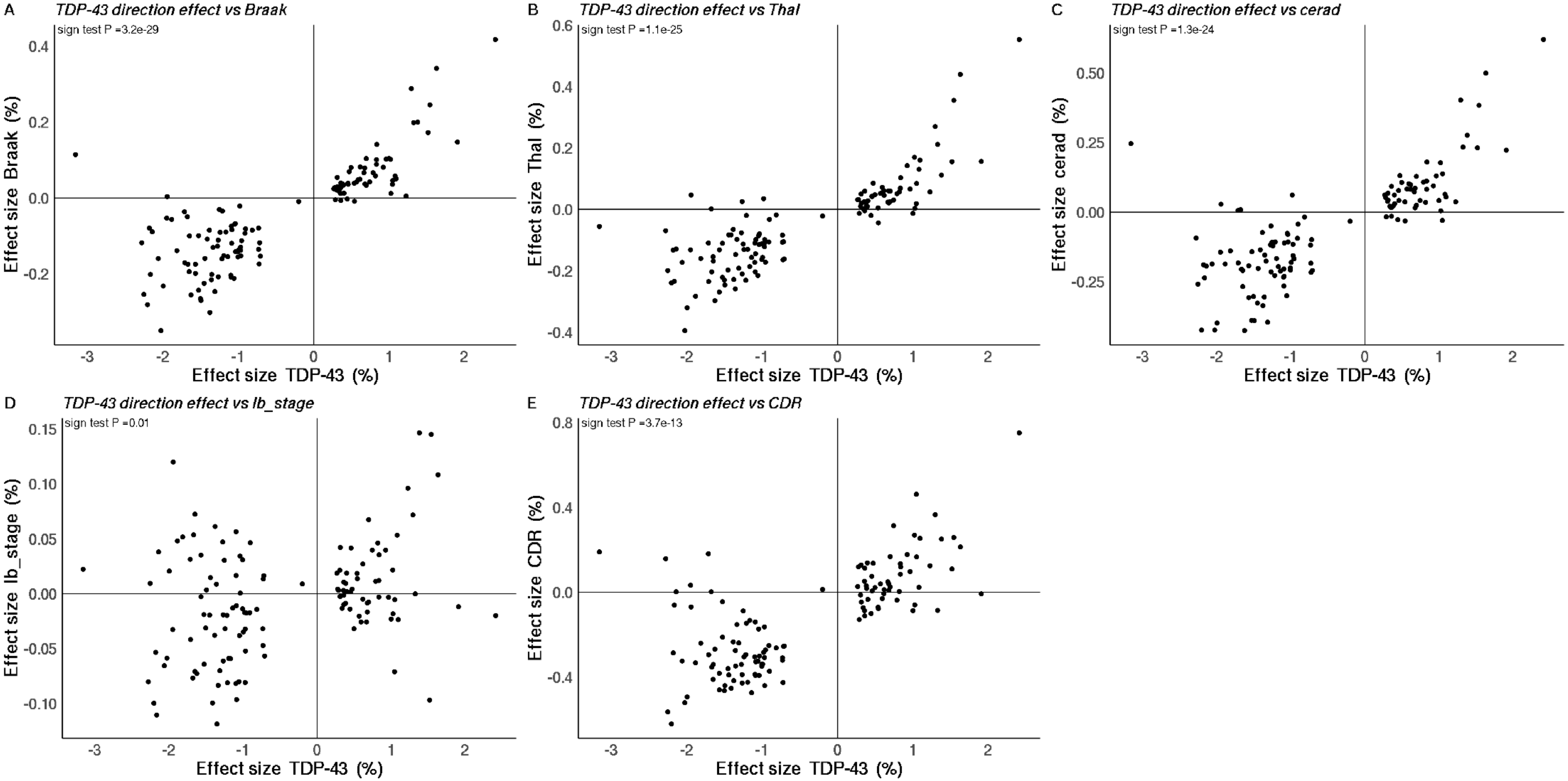


Figure 4.48: Direction of effect of the suggestive significant hits (p<5e-05) from the TDP-43 status EWAS were generally consistent with the other neuropathology measures. *Comparing the consistency of effect sizes (% change between presence TDP-43 vs none) between TDP-43 status and the other neuropathology and clinical dementia rating (CDR) EWAS. Sign test P = p-value of binomial sign test of effect sizes which statistically evaluates consistency across the measures.*

I then ran regional analysis on these results. Only TDP-43 had a significant DMR after controlling for the other neuropathology measures, which resided in *ACADS* (see **Table 4.24**; see **4.4.9** for a description of this gene).

Table 4.24: 1 DMR was associated with TDP-43 status after controlling for the other neuropathology measures.

CHR	start	end	n	Beta (%)	SE (%)	p.value	p.adjusted	Gene	Gene Region
12	121163023	121163497	7	-0.16	2.88e-02	1.99e-08	1.73e-02	*ACADS*	TSS200

4.4.12 WGCNA analysis: Clusters of methylated loci were associated with neuropathology

To identify clusters of probes that are co-methylated and hypothesized to share a common function, WGCNA analysis was performed which classified the filtered data set of 400,458 CpG probes into 14 separate modules (see **Figure 4.49**). These modules were then correlated with neuropathology as well as other traits including derived cell proportions, age, sex, *APOE* genotype and technical factors (e.g. experimental batch; see **4.3.7.1**).

The red (n =1647 DNAm sites), black (n = 1045 DNAm sites), tan (n = 242 DNAm sites) and brown (n = 13,557 DNAm sites) modules were all significantly positively correlated with Braak NFT stage, CERAD density, Thal stage and TDP-43 status (see **Figure 4.50**). The tan and brown modules were also positively correlated with Braak LB stage (see **Figure 4.50**). Braak NFT stage, CERAD density, Thal Phase and TDP-43 status were negatively correlated with the green (n =8803 DNAm sites), yellowgreen (n = 290 DNAm sites) and turquoise (n = 59,603 DNAm sites) modules. However, since these modules were very strongly correlated with other traits (predominantly derived cell proportions and batch), I ran a regression analysis between the modules using mixed effects regression models including age, sex, derived cell proportions and plate as fixed effects and brain region as a random effect. After controlling for covariates, Braak NFT stage remained significant (Bonferroni $p<0.05/28 = 0.002$) with the aforementioned modules apart from the tan module (see **Table 4.25**). CERAD density was significantly positively associated with the red module (p=4.30e-03) and significantly negatively associated with the turquoise (p=0.0013) and green modules (1.90e-03; see **Table 4.25**). Thal phase was significantly negatively associated with the green module (p= 1.30e-03). TDP-43 status was significantly positively associated the green module (p= 8.90e-03). After controlling for covariates Braak LB stage was not associated with any module.

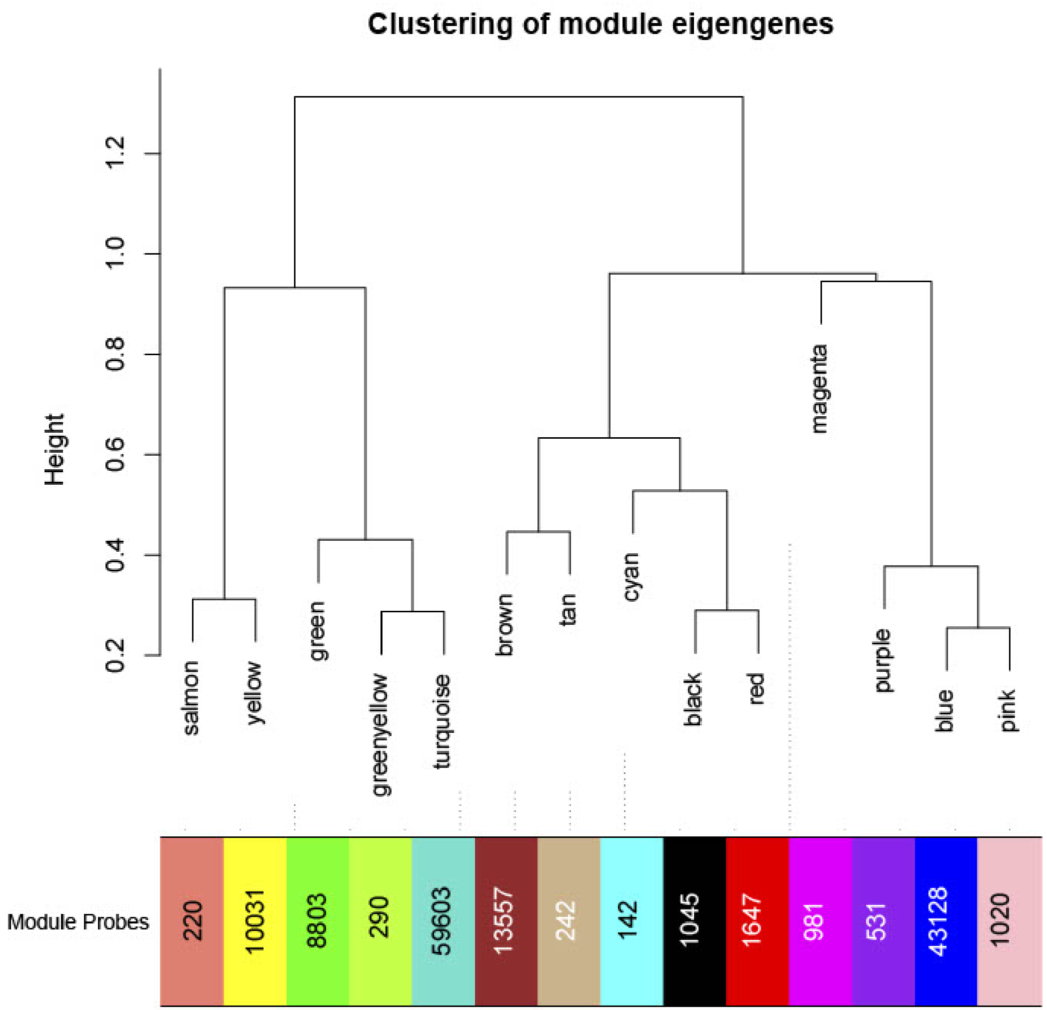
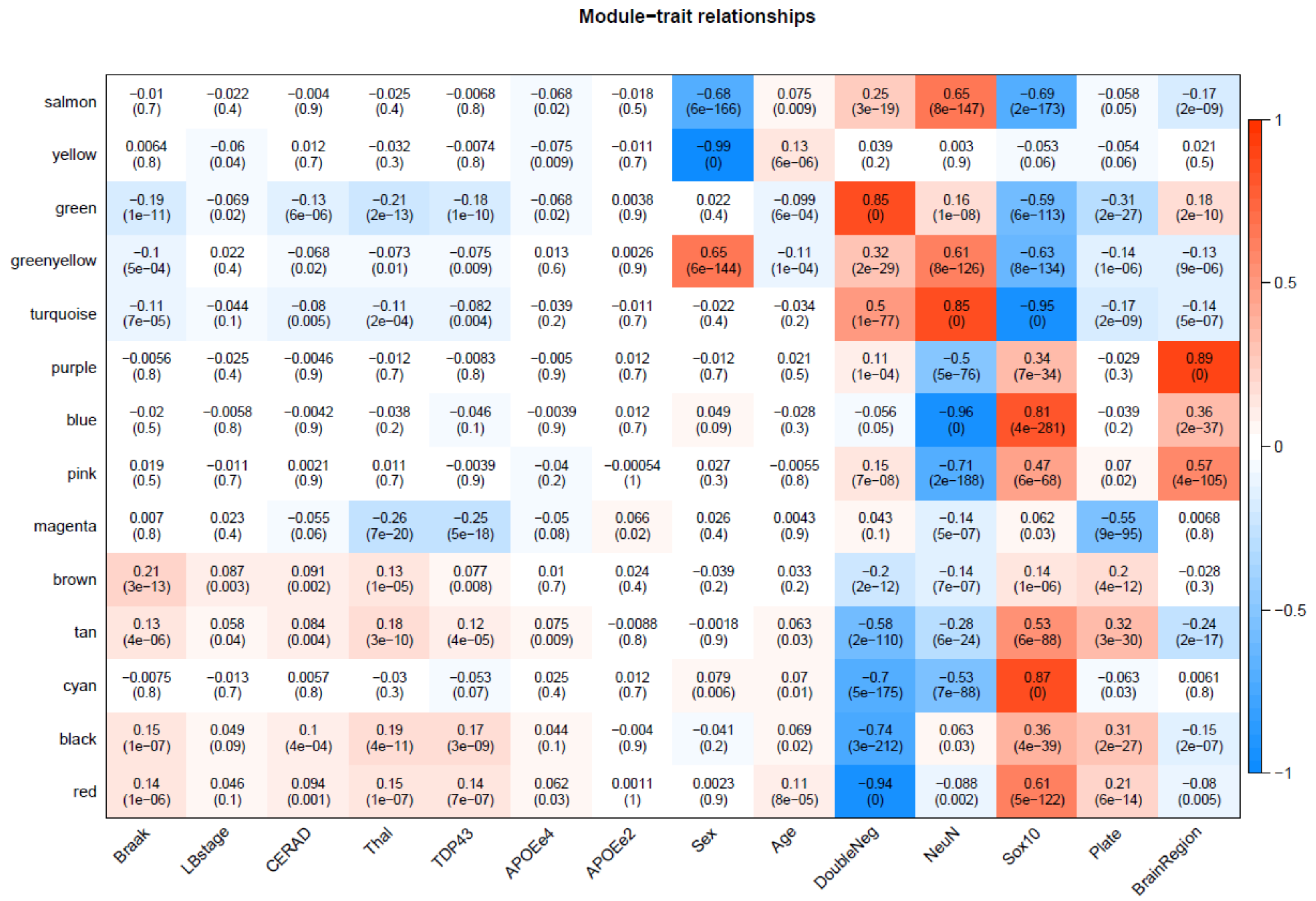


Figure 4.49 : Modules are hierarchically clustered based on calculated module eigengenes (representative of the DNA methylation values within each module). *The number of DNA methylation probes included in each module are indicated along the x-axis. The colour of each module was arbitrarily assigned.*



	Braak	LBstage	CERAD	Thal	TDP43	APOEe4	APOEe2	Sex	Age	DoubleNeg	NeuN	Sox10	Plate	BrainRegion
salmon	−0.01 (0.7)	−0.022 (0.4)	−0.004 (0.9)	−0.025 (0.4)	−0.0068 (0.8)	−0.068 (0.02)	−0.018 (0.5)	−0.68 (6e−166)	0.075 (0.009)	0.25 (3e−19)	0.65 (8e−147)	−0.69 (2e−173)	−0.058 (0.05)	−0.17 (2e−09)
yellow	0.0064 (0.8)	−0.06 (0.04)	0.012 (0.7)	−0.032 (0.3)	−0.0074 (0.8)	−0.075 (0.009)	−0.011 (0.7)	−0.99 (0)	0.13 (6e−06)	0.039 (0.2)	0.003 (0.9)	−0.053 (0.06)	−0.054 (0.06)	0.021 (0.5)
green	−0.19 (1e−11)	−0.069 (0.02)	−0.13 (6e−06)	−0.21 (2e−13)	−0.18 (1e−10)	−0.068 (0.02)	0.0038 (0.9)	0.022 (0.4)	−0.099 (6e−04)	0.85 (0)	0.16 (1e−08)	−0.59 (6e−113)	−0.31 (2e−27)	0.18 (2e−10)
greenyellow	−0.1 (5e−04)	0.022 (0.4)	−0.068 (0.02)	−0.073 (0.01)	−0.075 (0.009)	0.013 (0.6)	0.0026 (0.9)	0.65 (6e−144)	−0.11 (1e−04)	0.32 (2e−29)	0.61 (8e−126)	−0.63 (8e−134)	−0.14 (1e−06)	−0.13 (9e−06)
turquoise	−0.11 (7e−05)	−0.044 (0.1)	−0.08 (0.005)	−0.11 (2e−04)	−0.082 (0.004)	−0.039 (0.2)	−0.011 (0.7)	−0.022 (0.4)	−0.034 (0.2)	0.5 (1e−77)	0.85 (0)	−0.95 (0)	−0.17 (2e−09)	−0.14 (5e−07)
purple	−0.0056 (0.8)	−0.025 (0.4)	−0.0046 (0.9)	−0.012 (0.7)	−0.0083 (0.8)	−0.005 (0.9)	0.012 (0.7)	−0.012 (0.7)	0.021 (0.5)	0.11 (1e−04)	−0.5 (5e−76)	0.34 (7e−34)	−0.029 (0.3)	0.89 (0)
blue	−0.02 (0.5)	−0.0058 (0.8)	−0.0042 (0.9)	−0.038 (0.2)	−0.046 (0.1)	−0.0039 (0.9)	0.012 (0.7)	0.049 (0.09)	−0.028 (0.3)	−0.056 (0.05)	−0.96 (0)	0.81 (4e−281)	−0.039 (0.2)	0.36 (2e−37)
pink	0.019 (0.5)	−0.011 (0.7)	0.0021 (0.9)	0.011 (0.7)	−0.0039 (0.9)	−0.04 (0.2)	−0.00054 (1)	0.027 (0.3)	−0.0055 (0.8)	0.15 (7e−08)	−0.71 (2e−188)	0.47 (6e−68)	0.07 (0.02)	0.57 (4e−105)
magenta	0.007 (0.8)	0.023 (0.4)	−0.055 (0.06)	−0.26 (7e−20)	−0.25 (5e−18)	−0.05 (0.08)	0.066 (0.02)	0.026 (0.4)	0.0043 (0.9)	0.043 (0.1)	−0.14 (5e−07)	0.062 (0.03)	−0.55 (9e−95)	0.0068 (0.8)
brown	0.21 (3e−13)	0.087 (0.003)	0.091 (0.002)	0.13 (1e−05)	0.077 (0.008)	0.01 (0.7)	0.024 (0.4)	−0.039 (0.2)	0.033 (0.2)	−0.2 (2e−12)	−0.14 (7e−07)	0.14 (1e−06)	0.2 (4e−12)	−0.028 (0.3)
tan	0.13 (4e−06)	0.058 (0.04)	0.084 (0.004)	0.18 (3e−10)	0.12 (4e−05)	0.075 (0.009)	−0.0088 (0.8)	−0.0018 (0.9)	0.063 (0.03)	−0.58 (2e−110)	−0.28 (6e−24)	0.53 (6e−88)	0.32 (3e−30)	−0.24 (2e−17)
cyan	−0.0075 (0.8)	−0.013 (0.7)	0.0057 (0.8)	−0.03 (0.3)	−0.053 (0.07)	0.025 (0.4)	0.012 (0.7)	0.079 (0.006)	0.07 (0.01)	−0.7 (5e−175)	−0.53 (7e−88)	0.87 (0)	−0.063 (0.03)	0.0061 (0.8)
black	0.15 (1e−07)	0.049 (0.09)	0.1 (4e−04)	0.19 (4e−11)	0.17 (3e−09)	0.044 (0.1)	−0.004 (0.9)	−0.041 (0.2)	0.069 (0.02)	−0.74 (3e−212)	0.063 (0.03)	0.36 (4e−39)	0.31 (2e−27)	−0.15 (2e−07)
red	0.14 (1e−06)	0.046 (0.1)	0.094 (0.001)	0.15 (1e−07)	0.14 (7e−07)	0.062 (0.03)	0.0011 (1)	0.0023 (0.9)	0.11 (8e−05)	−0.94 (0)	−0.088 (0.002)	0.61 (5e−122)	0.21 (6e−14)	−0.08 (0.005)

Figure 4.50: Correlations between module eigengenes and traits of interest. *The module names are shown along the y-axis and the trait in which they were correlated with are shown along the x-axis. Correlation estimates are reported, with* p*-values in parentheses. Modules with a correlation* p*-value <0.05 (after controlling for other covariates) were selected for further analysis.*

Table 4.25: Several of the modules correlated with neuropathology remain significant after controlling for covariates. *The significantly correlated modules (Bonferroni p < 0.05/28 = 0.002) were taken forward for regression analysis to identify if the associations with the traits of interest remained after controlling for covariates (age, sex, cell proportions and batch).*

Module	Trait	BETA	SE	P
red	Braak NFT Stage	5.20e-04	1.50e-04	4.70e-04
red	CERAD density	6.50e-04	2.30e-04	4.30e-03
red	Thal Phase	3.50e-04	1.80e-04	0.06
red	TDP-43 Status	1.40e-03	7.20e-04	0.05
red	Braak LB Stage	1.20e-04	1.30e-04	0.37
black	Braak NFT Stage	9.50e-04	3.40e-04	5.00e-03
black	CERAD density	6.80e-04	5.30e-04	0.2
black	Thal Phase	6.00e-04	4.10e-04	0.15
black	TDP-43 Status	3.30e-03	1.70e-03	0.05
black	Braak LB Stage	4.60e-05	3.10e-04	0.88
tan	Braak NFT Stage	8.20e-04	3.40e-04	0.02
tan	CERAD density	5.20e-04	5.30e-04	0.33
tan	Thal Phase	7.50e-04	4.10e-04	0.07
tan	TDP-43 Status	1.10e-05	1.70e-03	0.99
tan	Braak LB Stage	3.90e-04	3.10e-04	0.22
brown	Braak NFT Stage	2.70e-03	5.70e-04	2.40e-06
brown	CERAD density	1.20e-03	9.00e-04	0.18
brown	Thal Phase	1.20e-03	6.70e-04	0.08
brown	TDP-43 Status	-5.50e-04	2.90e-03	0.85
brown	Braak LB Stage	6.50e-04	5.20e-04	0.21
turquoise	Braak NFT Stage	-5.30e-04	1.00e-04	1.30e-07
turquoise	CERAD density	-5.10e-04	1.60e-04	1.3e-03
turquoise	Thal Phase	-3.10e-04	1.20e-04	0.01
turquoise	TDP-43 Status	-1.10e-03	4.90e-04	0.02
turquoise	Braak LB Stage	-8.70e-05	9.20e-05	0.34
greenyellow	Braak NFT Stage	-5.80e-04	1.50e-04	1.90e-04
greenyellow	CERAD density	-4.80e-04	2.40e-04	0.05
greenyellow	Thal Phase	-3.40e-04	1.90e-04	0.07
greenyellow	TDP-43 Status	-1.40e-03	7.50e-04	0.06
greenyellow	Braak LB Stage	1.40e-04	1.40e-04	0.31
green	Braak NFT Stage	-1.40e-03	2.40e-04	2.20e-08
green	CERAD density	-1.20e-03	3.90e-04	1.90e-03
green	Thal Phase	-9.60e-04	3.00e-04	1.30e-03
green	TDP-43 Status	-3.20e-03	1.20e-03	8.90e-03
green	Braak LB Stage	-3.20e-04	2.30e-04	0.15

4.4.12.1 Pathway analysis of the significant modules

To help us better understand the functional role of the modules I performed gene ontology (GO) pathway analysis of genes annotated to DNAm sites in each module. I identified significant pathways for the tan (41 pathways; see **Figure 4.51**), black (13 pathways; see **Figure 4.52**), greenyellow (23 pathways; see **Figure 4.53**), turquoise (14 pathways; **see Figure 4.54**) and green modules (2 pathways; see **Figure 4.55**). Several of the pathways are relevant in the context of dementia and other neurobiological functions.

In the tan module several immune related pathways were identified including 'lymphocyte activation involved in immune response', 'B cell differentiation' 'B cell activation' and 'humoral immune response' (**Figure 4.51**). Mounting evidence suggests the immune system plays a role in the aetiology of AD and other dementias (Heppner, Ransohoff, & Becher, 2015). For example, humoral immune response (which relates to antibody production) by B cells to Aβ has been extensively studied in relation to AD; it has been identified that B cells from the blood of AD patients secrete antibodies that specifically recognise Aβ (Gaskin, Finley, Fang, Xu, & Fu, 1993). In addition several other AD related pathways were identified such as those involved in cholesterol metabolism (Picard *at al.*, 2018), gliogenesis (Rusznák, Henskens, Schofield, Kim, & Fu, 2016) and oligodendrocyte differentiation (Desai *at al.*, 2010). The black module was enriched for several pathways which have been implicated in dementia including 'regulation of cyclin-dependent protein serine/threonine kinase activity' and 'sterol transport'. Serine/threonine kinases are thought to be involved in signalling pathways mediated by familial AD mutations in *APP* and PSEN (Ryder, Su, & Ni, 2004). Additionally, *APOE*-e4 has been found to disrupt sterol metabolism in AD cases but not in controls (Bandaru *at al.*, 2009). The greenyellow module was enriched for pathways involving brain cells such as 'glial cell development',' ensheathment of neurons',' axon ensheathment' and 'myelination'. These pathways are disrupted in dementia patients. For example, myelin and the oligodendrocytes that produce it are damaged by Aβ, and this has been associated with AD (Bartzokis, 2004). Several immune related pathways were also enriched in the greenyellow, turquoise and green modules such as 'regulation of humoral immune response', 'defence response to bacterium', 'T cell receptor complex', 'platelet activation’, and ‘myeloid leukocyte differentiation’ further supporting a role for the immune system in dementia.

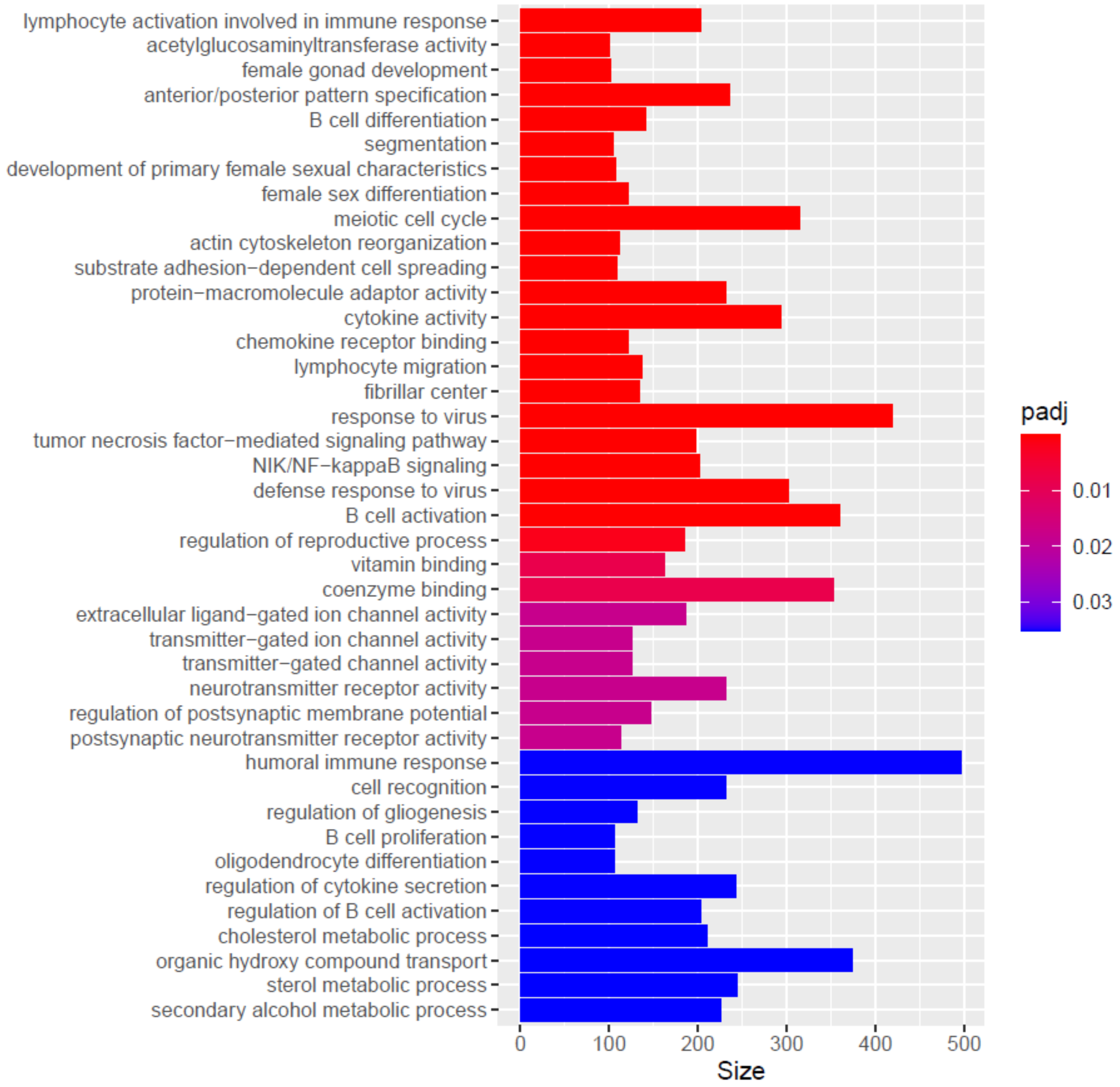


Figure 4.51: Pathway analysis of the tan module. *The y-axis is the gene ontology pathway. The x-axis is size, which represents the number of genes included in the pathway.*

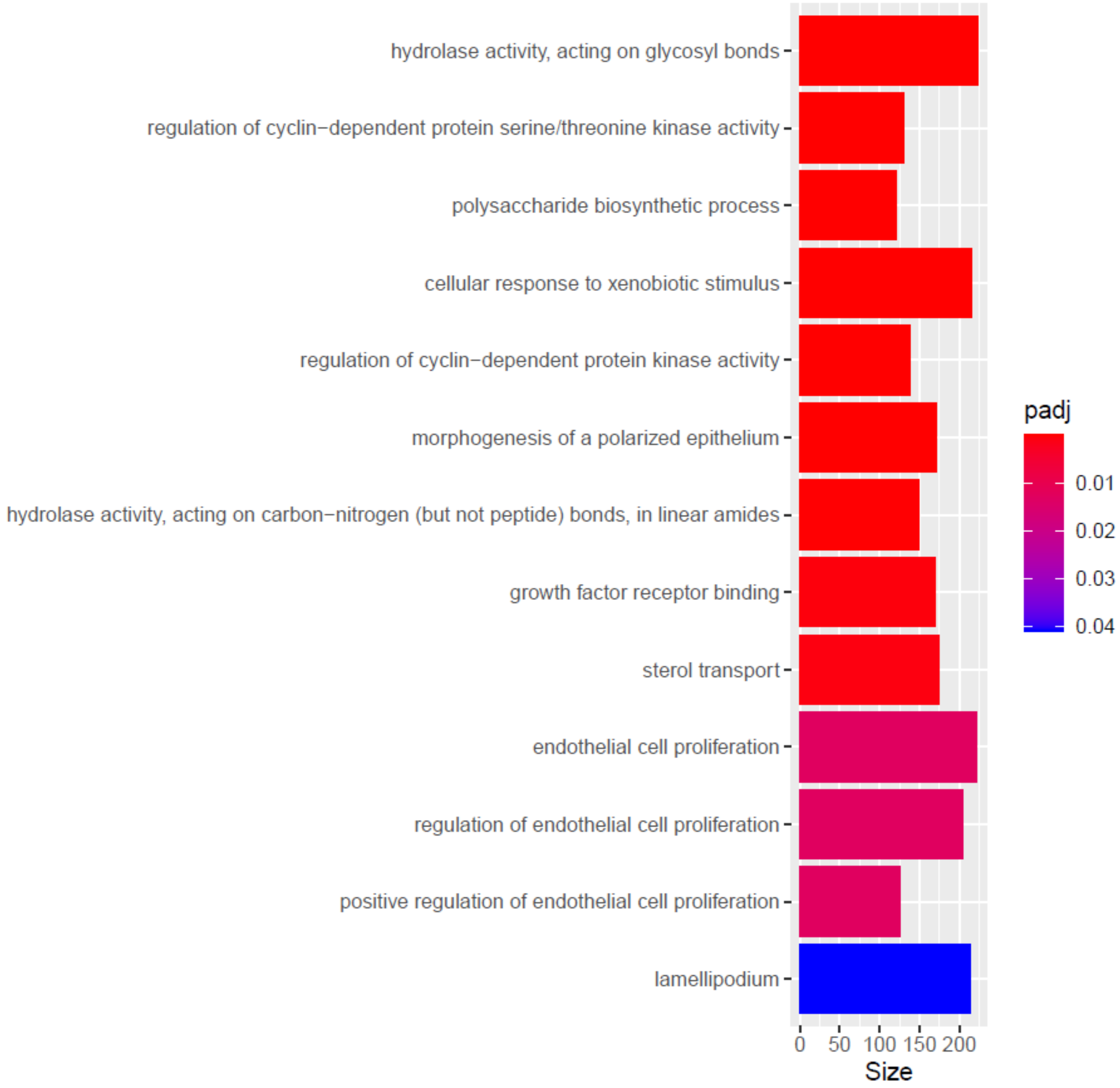


Figure 4.52: Pathway analysis of the black module. *The y-axis is the gene ontology pathway. The x-axis is size, which represents the number of genes included in the pathway.*

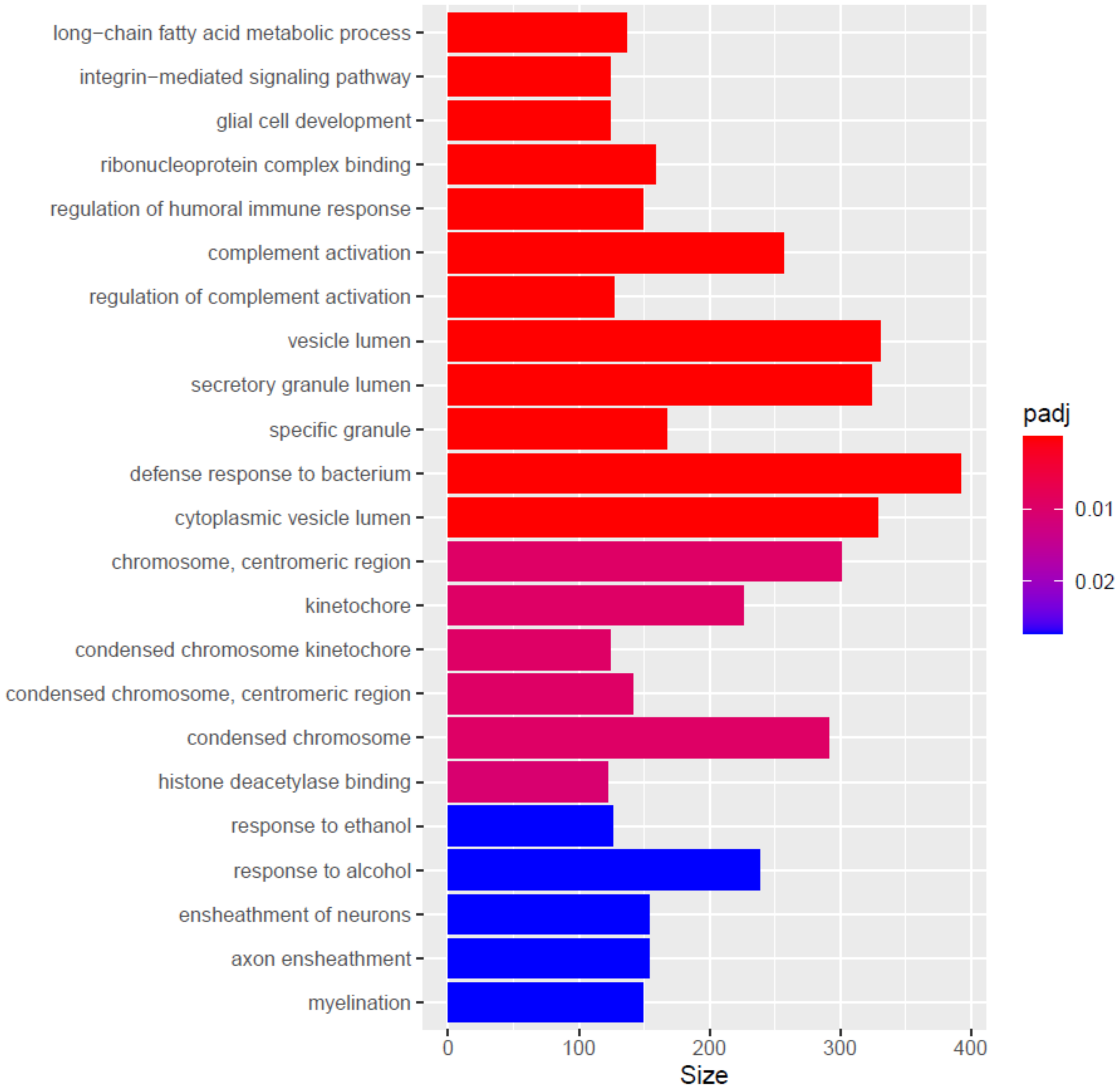


Figure 4.53: Pathway analysis of the greenyellow module. *The y-axis is the gene ontology pathway. The x-axis is size, which represents the number of genes included in the pathway.*

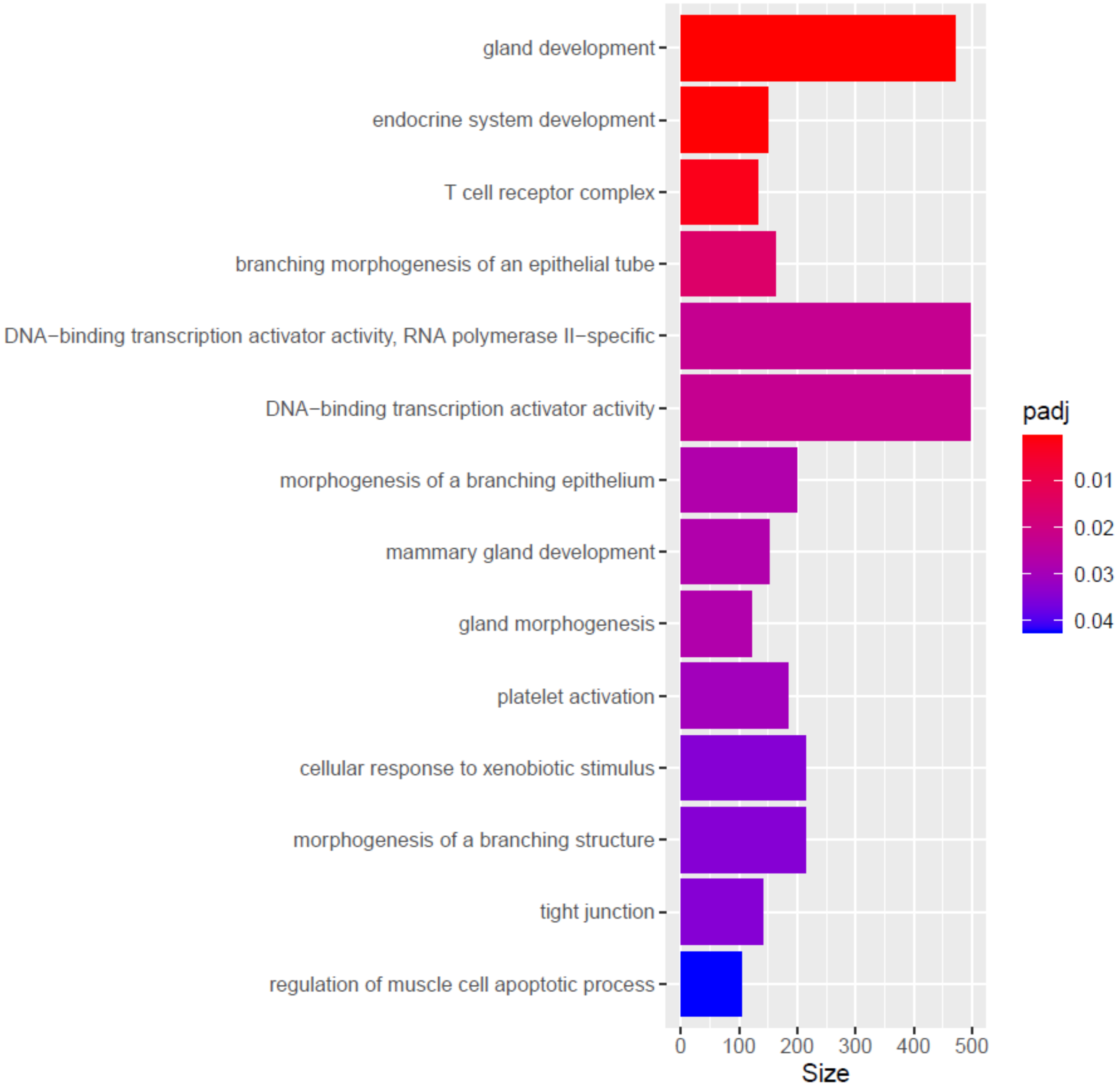


Figure 4.54: Pathway analysis of the turquoise module. *The y-axis is the gene ontology pathway. The x-axis is size, which represents the number of genes included in the pathway.*

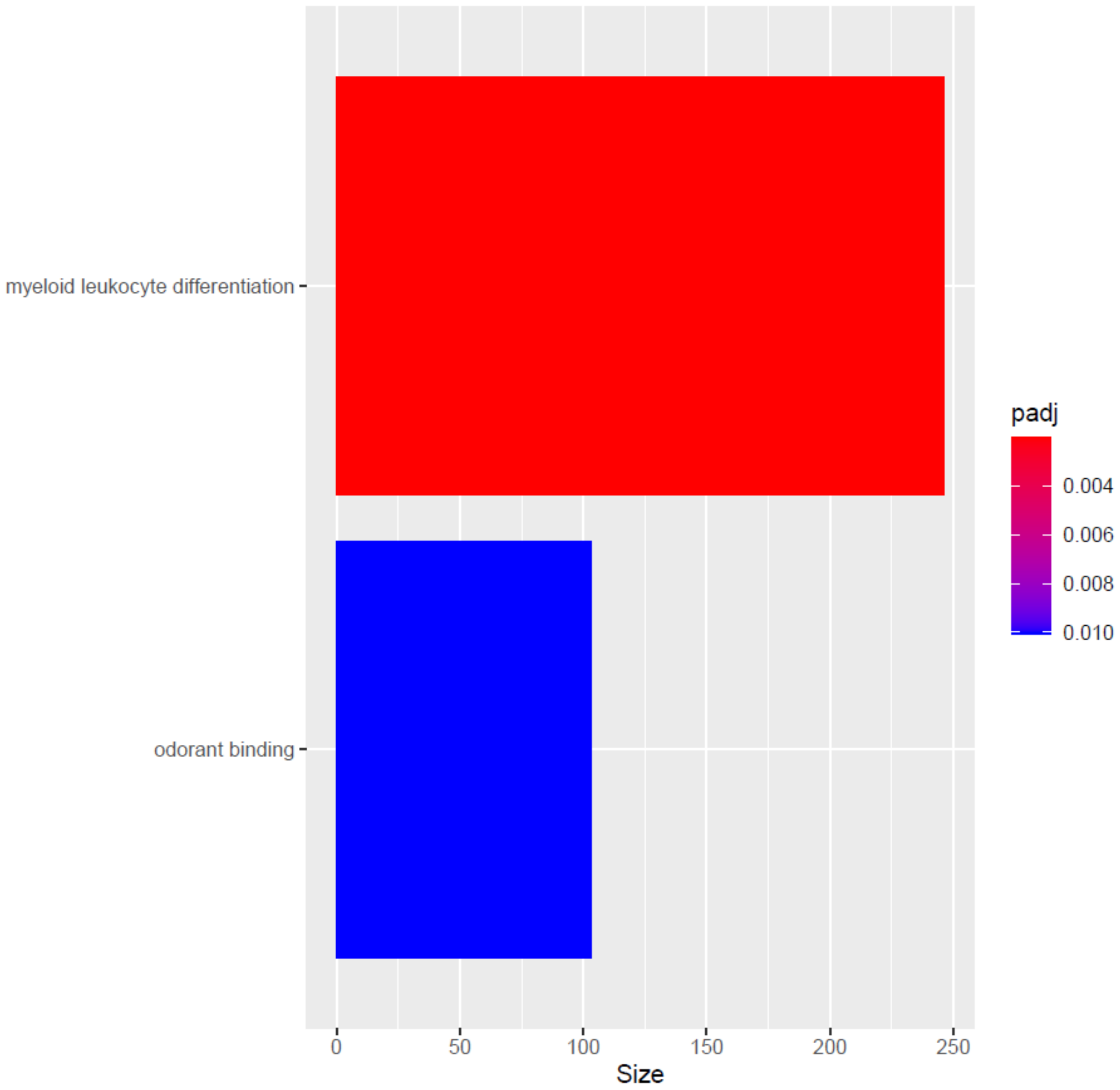


Figure 4.55: Pathway analysis of the green module. *The y-axis is the gene ontology pathway. The x-axis is size, which represents the number of genes included in the pathway.*

4.5 Discussion

4.5.1 Overview of results

In this chapter I examined the association between DNAm and five different neuropathology measures (Braak NFT stage, Thal phase, CERAD density, Braak LB stage and TDP-43 status) utilising DNAm data from 1221 samples from two cortical regions (PFC and OCC). I identified a number of DMPs and DMRs which were associated neuropathology. The majority DMPs showed consistent patterns of neuropathology-associated methylomic variation across the two cortical regions, with very few interactions being identified between neuropathology and brain region. This is likely due to the gained power when using the two regions collectively but is also likely to reflect cortex-wide effects as opposed to cortex-specific effects of neuropathology. Many of the DMPs identified were annotated to EPIC array specific probes which demonstrates the utility of the newer platform and the advantage of the increased power in comparison to the 450K array that has been used in previous studies.

Many of the DMPs were annotated to genes which have previously been implicated in dementia. This includes several DMPs and DMRs annotated to *HOXA* region which has consistently been reported to be hypermethylated in AD (Gasparoni *at al.*, 2018; R. G. Smith *at al.*, 2018; R. Smith *at al.*, 2021). It has been recognised that the *HOXA* gene cluster is involved in neuronal development, neuronal circuit organisation and the regulation of post mitotic neurons (Lizen *at al.*, 2017; Philippidou & Dasen, 2013). Methylomic variation of the HOX region has been associated with several neurodegenerative diseases including PD, Huntington's disease and C9ORF72-related dementia (Finch *at al.*, 2017; Hoss *at al.*, 2014; Labadorf *at al.*, 2015). Several DMPs and DMRs were associated with immune related genes (e.g. *TNFRSF1A* and *OSCAR*) and in combination with the pathway analysis of the neuropathology-associated WGCNA modules highlighting an abundance of immune pathways, these findings provide further evidence that immune regulation plays a role in the aetiology of AD and other dementias (Heppner *at al.*, 2015).

DMPs and DMRs annotated to *HOXA* and several other genes including *SH2B2* and *BCAR1* were associated with multiple AD neuropathology measures (Braak NFT Stage, Thal phase and CERAD density). Additionally, DMPs from all five neuropathology measures were characterised by a similar direction of effect, several overlapping DMPs were identified in the analysis including all neuropathology measures, and the significance of all DMPs diminished when including other neuropathology measures as covariates. Collectively, these findings suggest that DMPs and DMRs are driven by some aspect of neurodegeneration as opposed to the individual neuropathology measures. This suggests that regardless of dementia subtype, there are similar mechanisms involved in disease pathogenesis. These findings could be explained by the pleiotropy between different dementia subtypes. For example, SNPs within *HLA, MAPT* and *APOE* all contribute to increased risk for FTD, AD and PD (Ferrari *at al.*, 2017). Additionally, the fEOAD genes (*APP*, *PSEN1* and *PSEN2*) are found in PD cases suggesting genes involved in Mendelian neurodegenerative diseases also display pleiotropic effects (Ibanez *at al.*, 2018). There have been several studies suggesting that PD, DLB and AD share underlying mechanisms and there are strong genetic correlations between these diseases (Desikan *at al.*, 2015; Guerreiro *at al.*, 2016). Previous EWAS have also identified methylomic similarities in neurodegenerative diseases. Sanchez-Mut *at al.* (2016) focused on identifying common pathways involved in PD, DLB, AD and down-syndrome cases and identified multiple DMRs associated with disease. In their subsequent pathway analysis they reported significant over-representation in pathways related to brain function and immune response across all disorders tested. They also identified that *ANK1* was differentially expressed in DLB, which corresponds with results from previous AD EWAS indicating there is hypermethylation of *ANK1* (De Jager *at al.*, 2014; Gasparoni *at al.*, 2018; Lunnon *at al.*, 2014; Smith *at al.*, 2019), although recent research reported that the hypermethylation of *ANK1* in DLB could be confounded by AD pathology. Although it is possible to stratify by specific dementia subtypes, dementia in older people is usually a consequence of multiple pathologies (Kapasi, DeCarli, & Schneider, 2017; Thomas *at al.*, 2020) and this needs to be considered in EWAS of neurodegenerative diseases.

The evidence for pleiotropy suggests that common pathological mechanisms likely underlie neurodegenerative disorders. Although the proteins and specific brain regions involved in the aetiology of neurodegenerative diseases differ, in all cases the progressive accumulation of these deposits ultimately leads to neuronal cell death and brain atrophy (Ross & Poirier, 2004). Our results suggest that the pathogenesis of different types of dementia may be more similar than first hypothesised. This is perhaps the most important finding as EWAS generally focus on specific neuropathology measures depending on the dementia subtype they are investigating. This highlights the benefit of using the BDR dataset which has multiple neuropathology measures for each individual. Previous EWAS of dementia have enabled us to explore the molecular consequences of disease pathology however, previously findings should be interpreted with caution in relation to disease-specific claims. Of note, although the findings suggest there are common underlying mechanisms in neurodegeneration there may still be some specific methylomic differences between the neuropathology measures, although we were limited by power in these analyses. Moreover, I used neuropathology measures for the analyses and not definitive diagnoses of the dementia subtypes. If I had excluded definitive FTD, PD and DLB cases in the AD analyses and vice versa it is possible we may have identified some differences between the disorders.

4.5.2 Limitations

There are several limitations to this study. For one, bulk tissue was used which contains several cell types. It has been well established that cell proportions are altered in AD and dementia, with a reduction of neuronal and an upregulation of glial cells in comparison to cognitively normal controls (Gómez-Isla *at al.*, 1997). Although I have controlled for cell proportions using an algorithm based on nuclei-sorted DNAm data, the optimal approach would be to profile purified cell populations or utilise single-cell technologies in future studies.

I used the UCSC annotation provided by Illumina to identify the gene relating to each DMP which can lead to the annotation of overlapping genes, or no gene annotation, making is hard to establish the gene of interest. The data generated in **Chapter 5** section **5.4.16** (which characterised the relationship between DNA methylation and gene expression) and Hi-C data could potentially be used to link regulatory data to actual genes. It has revealed there are contacts between distant genomic regions within the same or across different chromosomes, which has many implications for gene regulation. Hi-C assesses this 3D chromatin conformation and can be used to elucidate how the spatial organisation of DNA affects gene regulation by identifying interactions and topologically associating domains (Kikuchi *at al.*, 2019; Pombo &

Dillon, 2015). These data could be combined to develop a more functionally relevant manifest for EWAS data.

As the distribution of samples across the neuropathology groups differed, the power to identify significant associations varied across the analyses. Most notably, LB pathology had very few cases in the latest stages of pathology. The correlations between the effects sizes of LB pathology DMPs and the other neuropathology measures were weak. This could suggest aetiological differences in LB pathology but we cannot draw these conclusions as there was not enough power to detect robust associations in this analysis. Future studies should incorporate a more even distribution of pathology for all measures included.

Although the experiment wide significance threshold (p<9e-08) was empirically derived via the simulation of null DNAm datasets using EPIC array data (Mansell *at al.*, 2019), it perhaps is too conservative due to the pleiotropy between neurodegenerative disorders; they are not fully independent (Desikan *at al.*, 2015; Ferrari *at al.*, 2017). The notion that this threshold is too stringent is further supported by the consistency in the direction of effect across the findings in this chapter with the findings from the recent meta-analysis conducted by our group (R. Smith *at al.*, 2021), which highlights how robust the associations are. Several of the suggestive significant (p<5e-05) sites in the meta-analysis did not surpass the experiment-wide significance threshold in their study but were found to be significant in our analysis and vice versa. This suggests they are likely to be true associations which were reported as non-significant (false negatives).

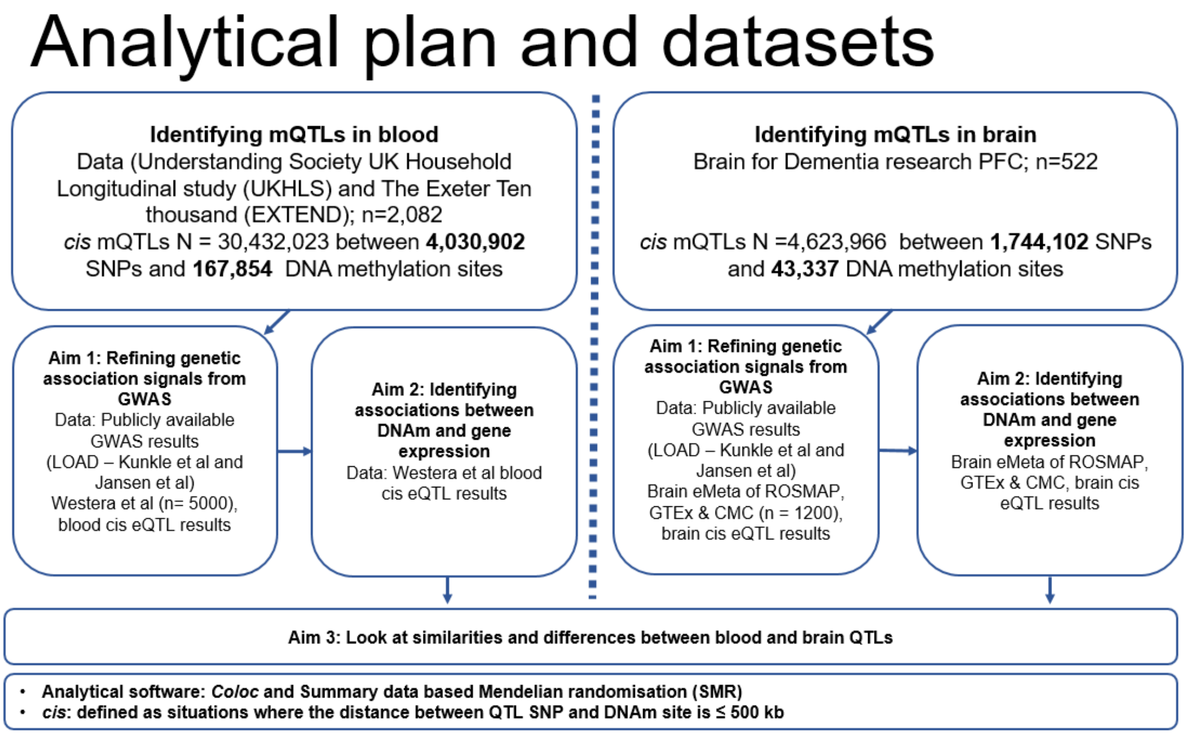
In the WGCNA analysis I did not regress out any covariates as they may offer biological insights. However, this means the modules identified may be driven by covariates with stronger associations such as derived cell proportions as opposed to neuropathology or by technical factors such as experimental batch. Removing technical covariates in future analysis would ensure that they are not driving the module generation. However, several of the pathways identified were relevant in the context of neurodegenerative disease, indicating that experimental batch had limited influence on the module generation.

One key limitation of epigenetic studies is the issue of causality. It is not possible to disentangle if the DMPs identified are driving disease pathogenesis, or if they are a consequence of the disease. However, since neuropathology in the OCC is low and the effects were consistent across brain regions, this indicates there is some causal aspect of disease rather than a consequence of the presence of neuropathology. There is also a chance that they are a result of other uncontrolled environmental factors such as medication.

4.5.3 Conclusion

To my knowledge, this is the most systematic dementia EWAS conducted to date. A number of novel loci which have not previously been implicated in dementia have been identified which warrant future exploration to help us better understand the aetiology of neurodegenerative disease. This study highlights the importance of utilising multiple neuropathology measures to help us better understand disease pathogenesis and suggests that independent pathologies may not currently provide disease specific information if other neuropathology measures or disease subtypes are not controlled for. This limits the usefulness of neuropathology-specific EWAS however it allows us to better understand the molecular mechanisms involved in general neurodegeneration. Future studies with multiple neuropathology measures should conduct similar analysis to validate these results.

5 Methylation quantitative trait loci (mQTL) analysis and summary data-based Mendelian randomisation (SMR)



Analytical plan and datasets

Identifying mQTLs in blood
Data (Understanding Society UK Household Longitudinal study (UKHLS) and The Exeter Ten thousand (EXTEND); n=2,082
cis mQTLs N = 30,432,023 between **4,030,902** SNPs and **167,854** DNA methylation sites

Identifying mQTLs in brain
Brain for Dementia research PFC; n=522

cis mQTLs N =4,623,966 between **1,744,102** SNPs and **43,337** DNA methylation sites

Aim 1: Refining genetic association signals from GWAS
Data: Publicly available GWAS results
(LOAD – Kunkle et al and Jansen et al)
Westera et al (n= 5000), blood cis eQTL results

Aim 2: Identifying associations between DNAm and gene expression
Data: Westera et al blood cis eQTL results

Aim 1: Refining genetic association signals from GWAS
Data: Publicly available GWAS results
(LOAD – Kunkle et al and Jansen et al)
Brain eMeta of ROSMAP, GTEx & CMC (n = 1200), brain cis eQTL results

Aim 2: Identifying associations between DNAm and gene expression
Brain eMeta of ROSMAP, GTEx & CMC, brain cis eQTL results

Aim 3: Look at similarities and differences between blood and brain QTLs

- **Analytical software: *Coloc* and Summary data based Mendelian randomisation (SMR)**
- ***cis*: defined as situations where the distance between QTL SNP and DNAm site is ≤ 500 kb**

Figure 5.1: Analytical plan and datasets for Chapter 5: Methylation quantitative trait loci (mQTL) analysis and summary data-based Mendelian randomisation (SMR).

5.1 Introduction

As the majority of GWAS variants associated with complex traits such as LOAD reside in non-coding regions – i.e. they do not index transcriptional changes which influence protein structure - it has been challenging to identify how they impact upon disease aetiology (Amlie-Wolf *at al.*, 2018; Giambartolomei *at al.*, 2018). It is hypothesised that GWAS variants act through gene regulation, which is supported by evidence showing that they are enriched in regulatory domains including enhancers and regions of open chromatin (Kikuchi *at al.*, 2019; Marzi *at al.*, 2018). A potential approach to explore the mechanisms by which non-coding risk variants regulate gene expression is through the integration of datasets that measure the association of sequence variation with molecular phenotypes, including DNA methylation quantitative trait loci (mQTLs; where a SNP is associated with DNA methylation [DNAm]) and gene expression quantitative trait loci (eQTLs; where a SNP is associated with gene expression). Quantitative trait loci with larger effects are predominantly *cis* acting, and are enriched in regulatory domains, supporting the hypothesis that they play a role in gene regulation. In addition, *cis*-mQTLs have been shown to be enriched amongst GWAS-loci, including LOAD GWAS variants (Min *at al.*, 2020). Characterising the relationship between genetic, epigenetic and transcriptomic variation can increase our understanding of the mechanisms driving disease pathogenesis in complex disease phenotypes including LOAD.

One method to characterise the relationship between genetic and epigenetic (or transcriptomic) variation is Bayesian-colocalisation. Bayesian-colocalisation explores whether two traits (i.e. disease and a given regulatory mark) are consistent with a shared genetic variant within a genomic region. Giambartolomei and colleagues (2014) developed a Bayesian test for colocalisation between multiple genetic association studies which utilises summary statistics. The method is Bayesian as it considers all the possible configurations of causal variants for the two traits and then calculates the support for each scenario using a Bayes factor (Giambartolomei *at al.*, 2014). The method is based on the assumption that there is a maximum of one causal variant per trait in each region and provides posterior probabilities (i.e. the probability of an event occurring after taking into consideration prior information) which can be easily interpreted. Bayesian-colocalisation accumulates evidence for five different mutually exclusive hypotheses (H_i):

- H_1) There is no association with either trait
- H_2) There is an association with trait one but not with trait two
- H_3) There is an association with trait two, but not with trait one
- H_4) There is an association with trait one and trait two but these are two independent SNPs
- H_5) There is an association with trait one and trait two, and there is a single causal SNP.

This method can be used to formally test if proximally located DNAm sites (or expressed genes) are influenced by the same causal variant. Hannon and colleagues (2018) used this method to better understand the relationship between genetic variation and DNAm in *cis*, identifying a complex relationship between the two. For example, they found that DNAm sites which co-localised with another site (i.e. where they were characterised by shared genetic effects) also co-localised with a median of three other DNAm sites.

There has been recent interest in utilising QTLs to help interpret the functional consequences of common genetic variation, particularly as the most proximally located gene to a lead GWAS SNP is not necessarily the ‘causal’ gene for disease. These methods aim to identify differential DNAm or gene expression associated with a trait due to pleiotropy (where a genetic variant has direct effects on both a trait and a molecular marker of gene regulation). Colocalisation itself – i.e. where the same genetic variant is causal for two traits - is not sufficient to definitively infer a pleiotropic relationship. This is important since a SNP could be associated with both DNAm and a phenotype, but this may reflect a situation where the top associated *cis* QTL is in linkage disequilibrium (LD; the correlation structure between proximal variants) with two independent causal variants – one SNP affecting DNAm and the other associated with phenotypic variation (see **Figure 5.2**). This is an active area of research and current methods which leverage QTLs to refine GWAS signals include Mendelian randomisation (MR) approaches including Summary-data based Mendelian Randomisation (SMR).

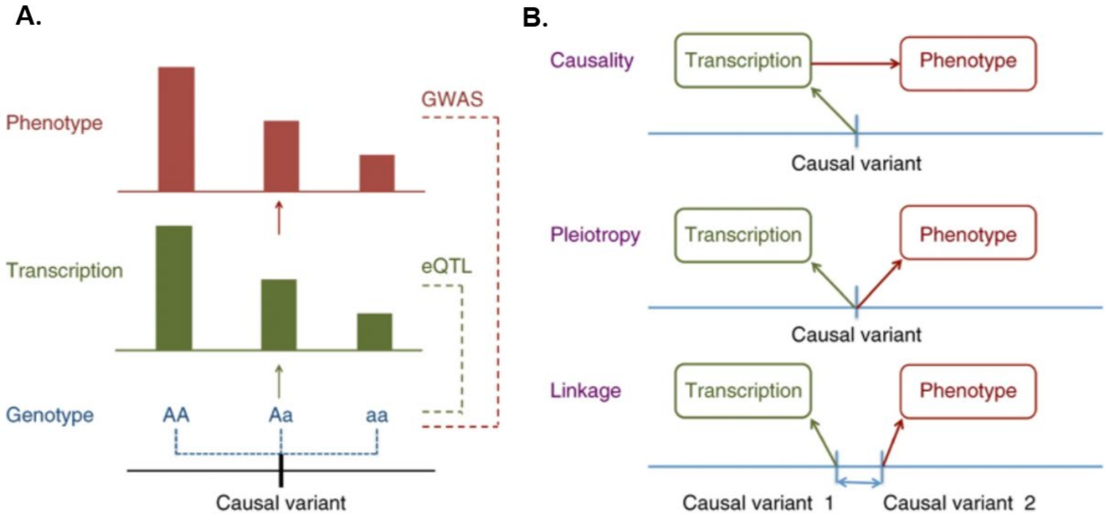


Figure 5.2 Association between gene expression and phenotypes via genetic control. ***(A)*** *A hypothetical model of causality where differences in phenotypic variation is driven by genetic variation, which is mediated through gene expression.* ***(B)*** *Three possible explanations for the observed association between a trait and gene regulation though genetic variation: causality (where the effect a genetic variant has on a trait is mediated by a molecular marker of gene regulation); pleiotropy (where a genetic variant has direct effects on both a trait and a molecular marker of gene regulation); and linkage (where there are two distinct variants in LD where one of these variants has an effect on transcription and the other on the phenotype). Figure taken directly from Zhou at al., (2016).*

SMR was developed to identify whether there is pleiotropy between genetic variants, molecular markers (e.g. DNAm at a specific site or the expression of a specific gene) and complex traits or disease (Zhu *at al.*, 2016). This method is based on the premise of Mendelian randomisation (MR), which uses genetic variation as an instrumental variable to investigate the ‘causal’ relationships between phenotypes in observational data (Burgess, Dudbridge, & Thompson, 2016). The MR approach is usually used to test for the causative effect of a modifiable risk factor on health outcomes but in SMR the approach is used to test if the effect size of a SNP on a trait identified from a GWAS is mediated by the expression level of a gene (Zhu *at al.*, 2016). For example, if expression of a gene is influenced by a SNP (i.e. an eQTL) then there will be differences in gene expression among individuals who are homozygous (e.g. AA, aa; where ‘A’ and ‘a’ are the different alleles) or heterozygous (Aa) for specific genetic variants (e.g. see **Figure 5.2(A)**). If the expression levels have a significant effect on a trait then these differences will be observable in the three distinct genetic groups.

The aim of MR is to estimate the ‘causal’ effect for an exposure on an outcome using one or more genetic instruments as shown in **Figure 5.3.** X is the exposure (in MR this is usually a modifiable risk factor and in SMR this is a molecular marker of gene regulation such as gene expression or DNAm), Y is the outcome (which is the phenotype of interest, e.g. LOAD) and Z is the genetic instrument (e.g. one or more SNPs). β_{XY} (defined $\beta_{XY} = \beta_{ZY}/\beta_{ZX}$) is the effect size of X on Y (this is the slope of Y regressed onto the genetic value of X and is the relationship of interest in MR analysis), β_{ZX} is the effect of Z on X and β_{ZY} is the effect of Z on Y.

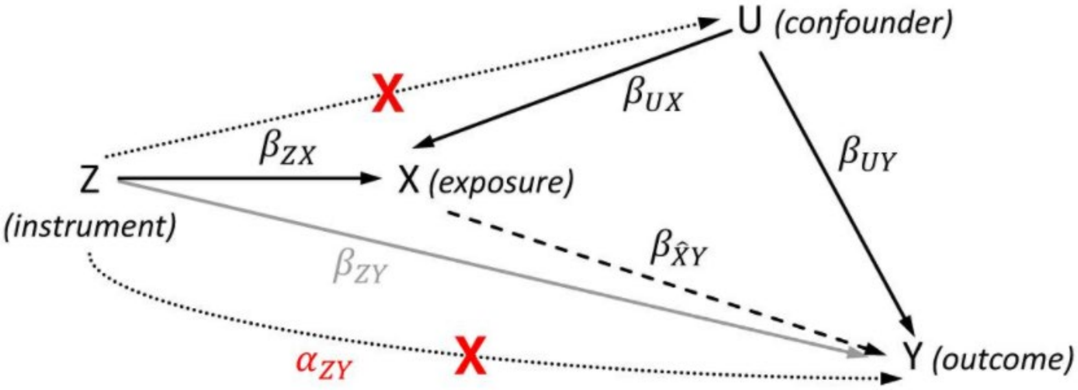


Figure 5.3: Mendelian Randomisation. *Y is the phenotype of interest (e.g. LOAD), X is the exposure (gene expression/ DNA methylation) and Z is a genetic variant (the genetic instrument). βXY is the effect size of X on Y (this is the slope of Y regressed onto the genetic value of X), βZX is the effect of Z on X and βZY is the effects of Z on Y. β_{UY} is the effect size of U on Y. In MR analysis βXY (defined βXY = βZY/βZX) is interpreted to be the effect of X on Y where there are non-genetic confounders. Figure taken directly from Teumer (2018).*

Zhu and colleagues (Zhu *at al.*, 2016) developed SMR to integrate summary-level data (e.g. effect sizes) from GWAS as well as eQTL and mQTL studies, utilising data across the genome. Under the assumption of causality (where the exposure is causally related to the outcome) or pleiotropy (where a genetic variant has direct effects on both a trait and a molecular marker of gene regulation), SMR analysis is analogous to MR. In SMR analysis β_{XY} corresponds to the association (either with a positive or negative sign) between the molecular marker and trait which is not influenced by non-genetic confounders. The SMR method for calculating the pleiotropic effect of the exposure on the outcome is shown in **Figure 5.4A**. Since only one instrumental variable (i.e. a specific SNP) is considered in each calculation, a Wald ratio (Wald, 1940) is used to estimate β_{XY} (referred to as bEWAS in **Figure 5.4**) which is simply the influence of the SNP-outcome effect divided by the SNP-exposure effect. There is a second step to the SMR process: the heterogeneity in dependent instruments (HEIDI) test, which is used to detect pleiotropy from linkage (**Figure 5.4B and Figure 5.4C**). β_{XY} estimated from SMR will be identical for any variant in LD with the causal variant suggesting there is a single causal variant and the association statistics remain the same irrespective of the genetic instrument used (see **Figure 5.4B**). Differences in β_{XY} implies there is a greater likelihood of linkage, rather than pleiotropy, suggesting there are distinct causal SNPs for the exposure and the trait (see **Figure 5.4C**). Of note, this approach cannot distinguish between linkage and pleiotropy if the two causal variants are in perfect LD as power is inversely proportional to the strength of the correlation between the two causal variants.

Simulations by Zhu and colleagues showed that MR and SMR analysis based on one genetic variant cannot distinguish between causality and pleiotropy regardless of whether or not if the effect of the genetic instrument on the exposure is direct or mediated by a latent variable (Zhu *at al.*, 2016). Therefore 'pleiotropy' is a more accurate term in SMR analysis than 'causality' as stated by the authors of the tool; this is to avoid misinterpretation of the results and is the term I have adopted throughout this Chapter.

SMR has been applied to multi-omics datasets. For example, Hannon and colleagues (2018) utilised the SMR tool to characterise the intricate relationship between genetic, epigenetic and transcriptomic variation in >60 traits, identifying ~1700 pleiotropic associations between 36 complex traits and >1200 DNAm sites. They also identified ~6,800 pleiotropic associations between >5,400 DNAm sites and the transcription of >1700 genes. SMR can be used to prioritize potentially functionally relevant genes within previously identified regions of association by GWAS and has the potential to highlight loci that currently do not have sufficient statistical power to reach genome-wide significance in GWAS.

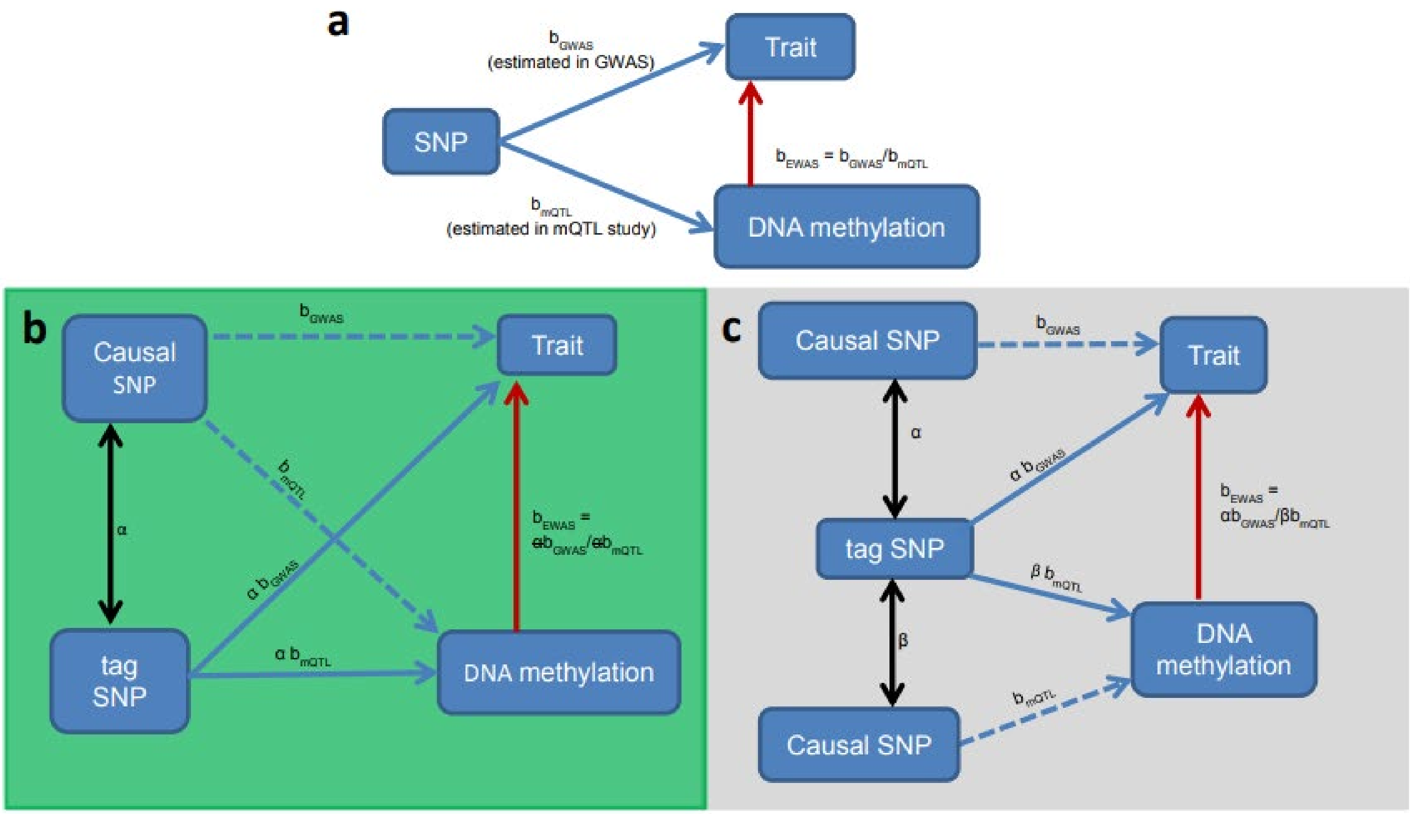


Figure 5.4: The Summary-data based Mendelian randomisation (SMR) approach. ***(A)*** *The first step of SMR analysis tests for an association between DNA methylation and a trait (e.g. Alzheimer's disease). Blue solid arrows represent known information (GWAS or mQTL results). Red arrows represent the relationship being derived.* ***(B)*** *and* ***(C)*** *represent stage two of SMR analysis which aims to distinguish between pleiotropy* ***(B)*** *and linkage* ***(C)****. Dashed blue arrows indicate the true causal associations estimated via "tag SNPs". "Tag SNPs" are highly correlated (shown by solid black arrows) with the "causal SNP" quantified as α or β. In* ***(B)*** *the same causal SNP is associated with both DNA methylation and the trait of interest; there is only one correlation statistic (α), which is cancelled out when estimating the effect bEWAS. In* ***(C)*** *there are two causal SNPs; one for DNA methylation and one for the trait, and therefore two correlations with the "tag SNP" (α and β), which do not cancel each other out. In this scenario the estimate of bEWAS will exhibit heterogeneity when different "tag SNPs" are tested, whereas in the scenario depicted in* ***(B)*** *the estimate of bEWAS will be consistent regardless of the choice of "tag SNP". SNP – single nucleotide polymorphism; GWAS - genome-wide association study; EWAS -epigenome wide association study. Figure taken directly and legend adapted from (Hannon at al., 2018).*

5.2 Chapter aims

The main aim of this chapter is to utilise methylation and expression quantitative trait loci to prioritise putative causal loci within large genomic regions associated with LOAD in both the human cortex and whole blood. The specific aims of this chapter are:

1. Building on work previously conducted within our group by Hannon and colleagues, I aim to generate two large mQTL datasets using EPIC array data in whole blood and the cortex.
2. Using Bayesian colocalisation I aim to formally test if proximally located DNAm sites are influenced by the same causal variant.
3. I will use the mQTL databases within the SMR framework to refine genetic association data from publicly available LOAD GWAS datasets in order to prioritize genes for future functional work.
4. Using the SMR approach I aim to identify pleiotropic relationships between DNAm and variable gene expression using publicly available whole-blood and cortex eQTL data.
5. Using the results generated from Aims 1 to 4, I will identify similarities and differences between the whole blood and cortex results.

5.3 Methods

5.3.1 Whole Blood mQTL data

The whole blood mQTL dataset was generated using data from the UK Household Longitudinal Study (UKHLS) and the Exeter 10,000 Study (EXTEND). Briefly, the UKHLS was established in 2009 and is a longitudinal panel survey of 40,000 UK households from England, Scotland, Wales and Northern Ireland (Buck & McFall, 2011)(see chapter 3 section **3.2.3** for more details). EXTEND is a research biobank funded by the National Institute for Health Research (NIHR) and is a population study of >10,000 individuals >18 years of age who live within 25 miles of Exeter (Devon, UK; https://exetercrfnihr.org/about/exeter-10000/).

5.3.1.1 DNA methylation data

DNAm data was generated within our laboratory for each of the cohorts. The EZ-96 DNA Methylation-Gold kit (Zymo Research; Cat No# D5007) was used to treat ~500 ng of DNA from each sample with sodium bisulfite. DNAm data were generated using the Illumina EPIC DNAm array as described previously (see Chapter 2 section **2.1.2**).

5.3.1.2 DNA methylation data pre-processing

Unless otherwise reported, all statistical analysis was conducted in the R statistical environment (version 3.5.2; https://www.r-project.org/). Raw data for both datasets were used, prior to any QC or normalisation, and processed using the *wateRmelon* (Pidsley *at al.*, 2013) and *bigmelon* (Gorrie-Stone *at al.*, 2019) packages. The stringent DNAm QC pipeline described in Chapter 3 section **3.2.4** was used for both datasets. Smoking score was estimated using the algorithm which is based on the DNAm profile at 183 sites known to be associated with smoking (Elliott *at al.*, 2014). For more details on the QC pipeline see Chapter 2 section **2.1.4.**

5.3.1.3 Genotyping and Imputation

5.3.1.3.1 UKHLS

UKHLS samples were genotyped using the Illumina Infinium HumanCoreExome BeadChip Kit as previously described (Prins *at al.*, 2017). This array contains a set of >250,000 highly informative genome-wide tagging SNPs as well as a panel of functional exonic markers, including a large proportion of low-frequency (MAF 1%–5%) and rare (MAF < 1%) variants. Genotype calling was performed with the *gencall* algorithm within Illumina GenomeStudio (https://emea.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html). Samples with matched DNAm data were selected and variants were refiltered prior to imputation. *PLINK1.9* was used for removing samples if: (1) they had > 5% missing data, (2) their genotype predicted sex using X chromosome homozygosity was discordant with their reported sex (excluding females with an F value > 0.2 and males with and F value < 0.8), (3) they had excess heterozygosity (>3 SD from the mean), (4) they were related to another individual in the sample (pi hat > 0.2), where one individual from each pair of related samples were randomly excluded, or (5) they were classed as non-European, determined by merging the UKHLS genotypes with data from HapMap Phase 3 (http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html), linkage disequilibrium (LD) pruning the overlapping SNPs such that no pair of SNPs within 1500 bp had $r^2 > 0.20$ and visually inspecting the first two genetic principal components along with the known ethnicities of the HapMap sample to define European samples. These data were then imputed with the 1000 Genomes phase 3 version reference panels SHAPEIT and minimac3. Best-guess genotypes were called, and variants were filtered to those with a minor-allele frequency >0.01 and an R^2 INFO score >0.8 (info score = information metric between 0-1 where 1 = a SNP which has been imputed to high certainty). Since variants were named in 1000 Genomes format using their locations (“chr:pos”) and variant type (SNP/INDEL), duplicate variants were also excluded. Principal components were calculated from the imputed genotype data via GCTA (a software tool for genome-wide complex-trait analysis) (Yang, Lee, Goddard, & Visscher, 2011). The imputed genetic variants were then filtered so that variants characterised by >5% missing values, a Hardy-Weinberg equilibrium p-value <0.001, a minor-allele frequency of <5%, and a minimum of five observations in each genotype group were excluded. Variants were named using their genomic locations (“chr:pos”) and variant type (SNP/INDEL) and therefore duplicate variants were excluded.

5.3.1.3.2 EXTEND

Genotyping was performed on the Illumina Infinium Global Screening Array (GSA) which incorporates 686,082 tagging SNPs and includes a panel of markers with established clinical associations based on the ClinVar database (Landrum *at al.*, 2016). The GSA has been optimised for genomic coverage and imputation performance in the five defined super populations (Africans, mixed Americans, East Asians, Europeans, South Asia). Genotype calling was performed using GenomeStudio (v2.0, Illumina) and QC was completed using *PLINK1.9* (Chang *at al.*, 2015). Individuals were excluded if either (1) they had > 5% missing data, (2) their genotype predicted sex using X chromosome homozygosity was discordant with their reported sex (excluding females with an F value > 0.2 and males with and F value < 0.8), (3) they had excess heterozygosity (>3 SD from the mean), (4) they were related to another individual in the sample (pi hat > 0.2), where one individual from each pair of related samples was randomly excluded or (5) they were classed as non-European, determined by merging the EXTEND genotypes with data from HapMap Phase 3 (http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html), linkage disequilibrium (LD) pruning the overlapping SNPs such that no pair of SNPs within 1500 bp had $r^2 > 0.20$ and visually inspecting the first two genetic principal components along with the known ethnicities of the HapMap sample to define European samples. The genetic data were then recoded as vcf files before uploading to the Michigan Imputation Server (Das *at al.*, 2016) (https://imputationserver.sph.umich.edu/index.html#!) which uses Eagle2 (Loh *at al.*, 2016) to phase haplotypes, and Minimac4 (https://genome.sph.umich.edu/wiki/Minimac4) with the most recent 1000 Genomes reference panel (phase 3, version 5). Imputed genotypes were then filtered with *PLINK2.0alpha* (https://www.cog-genomics.org/plink/2.0/), excluding SNPs with an R^2 INFO score < 0.5 and recoded as binary PLINK format. Proceeding with *PLINK1.9*, samples with >5% missing values, and SNPs with >2 alleles, >5% missing values, Hardy-Weinberg equilibrium $p < 0.001$, or a minor allele frequency of <5% were excluded. Since variants were named in 1000 Genomes format using their genomic locations ("chr:pos") and variant type (SNP/INDEL), duplicate variants were excluded.

5.3.1.4 Whole Blood mQTL dataset

Data from UKHLS and EXTEND were combined, keeping the overlapping samples between the genetic and DNAm data which passed the stringent QC criteria for both data types. Genetic data were combined keeping the SNPs in common between the two datasets. Genetic QC was conducted again on the combined datasets so that variants characterised by >5% missing values, a Hardy-Weinberg equilibrium p-value <0.001, a minor-allele frequency of <5% and <5 observations in each genotype group were excluded. Sample relatedness was calculated and related individuals were excluded (relatedness exclusion criteria: pi hat > 0.2). Principal components (PCs) were recalculated using the *--pca* flag in *PLINK1.9* (Chang *at al.*, 2015) for inclusion as covariates in QTL analyses (10 PCs included). Raw DNAm data (in the form of idat files) for the samples in each cohort which passed the initial DNAm QC and the combined genotyping QC were read into R. A brief QC for the DNAm data was conducted again: the *pfilter* function in *wateRmelon* was used to exclude samples with >1 % of probes with a detection P value > 0.05 and probes with >1 % of samples with detection P value > 0.05 and data was normalised using *dasen*. Cross-hybridizing probes, probes with a common SNP (European population minor-allele frequency > 0.01) within 10 base pairs of the DNAm site or a single base extension and sex chromosomes were excluded from the QTL analysis (McCartney *at al.*, 2016). The final QC'd set of data included 2082 samples, of which 44% were male, the age range was between 19-98 years and a there was a mean age of 57.49 years (UKHLS: N=1,099, 41% male, age range =28-98 years, mean age = 58.50 years; EXTEND: N= 983, 47.5% male, age range =19-80 years, mean age = 56.42 years). 765,013 DNAm probes and 5,359,678 genetic variants passed QC for analysis.

5.3.2 Cortex mQTL dataset

5.3.2.1 DNA methylation data

Cortex mQTLs were generated using post-mortem prefrontal cortex (PFC) samples from the Dementia Research (BDR) cohort (see Chapter 3, section **3.2.2**). Briefly, the BDR cohort was established with the aim of generating a large comprehensive neuropathological dataset from multiple brain banks using standardised procedures to enable the investigation of dementia through detailed phenotypic and multi-omics datasets (Francis, Costello, & Hayes, 2018). For more details on this cohort see Chapter 4 section **4.3.1.**

5.3.2.2 Genotyping and Imputation

DNA was isolated using a standard phenol chloroform method on 100 mg of cortex tissue. DNA quality was assessed using the Agilent 2200 TapeStation DNA integrity number and quantified using Nanodrop 3300 spectrometry. Genotyping was performed on the NeuroChip array which is a custom Illumina genotyping array with an extensive genome-wide backbone (*n* = 306,670 variants) and custom content covering 179,467 variants specific to neurological diseases (Blauwendraat *at al.*, 2017). Genotype calling was performed using GenomeStudio (v2.0, Illumina) and quality control (QC) was completed using PLINK1.9 (Chang *at al.*, 2015). Individuals were excluded if either 1) they had > 5% missing data, 2) their genotype predicted sex using X chromosome homozygosity was discordant with their reported sex (excluding females with an F value > 0.2 and males with and F value < 0.8), 3) they had excess heterozygosity (>3 SD from the mean), 4) they were related to another individual in the sample (pi hat > 0.2), where one individual from each pair of related samples was excluded considering data quality and phenotype, or 5) they were classed as non-European, determined by merging the BDR genotypes with data from HapMap Phase 3 (http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html), linkage disequilibrium (LD) pruning the overlapping SNPs such that no pair of SNPs within 1500 bp had $r^2 > 0.20$ and visually inspecting the first two genetic principal components along with the known ethnicities of the HapMap sample to define European samples (**See Figure 5.5**). Prior to imputation SNPs with high levels of missing data (>5%), Hardy-Weinberg equilibrium p < 0.001 or minor allele frequency <1% were excluded. The genetic data were then recoded as vcf files before uploading to the Michigan Imputation Server (Das *at al.*, 2016) (https://imputationserver.sph.umich.edu/index.html#!) which uses Eagle2 (Loh *at al.*, 2016) to phase haplotypes, and Minimac4 (https://genome.sph.umich.edu/wiki/Minimac4) with the most recent 1000 Genomes reference panel (phase 3, version 5). Imputed genotypes were then filtered with PLINK2.0alpha, excluding SNPs with an R^2 INFO score < 0.5 and recoded as binary PLINK format. Proceeding with PLINK1.9, samples with >5% missing values, and SNPs with >2 alleles, >5% missing values, Hardy-Weinberg equilibrium p < 0.001, or a minor allele frequency of <5% were excluded. Variants were named using their genomic locations (“chr:pos”) and variant type (SNP/INDEL) and therefore duplicate variants were excluded. The final quality controlled imputed set of genotypes contained 6,607,832 variants.

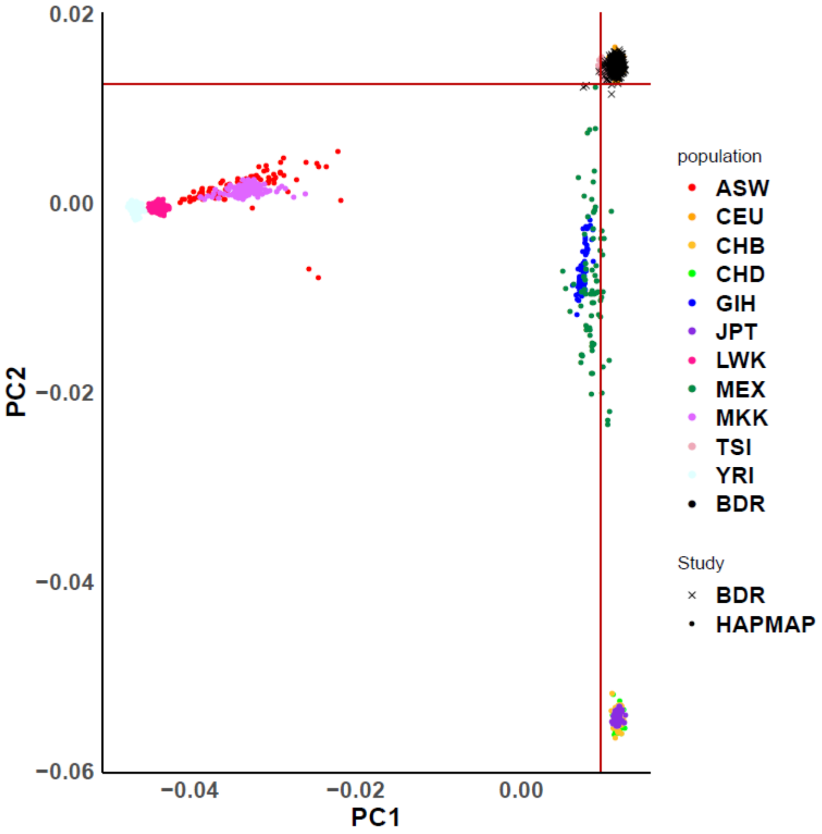


Figure 5.5: Visually ascertaining ancestry outliers in the BDR sample using principal components. *BDR was merged with HapMap3 and ancestry was ascertained based on the first two principal components. ASW = African ancestry in Southwest USA; CEU = Utah residents with Northern and Western European ancestry from the CEPH (The Centre d'Etude du Polymorphism Humain) collection; CHB = Han Chinese in Beijing, China; CHD = Chinese in Metropolitan Denver, Colorado; GIH = Gujarati Indians in Houston, Texas; JPT =Japanese in Tokyo, Japan; LWK = Luhya in Webuye, Kenya; MXL = Mexican ancestry in Los Angeles, California; MKK = Maasai in Kinyawa, Kenya; TSI = Toscani in Italia; YRI = Yoruba in Ibadan, Nigeria; BDR = Brains for Dementia Research; and HAPMAP = The International HapMap Project.*

5.3.2.3 Final cortex QTL dataset

Overlapping samples between the genetic and DNAm data which passed the stringent QC criteria for both data types were included in the final cortex mQTL dataset. Cross-hybridizing probes, probes with a common SNP (European population minor-allele frequency > 0.01) within 10 base pairs of the DNAm site or a single base extension and sex chromosomes were excluded from the QTL analysis (McCartney *at al.*, 2016). The final QC'd set of data included 522 samples (53% male, age range = 41-104 years, mean age = 83.44 years), 763,451 DNAm probes and 6,607,832 genetic variants.

5.3.3 Whole blood eQTL dataset

I used a publicly available expression quantitative trait loci (eQTL) dataset generated by Westra and colleagues (2015). They conducted an eQTL meta-analysis of 5,311 samples from whole blood, where gene expression was measured on Illumina expression arrays (annotation file for the Illumina HumanHT-12 v3.0 Gene Expression BeadChip; https://support.illumina.com/downloads/humanht-12_v3_product_files.html) and the genetic data (SNP data) was imputed using the HapMap2 reference panel (International HapMap Consortium, 2005). I only included probes where the p-value for the top *cis*-eQTL was p<5e-08. I removed SNPs (eQTLs) with minor allele frequency MAF <0.01). After filtering, there were 4,874 probes (tagging 4180 genes) and 757,479 SNPs.

5.3.4 Cortex eQTL dataset

I used a publicly available cortex eQTL dataset downloaded from the SMR website (https://cnsgenomics.com/software/smr/#DataResource) originally generated by Qi and colleagues (Qi *at al.*, 2018). These data were generated from a meta-analysis of three cortex eQTL datasets: the Common Mind Consortium (CMC) (Fromer *at al.*, 2016), Genotype-Tissue Expression (GTEx) (GTEx Consortium, 2015) and The Religious Orders Study and Memory and Aging Project (ROSMAP). Gene expression levels of CMC and GTEx were quantified by RNA sequencing, and the annotation was from GENCODE Version 19 (Harrow *at al.*, 2012). The GTEx summary data are available at dbGaP (http://www.gtexportal.org/home/). The sample size of the GTEx brain tissue goes up to 125 and consists of up to 24,762 transcripts across 10 cortex regions and ~6.5 million imputed SNPs (imputed using 1000 Genomes). CMC data were generated from PFC tissue from 467 cortex samples and the eQTL summary data consists of 14,366 transcripts and ~1.1 million imputed ~SNPs (imputed using 1000 Genomes). The ROSMAP eQTL data were generated from PFC tissue and consists of 494 individuals, 12,979 transcripts and ~6.4 million SNPs. The final meta-analysis consisted of 1194 individual, 28,538 expression sites and 7,425,225 SNPs. I limited my analyses to probes where the p-value for the top *cis*-eQTL was p<5e-08, leaving 7,370 gene expression sites (tagging 7,345 genes).

5.3.5 Generating mQTLs in Whole Blood

I performed mQTL analysis for each chromosome whereby I tested 765,013 DNAm probes against 5,359,678 SNPs using the R package *MatrixEQTL* (Shabalin, 2012). This package was built to enable fast computation of QTLs and only saves QTLs more significant than a certain threshold which I set to p<1e-05. I fitted an additive linear model, testing if the number of alleles (which were coded as 0, 1 or 2) predicted DNAm at each DNAm site. The covariates included in the analysis were age, sex, six estimated cellular composition variables (B cells, CD8 T cells, CD4 T cells, monocytes, granulocytes, and natural killer T cells), a binary batch variable for cohort, and ten principal components from the genotype data to control for population stratification and ethnicity differences. Since the majority of mQTLs with effects which are detectable in our sample are *cis* acting (as we are limited by samples size, and therefore power), I limited my analysis to *cis* mQTLs; defined as situations where the distance between QTL SNP and DNAm site is ≤ 500 kilobases (kb). In order to identify the number on independent associations for each DNAm sites in cases where they were associated with >1 mQTL I used the clump command in *PLINK1.9* with the following parameters: --clump-p1 1e-8 --clump-p2 1e-8 --clump-r2 0.1 --clump-kb 250.

5.3.6 Generating mQTLs in Cortex tissue

I performed a genome-wide mQTL analysis whereby I tested 763,451 DNAm probes and 6,607,832 SNPs using the R package *MatrixEQTL* (Shabalin, 2012) using the same parameters I set for the whole blood QTL analysis (see above section **5.3.5**). The covariates included in the analysis were age, sex, neuronal cell proportions (derived using the CETS algorithm (Guintivano *at al.*, 2013) and ten principal components from the genotype data to control for population stratification and ethnicity differences. Since the majority of mQTLs are *cis* acting, I limited my analysis to *cis* mQTLs. In order to identify the number on independent associations for each DNAm sites in cases where they were associated with >1 mQTL, I used the clump command in *PLINK1.9* using the same parameters as described above (see section **5.3.5**).

5.3.7 Enrichment analyses

Enrichment analyses were conducted to identify if mQTLs were overrepresented in certain genic or CpG island (CGI) features. DNAm sites were annotated to genic and CGI features based on the Illumina manifest. Sites are defined as being annotated to genes if they are within the gene body or < 1,500 base pairs (bp) from the transcription start site. Sites are defined as CGIs if they are located within the boundaries of a CpG island, to a shore if they are 2,000 bp from an island and a shelf if they are 2000-4000 bp from a CGI. I generated frequency tables to identify how many sites were annotated to each feature. A Fisher's exact test was used to determine if there are non-random associations between two categorical variables (i.e. if they if a certain genic feature is over-represented this suggests there is enrichment for this category).

5.3.8 Bayesian colocalisation

I applied Bayesian colocalisation to both the whole blood and the cortex datasets to characterise the relationship between genetic variants which influence DNAm and identify if there is an underlying regional correlation structure. In order to run the colocalisation analysis all SNPs need to be included, not just those which surpass the mQTL significant threshold. Therefore, I re-ran the mQTL analysis for the DNAm sites which were associated with a significant mQTL (p < 1e-10; whole blood DNAm sites n = 167,854; cortex DNAm sites n = 42,926) including all SNPs. Using the DNAm sites which were associated with at least one significant mQTL (p < 1e-10), all pairs of DNAm sites located on the same chromosome and which were within 250 kb of each other were tested for co-localisation. Co-localisation analysis was performed with the R *coloc* package.

Using the mQTL results I input the regression coefficients, their variances, and SNP minor-allele frequencies. Prior probabilities were set as their default values. This methodology allowed me to quantify the support across the results of each GWAS for five hypotheses (*Hi*) by calculating the posterior probabilities (*PPi)* as described in section **5.1.**

The strength of the association can be determined using the posterior probabilities (PP) which are outputs from *Coloc.* Based on criteria derived by Gou and colleagues

(Guo *at al.*, 2015), the posterior probability (PP) provides evidence for a co-localized association within the same genomic region if PP3 + PP4 > 0.99. There is “suggestive” evidence to support associations between both DNAm sites with the same causal mQTL variant if PP3 + PP4 > 0.99 & PP4/PP3 > 1. There is “convincing” evidence to support associations between both DNAm sites with the same causal mQTL variant of PP3 + PP4 > 0.99 & PP4/PP3 > 5.

5.3.9 SMR analysis

SMR analyses were performed using publicly available software (https://cnsgenomics.com/software/smr/#SMR) (Zhu *at al.*, 2016). SMR uses a Mendelian randomisation approach, where the most significant QTL (mQTL/eQTL) SNP (which was tested within the AD GWAS) is used as a genetic instrument. SMR analysis involves two steps. First, a two-sample MR is performed with the two-step least squares approach (2SLS), using the effect size of the top *cis*-QTL SNP and its corresponding effect in the GWAS. Second, the SMR tool then tests for heterogeneity effects by using alternative SNPs as the instrumental variable. If there is one singular causal variant the association statistics remain the same irrespective of the genetic instrument used. However, if there are two distinct causal variants there will be variation in the results. In order to distinguish pleiotropy from linkage the HEIDI test (HEIDI p > 0.01) is applied. If the HEIDI p > 0.01 this suggest there is a pleiotropic effect on a GWAS trait and DNAm. In the analyses I used the default SMR settings which excluded SNPs in LD with the top *cis*-mQTL at r^2 >0.9; SNPs in near perfect LD with the top *cis*-mQTL are not informative for the HEIDI test (Zhu *at al.*, 2016).

5.3.9.1 LOAD GWAS

I used the two most recent and publicly available LOAD GWAS datasets (Jansen *at al.*, 2019; Kunkle *at al.*, 2019) for the SMR analysis, both of which were released in 2019. Of note, there have been several LOAD GWAS preprints published within the last six months on medrxiv (Bellenguez *at al.*, 2020; de Rojas *at al.*, 2020; Wightman *at al.*, 2020) (see Chapter 1 section **1.2** for more details), however these data are not yet publicly available. One further GWAS has very recently been published (Schwartzentruber *at al.*, 2021), although the summary statistics were not available at the time I conducted the analysis for this Chapter.

First, I used the Kunkle *at al.* (2019) GWAS, which included clinically and autopsy-documented LOAD cases (35,274 cases and 59,163 controls). Kunkle and colleagues conducted a GWAS meta-analysis of non-Hispanic whites from the International Genomics of Alzheimer's Project (IGAP) and 25 LOAD risk loci were identified. Second, I used the Jansen *at al.* (2019) GWAS which was based both on clinically diagnosed AD and AD-by proxy (based on family history) cases (71,880 cases, 383,378 controls) and 29 LOAD risk loci were identified. See Chapter 1 section **1.2** for more details on these GWAS.

I used both of these LOAD GWAS in my analyses as together they provide the most comprehensive list of AD SNPs currently publicly available. Of note, the genetic correlation between clinically diagnosed AD and self-reported parental history of AD has been reported in previous studies to be between 0.66-0.91 (Marioni *at al.*, 2018), suggesting the AD by-proxy measure captures much of the same underlying genetic architecture as case-control studies while increasing power to detect LOAD risk loci. To confirm these previous findings, I calculated the correlation between the effect sizes of the two 2019 LOAD GWAS, using a clumped set of variants (clumped using *PLINK1.9* (Chang *at al.*, 2015) and the --clump flag, with the default parameters), keeping only overlapping and nominally significant probes (p<0.05) from each GWAS. There was a correlation of 0.89 between the effect sizes of the two GWAS (see **Figure** 5.6). In addition, 18 genome-wide significant (p<5e-08) SNPs overlap between the two GWAS. Although the GWAS are generally similar there are some differences between them and therefore including both in the SMR analysis provides the opportunity to potentially prioritise more AD genes.

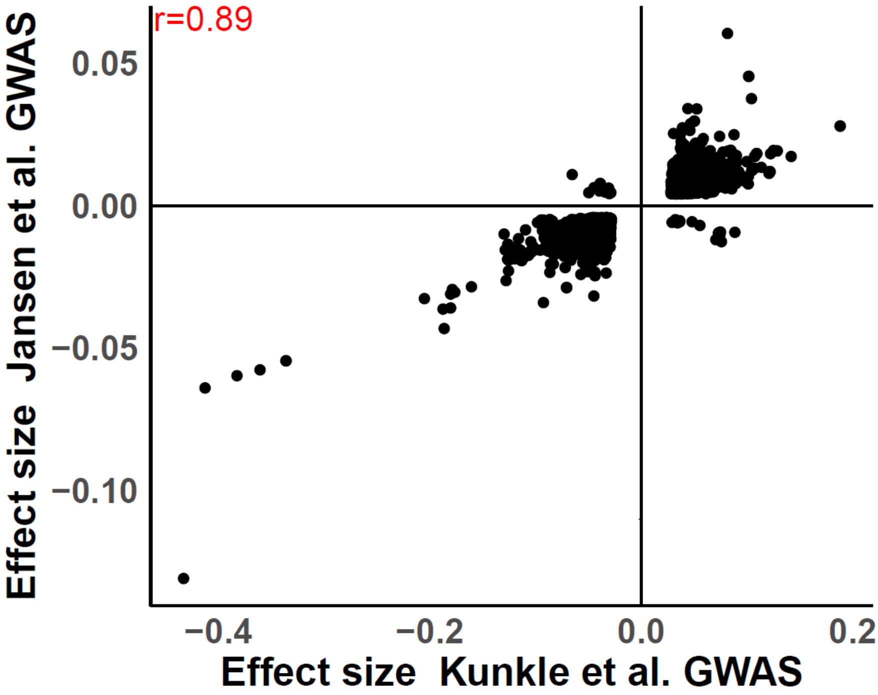


Figure 5.6: The Kunkle at al. and Jansen at al. LOAD GWAS effect sizes are highly correlated (r=0.89). *Shown is a plot of effects sizes for the Kunkle at al. LOAD GWAS (x-axis) against the Jansen at al. LOAD GWAS effect sizes (y-axis). Overlapping SNPs between the two GWAS and SNPs reaching nominal significance (p<0.05) in both were included in the analysis.*

5.3.9.2 mQTL enrichment analysis for variants associated with LOAD

I performed mQTL enrichment analyses using GARFIELD (Iotchkova *at al.*, 2016, 2019), a software tool which tests for enrichment of GWAS associated variants in genomic annotation categories. The workflow applied in GARFIELD is shown in **Figure 5.7.** Briefly, this method requires the user to provide GWAS p-values for all variants and an annotation file which indicates if these variants are located in the functional categories being tested. The first step involves 'LD pruning', which is performed using a greedy algorithm to extract independent variants based on LD clumping ($r^2 \geq 0.01$) and distance information. The second step is the 'LD tagging annotation step' – each variant (or a variant in high LD with that variant; $r^2 \geq 0.8$) with a regulatory annotation is annotated if it overlaps the features. Next, the odds ratio (OR) and enrichment p-values at different GWAS p-value thresholds are calculated using a logistic regression for each annotation. To identify the number of independent annotations eigenvalues are calculated using a correlation matrix of the annotation overlap matrix (as shown in **Figure 5.7**) and subsequently a Bonferroni correction is applied at the 95% significance threshold.

Both LOAD GWAS datasets were reformatted to be input into GARFILED. The mQTLs were used to define an annotation category. I tested for the enrichment among whole blood and cortex mQTLs for variants associated with LOAD at four p-value thresholds ($p < 5e-05$, 5 e-06, 5 e-07, 5 e-08). As I ran four analyses (whole blood mQTLs and cortex mQTLs with the Jansen and Kunkle GWAS) I used a relevant Bonferroni correction threshold ($p<0.05/4 = 0.0125$).

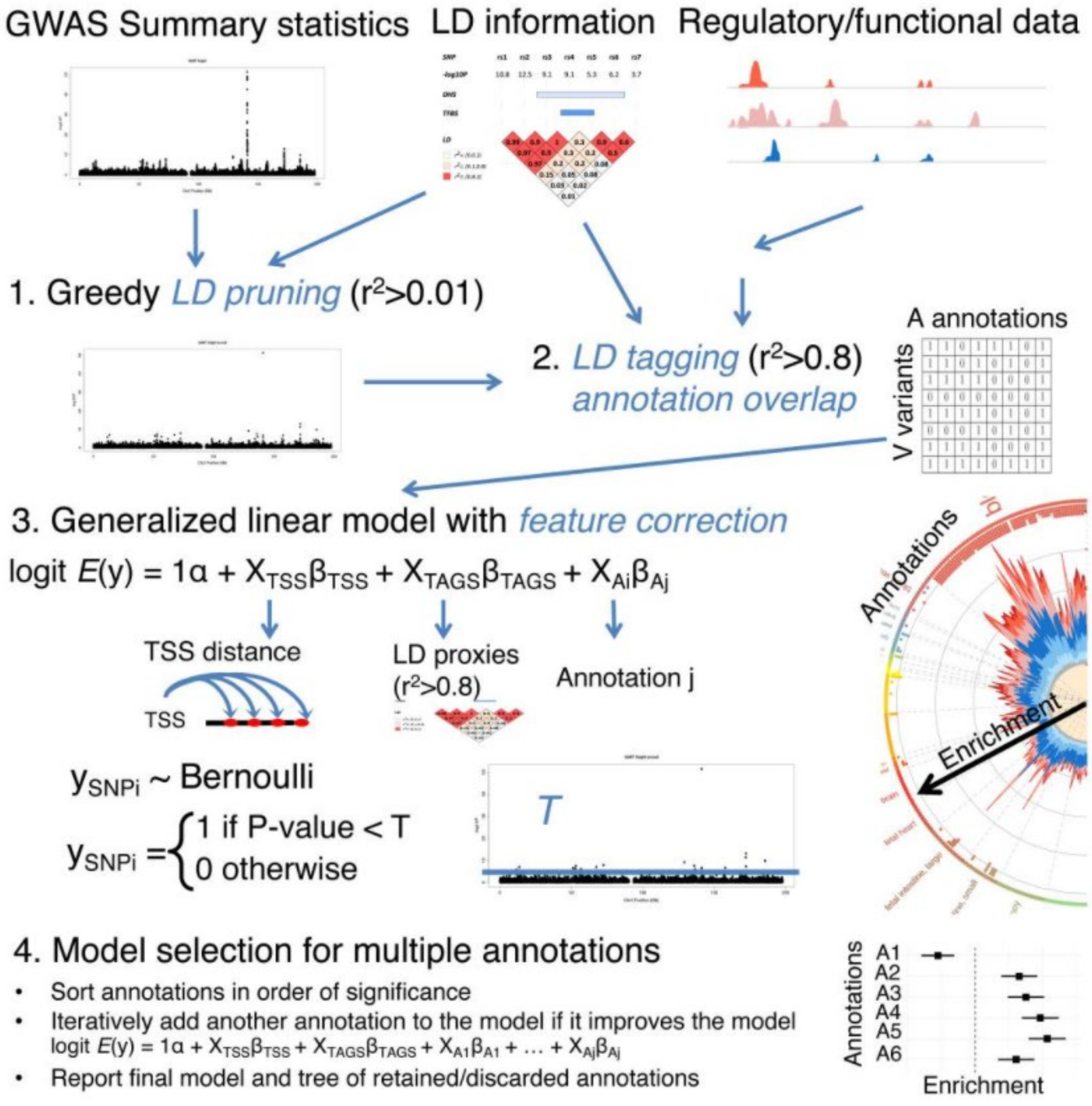

GWAS Summary statistics
LD information
Regulatory/functional data
1. Greedy LD pruning (r²>0.01)
A annotations
2. LD tagging (r²>0.8)
annotation overlap
V variants
3. Generalized linear model with feature correction
logit E(y) = 1α + X_TSS β_TSS + X_TAGS β_TAGS + X_Aj β_Aj
TSS distance
TSS
LD proxies (r²>0.8)
Annotation j
Annotations
Enrichment
y_SNPi ~ Bernoulli
y_SNPi = 1 if P-value < T, 0 otherwise
T
4. Model selection for multiple annotations
Sort annotations in order of significance
Iteratively add another annotation to the model if it improves the model
logit E(y) = 1α + X_TSS β_TSS + X_TAGS β_TAGS + X_A1 β_A1 + … + X_Aj β_Aj
Report final model and tree of retained/discarded annotations
A1
A2
A3
A4
A5
A6
Annotations
Enrichment

Figure 5.7: Outline of the GARFIELD method. *Top panel: three inputs (annotation, p-value and linkage disequilibrium (LD) data) are used for the first two analytical steps (LD pruning and variant functional annotation), which result in a binary annotation overlap matrix of V pruned variants and A annotations. Middle panel: a logistic regression approach is used for testing for enrichment at a GWAS significance p-value threshold T while controlling for confounding features such as TSS distance and number of LD proxies. Bottom panel: model selection procedure for multiple annotations. Figure and legend taken directly from (Iotchkova at al., 2016).*

5.3.9.3 SMR analyses in whole blood

5.3.9.3.1 Identifying putative pleiotropic relationships between DNAm and LOAD

I used the significance threshold of p<1e-10 to select whole blood mQTLs to use as genetic instruments for the 167,854 probes that were included in the SMR analysis. I reformatted the AD GWAS SNPs to be in the 1000 Genomes format (chr:bp) to align them with the mQTL output - these reformatted files were used for all of the SMR analyses in this chapter. SMR analyses were run using both the Kunkle *at al.* (2019) and the Jansen *at al.* (2019) GWAS summary statistics. Significant pleiotropic associations between DNAm and LOAD were selected as those with SMR p < 3.5e-07 (corrected for 167,854 DNAm sites tested in the SMR analysis) and HEIDI p > 0.01.

5.3.9.3.2 Identifying putative pleiotropic relationships between gene expression and LOAD

I used SMR with a publicly available whole blood eQTL dataset generated by Westra and colleagues (n = 5,111) which identified eQTLs for 4,874 probes (annotated to 4180 genes). SMR analyses were run using both the Kunkle *at al.* (2019) and the Jansen *at al.* (2019) GWAS summary statistics. Significant pleiotropic associations between gene expression and AD were selected as those with SMR p < 8.38e-06 (corrected for 4,874 gene expression probes tested) and HEIDI p > 0.01.

5.3.9.3.3 Identifying putative pleiotropic relationships between DNAm and gene expression

I also used SMR to identify relationships between DNAm and gene expression. The Westra data were used as the eQTL dataset. All DNAm-expression combinations were tested where (1) the DNAm site had a significant mQTL (P< 1e-10); (2) the gene had a significant eQTL (P<5e-8); and (3) there was a significant common genetic variant tested within 500 kb of the gene expression probe and DNAm site. The standard Illumina manifest used to annotate the EPIC DNAm data is not necessarily the most functionally relevant as it is based on the nearest gene. I therefore used the mQTL-eQTL SMR results to refine the gene annotations for DNAm sites. Significant pleotropic associations between DNAm and gene expression were selected as those with SMR p < 4.47e-07 (corrected for 119,352 pairs of probes tested) and HEIDI p >

0.01. I conducted enrichment analysis of the mQTL-eQTL results as described in **5.3.7**.

5.3.9.3.4 Identifying putative pleiotropic relationships between DNAm, Gene Expression and LOAD

To identify pleiotropic relationships between DNAm, expression and LOAD (i.e. to test if DNAm, gene expression and LOAD are associated because of a shared causal variant), I utilised the mQTL-eQTL, DNAm-LOAD and expression-LOAD results, identifying situations where the association signals were consistent across the three analyses at a locus. First, I tested these relationships using the whole blood SMR results which were generated using the Kunkle *at al.* GWAS. I then repeated the analysis using the SMR results generated using the Jansen *at al.* GWAS. I included SMR significant mQTL results but relaxed the threshold of the SMR expression results (eQTL p SMR < 1e-4) since an mQTL-eQTL relationship would strengthen the hypothesis of a gene being involved in the pathogenesis of LOAD.

5.3.9.3.5 Pathway analysis

An important downstream analysis after genome-wide DNAm studies (e.g. EWAS and SMR) is gene set analysis, whereby significant DMPs and genes can be related to known biological functions. I ran a gene ontology (GO) pathway analysis using the *methylglm* function within the *methylGSA* package (2019). *methylglm* adjusts for the number of DNAm sites in the logistic regression model (see Chapter 2 section **2.3.4** for more details). I used the default settings of the package. Pathway analysis was run for both mQTL SMR analyses (Kunkle blood DNAm and Jansen blood DNAm). I included DNAm sites which had an SMR p-value < 5e-5 and passed the HEIDI test.

5.3.9.4 SMR analyses in the Cortex

5.3.9.4.1 Identifying putative pleiotropic relationships between DNAm and LOAD

I applied SMR to the cortex mQTLs. I used the significance threshold of p<1e-10 for the mQTLs in the cortex sample to identify genetic instruments for the 42,926 probes that were included in the SMR analysis. SMR analyses were run using both the Kunkle *at al.* (2019) and the Jansen *at al.* (2019) GWAS summary statistics. Significant pleotropic associations between DNAm and LOAD were selected as those with SMR p < 1.15e-06 (corrected for 42,926 DNAm probes tested) and HEIDI p > 0.01.

5.3.9.4.2 Identifying putative pleiotropic relationships between gene expression and LOAD

I applied SMR to a publicly available cortex eQTL dataset (n=1,194) where eQTLs are available for 7,370 gene expression sites (annotated to 7,345 genes). Data were downloaded from the SMR site (https://cnsgenomics.com/software/smr/#DataResource) and represent results from a meta-analysis of three cortex eQTL datasets: GTEx cortex (GTEx Consortium 2017), CMC (Fromer *at al.* 2016), and ROSMAP (Ng *at al.* 2017). SMR analyses were run using both the Kunkle *at al.* (2019) and the Jansen *at al.* (2019) GWAS summary statistics. Significant pleotropic associations between gene expression and LOAD were selected as those with SMR p < 7.4e-06 (corrected for 6,740 gene expression probes tested) and HEIDI p > 0.01.

5.3.9.4.3 Identifying putative pleiotropic relationships between DNAm and Gene Expression

I used SMR to identify relationships between DNAm and gene expression in cortex using the methodology described in **5.3.9.3.3**. Significant pleotropic associations between DNAm and gene expression were selected as those with SMR p < 2.14e-6 (corrected for 23,333 pairs of probes tested) and HEIDI p > 0.01. I conducted enrichment analysis of the mQTL-eQTL results as described in **5.3.7**

5.3.9.4.4 Identifying putative pleiotropic relationships between DNAm, Gene Expression and LOAD

To identify pleiotropic relationships between DNAm, expression and LOAD I utilised the mQTL-eQTL, DNAm-LOAD and expression-LOAD cortex results, identifying situations where the association signals were consistent across the three analyses using the methodology described in **5.3.9.3.4**.

5.3.9.4.5 Pathway analysis

Pathway analysis was applied to the results from each of the cortex mQTL SMR analyses (Kunkle cortex DNAm, and Jansen cortex DNAm) as described above in section **5.3.9.3.5**.

5.3.10 Comparing Blood and Cortex SMR results

I compared the SMR results from blood and cortex. First, I had to take into consideration whether or not the same genes and DNAm sites were tested, and therefore limited the results to genes of DNAm sites which were tested across the datasets. Using the mQTL SMR results I evaluated the consistency of effect sizes across the analysis including SMR associations which had nominal significance ($p<0.05$). A binomial sign test was applied to gauge the significance of this effect. This analysis was repeated for the eQTL SMR results.

5.4 Results

5.4.1 Overview of results

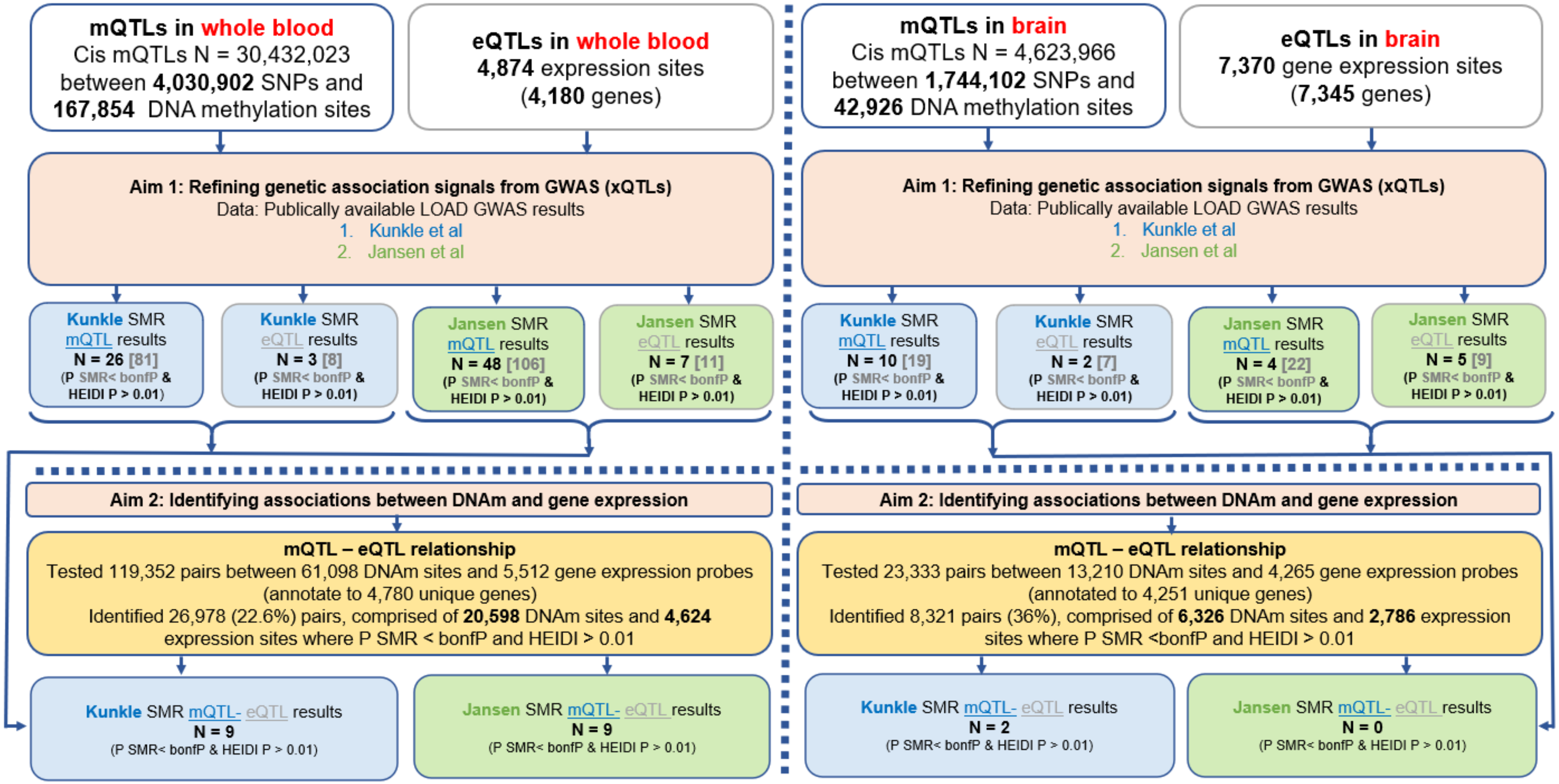

mQTLs in whole blood
Cis mQTLs N = 30,432,023 between 4,030,902 SNPs and 167,854 DNA methylation sites
eQTLs in whole blood
4,874 expression sites (4,180 genes)
mQTLs in brain
Cis mQTLs N = 4,623,966 between 1,744,102 SNPs and 42,926 DNA methylation sites
eQTLs in brain
7,370 gene expression sites (7,345 genes)
Aim 1: Refining genetic association signals from GWAS (xQTLs)
Data: Publically available LOAD GWAS results
1. Kunkle et al
2. Jansen et al
Aim 1: Refining genetic association signals from GWAS (xQTLs)
Data: Publically available LOAD GWAS results
1. Kunkle et al
2. Jansen et al
Kunkle SMR mQTL results N = 26 [81] (P SMR< bonfP & HEIDI P > 0.01)
Kunkle SMR eQTL results N = 3 [8] (P SMR< bonfP & HEIDI P > 0.01)
Jansen SMR mQTL results N = 48 [106] (P SMR< bonfP & HEIDI P > 0.01)
Jansen SMR eQTL results N = 7 [11] (P SMR< bonfP & HEIDI P > 0.01)
Kunkle SMR mQTL results N = 10 [19] (P SMR< bonfP & HEIDI P > 0.01)
Kunkle SMR eQTL results N = 2 [7] (P SMR< bonfP & HEIDI P > 0.01)
Jansen SMR mQTL results N = 4 [22] (P SMR< bonfP & HEIDI P > 0.01)
Jansen SMR eQTL results N = 5 [9] (P SMR< bonfP & HEIDI P > 0.01)
Aim 2: Identifying associations between DNAm and gene expression
mQTL – eQTL relationship
Tested 119,352 pairs between 61,098 DNAm sites and 5,512 gene expression probes (annotate to 4,780 unique genes)
Identified 26,978 (22.6%) pairs, comprised of 20,598 DNAm sites and 4,624 expression sites where P SMR < bonfP and HEIDI > 0.01
Aim 2: Identifying associations between DNAm and gene expression
mQTL – eQTL relationship
Tested 23,333 pairs between 13,210 DNAm sites and 4,265 gene expression probes (annotated to 4,251 unique genes)
Identified 8,321 pairs (36%), comprised of 6,326 DNAm sites and 2,786 expression sites where P SMR <bonfP and HEIDI > 0.01
Kunkle SMR mQTL- eQTL results N = 9 (P SMR< bonfP & HEIDI P > 0.01)
Jansen SMR mQTL- eQTL results N = 9 (P SMR< bonfP & HEIDI P > 0.01)
Kunkle SMR mQTL- eQTL results N = 2 (P SMR< bonfP & HEIDI P > 0.01)
Jansen SMR mQTL- eQTL results N = 0 (P SMR< bonfP & HEIDI P > 0.01)

Figure 5.8: Overview of results for chapter 5: Methylation quantitative trait loci (mQTL) analysis and summary data-based Mendelian randomisation (SMR). *The results N are the number of sites which passed both Bonferroni significant and the HEIDI test, the number in the grey square brackets represent the Bonferroni significant hits without the HEIDI correction. eQTL = expression QTL; LOAD = late onset Alzheimer's disease; bonfP = Bonferroni corrected P value and DNAm = DNA methylation.*

5.4.2 DNA methylation quantitative trait loci in whole blood

I generated a database of DNAm mQTLs in whole blood tissue using the *MatrixEQTL* package (Shabalin, 2012), testing 5,359,678 SNPs against 765,013 DNAm sites which were on the Illumina EPIC array and passed stringent QC criteria (see **Methods 5.2**). At a highly conservative Bonferroni threshold (p < 5e-08/765,013 = 6.5e-14) I identified 23,859,815 significant *cis* (distance between the QTL SNP and DNAm site ≤ 500Kb) mQTL associations between 3,731,711 SNPs and 142,964 DNAm sites. This is 88% more mQTLs than the UKHLS study alone (12,689,548 mQTLs were identified in the study by Hannon and colleagues (2018)) which reflects the increased power. This is currently the most comprehensive mQTL dataset in EPIC blood samples. There was a mean percentage change in DNAm per additional reference allele of 2.5% (standard deviation [SD] =2.6%) across the associated mQTL sites. DNAm sites associated with genetic variants were associated with a median of 81 mQTLs (interquartile range [IQR] = 27-167), likely reflecting linkage disequilibrium – the correlation structure between proximal variants. Each mQTL variant was associated with a median of 3 DNAm sites (IQR=1-7). Examples of mQTLs identified in whole blood are shown in **Figure 5.9A** and **Figure 5.10A.**

The set of mQTLs used throughout this chapter are based on associations meeting a more relaxed "discovery" threshold of p<1e-10 as defined in the previous SMR paper conducted by our group (Hannon *at al.*, 2018). At this threshold I identified 30,432,023 significant *cis* mQTL associations between 4,030,902 SNPs and 167,854 DNAm sites. The inclusion of EXTEND with UKHLS led to the identification of 80% more QTLs than using UKHLS alone (17,051,673 mQTLs were identified in the study by Hannon and colleagues (2018) at the "discovery" threshold). There was a mean percentage change in DNAm per additional reference allele of 2.2% (standard deviation [SD] =2.4%) across the associated mQTL sites. DNAm sites associated with genetic variation were associated with a median of 89 mQTLs (interquartile range [IQR] = 29-224). Each mQTL variant was associated with a median of 4 DNAm sites (IQR=2-9). To identify independent associations with each DNAm site I conducted LD clumping of the mQTLs; there were a total of 262,461 independent associations (0.86% of the total number of un-clumped mQTL associations) with a median of 1 (IQR=1–2) mQTL associated with each DNAm site. The whole blood mQTLs were enriched in intergenic regions (OR=2.61; p=2.23e-308), gene bodies (odds ratio [OR] =1.85; p=2.23e-308), transcription start sites (OR=1.51; p= 2.71e-147), and in CpG shores (OR=1.91, p=2.23e-308), shelves (OR=1.5; p= 1.63e-71) and seas (OR=2.91; p=2.23e-308). In contrast, they were significantly less likely to be located within CGIs (OR=0.94; p= 1.80e-05).

5.4.3 DNA methylation quantitative trait loci in cortex tissue

I ran the same analysis pipeline to generate a database of mQTLs in cortex tissue using the *MatrixEQTL* package (Shabalin, 2012) as described above (section **5.4.2**). In the prefrontal cortex (PFC) I tested 6,607,832 SNPs against 763,451 DNAm sites passing our stringent QC criteria. At a conservative Bonferroni threshold (p< 5e-08/763,451 = 6.5e-14). I identified 3,153,315 significant *cis* mQTL associations between 1,347,502 SNPs and 32,420 DNAm sites, which is about 13% of the number of mQTLs identified in whole blood at this threshold due to the smaller sample size. There was a mean percentage change in DNAm per additional reference allele of 4.8% (standard deviation [SD] =3.3%) across the associated mQTL sites. This is higher than the whole blood mQTL mean percentage change, reflecting the lower power and therefore only stronger mQTL effects can be identified. DNAm sites associated with genetic variation were associated with a median of 45 mQTLs (IQR = 15-112) and each mQTL variant was associated with a median of one DNAm sites (IQR=1-2. Examples of mQTLs identified in the cortex are shown in **Figure 5.9B** and **Figure 5.10B.**

The set of cortex mQTLs used throughout this chapter are based on associations meeting a more relaxed “discovery” threshold of p<1e-10 as defined in the previous SMR paper conducted by our group (Hannon *at al.*, 2018). I identified 4,623,966 significant *cis* mQTL associations between 1,744,102 SNPs and 42,926 DNAm sites using this threshold, which is considerably lower (15%) of the number of mQTLs identified in the whole blood mQTL analysis at the same threshold. There was a mean percentage change in DNAm per additional reference allele of 4.25% (SD=3.0%) across the associated mQTL sites. DNAm sites associated with genetic variation were associated with a median of 49 mQTLs (IQR = 15-123. Each mQTL variant was associated with a median of one DNAm sites (IQR=1-3). To identify independent associations with each DNAm site I conducted LD clumping of the mQTLs; there were a total of 82,296 independent associations (1.25% of the total number of un-clumped mQTL associations) with a median of a single mQTL (IQR=1-1) associated with each DNAm site, which is about 30% of the number of independent whole blood mQTLs. The cortex mQTLs showed a similar enrichment pattern to the whole blood mQTLs being enriched in intergenic regions (OR=2.01; p=2.23e-308), gene bodies (odds ratio [OR] =1.61; p=2.23e-308), transcription start sites (OR=1.24; p=3.04e-53), and in CpG shores (OR=1.69, p=2.23e-308), shelves (OR=1.12; p=2.49e-09) and seas (OR=2.45; p=2.23e-308), with an underrepresentation in CGIs (OR=0.84; p= 4.18E-34).

5.4.4 Comparison between whole blood and cortex mQTLs

There was an overlap of 2,686,279 mQTLs between the two mQTL datasets at the “discovery” threshold (p<1e-10). There was a strong correlation of individual mQTL effect sizes across tissues (r=0.79, binomial sign test p= 2.23e−308; see **Figure 5.9**), although the direction differed for some variants, providing evidence that there are some tissue-specific differences (see **Figure 5.10**). Since there was more power to detect whole blood mQTLs we can only identify if there are cortex-specific mQTLs; we cannot rule out that the mQTLs identified in the whole blood analysis would not be identified in the cortex if we had equal power. Of note, some of the identified cortex mQTLs were not tested for in the whole blood mQTL, although 3,971,969 (86%) were. Of these cortex mQTLs, 1,285,690 (32%) were not identified as a significant mQTL in whole blood. This provides further evidence that there are tissue specific mQTL effects.

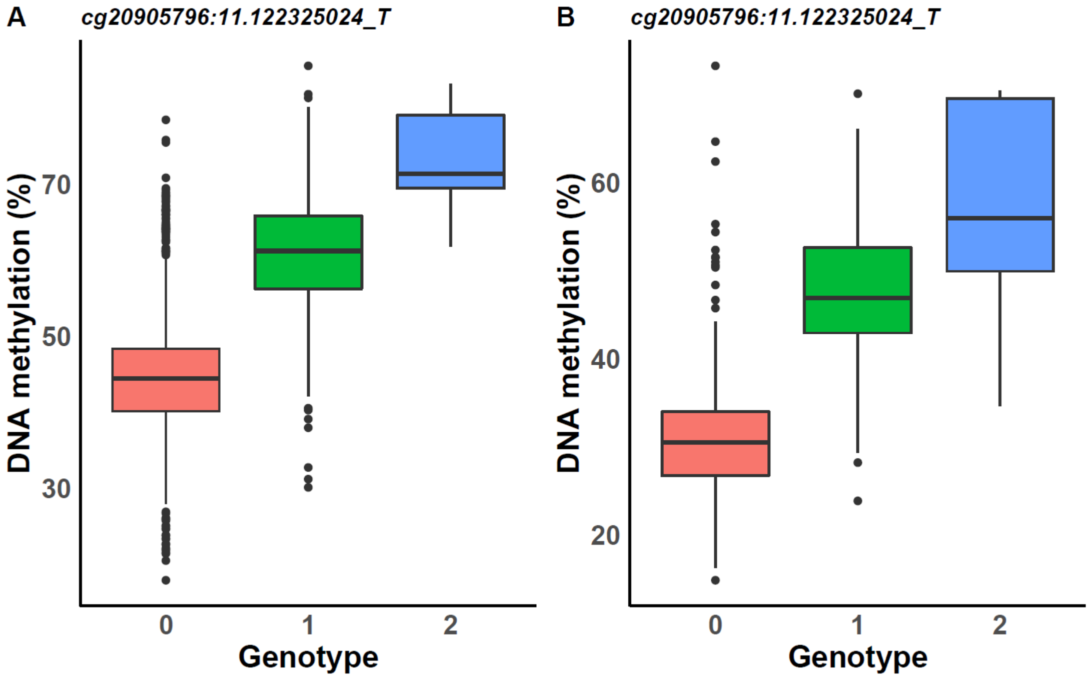


***Figure 5.9: Shown is an example of an mQTL between 11:122325025 and cg20905796 in (A) the whole blood and (B) the cortex**. The direction of effect is consistent in whole blood and the cortex (hypermethylated with increasing number of effect alleles).The x-axis represents the genotype group (number of effect alleles - in this case T) and the y-axis represent the % DNA methylation for that cg site for each individual in the different genotype groups.*

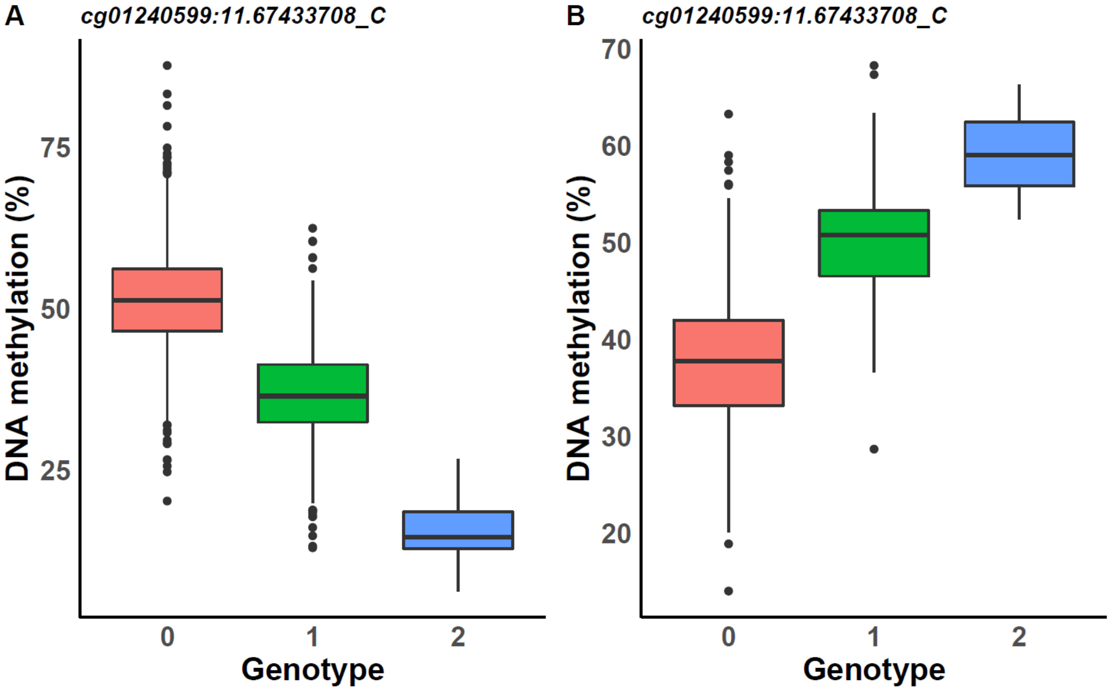


***Figure 5.10: Shown is an example of an mQTL between 11.67433708 and cg01240599 in (A) the whole blood and (B) the cortex**. The direction of effect differs in whole blood (hypomethylated with increasing number of effect alleles) and the cortex (hypermethylated with increasing number of effect alleles).The x-axis represents the genotype group (number of effect alleles - in this case C) and the y-axis represent the % DNA methylation for that cg sites for each indiviual in the different genotype groups.*

5.4.5 mQTL associations influence DNAm at sites which are heritable

Using the results from a twin study (van Dongen *at al.*, 2016) where DNAm was derived in whole blood using the Illumina 450K array, I investigated if the heritability (h_2) of DNAm sites which were associated with at least one mQTL variant were more strongly influenced by additive genetic factors than other tested DNAm sites. In the whole blood mQTL dataset, mQTL sites had a higher heritability (median h_2 =42%; IQR= 26%-59%) than all DNAm sites (median h_2=12%; IQR= 4.2%-30%; Mann-Whitney P<2.23E−308). Correspondingly, in the PFC mQTL dataset mQTL sites also had a higher heritability (median h_2=41%; IQR= 17%-65%) compared to all DNAm sites (median h_2=12%; IQR= 4.2%-30%; Mann-Whitney p < 2.23E−308).

5.4.6 mQTLs are enriched in LOAD associated variants

I ran enrichment analysis using GARFIELD to assess if whole blood and cortex mQTLs are enriched for LOAD-associated variants at four p-value thresholds (pTs; pT < 5e-05, 5 e-06, 5 e-07, 5 e-08). There was an enrichment of LOAD-associated variants amongst whole blood and cortex mQTLs for at least two GWAS pTs for all analyses (see **Figure 5.11** and **Table 5.1**). Whole blood mQTLs were significantly enriched (p < 0.0125) for LOAD variants across all Kunkle and Jansen GWAS pTs, with the most significant enrichment occurring at the less stringent GWAS threshold - pT<5e-05 (Kunkle OR=2.27, p=6.19e-08; Jansen OR=2.74, p=2.79e-18) (see **Figure 5.11** and **Table 5.1**). There was an enrichment of Jansen LOAD-associated variants amongst the cortex mQTLs across all GWAS thresholds, with the most significant being pT<5e-05 (OR=2.76; p=1.37e-15). There was an enrichment of Kunkle LOAD-associated variants amongst the cortex mQTLs at two GWAS thresholds (pT<5e-05 and 5e-06), with the most significant being pT<5e-05 (OR=2.34; p=3.16e-07). Here I found that genetic variants exhibiting genome-wide significant association with LOAD showed a threefold enrichment amongst whole blood and cortex mQTLs, directly implicating altered gene regulation in the aetiology of LOAD, and providing evidence that these mQTL datasets are suitable for SMR analysis of LOAD.

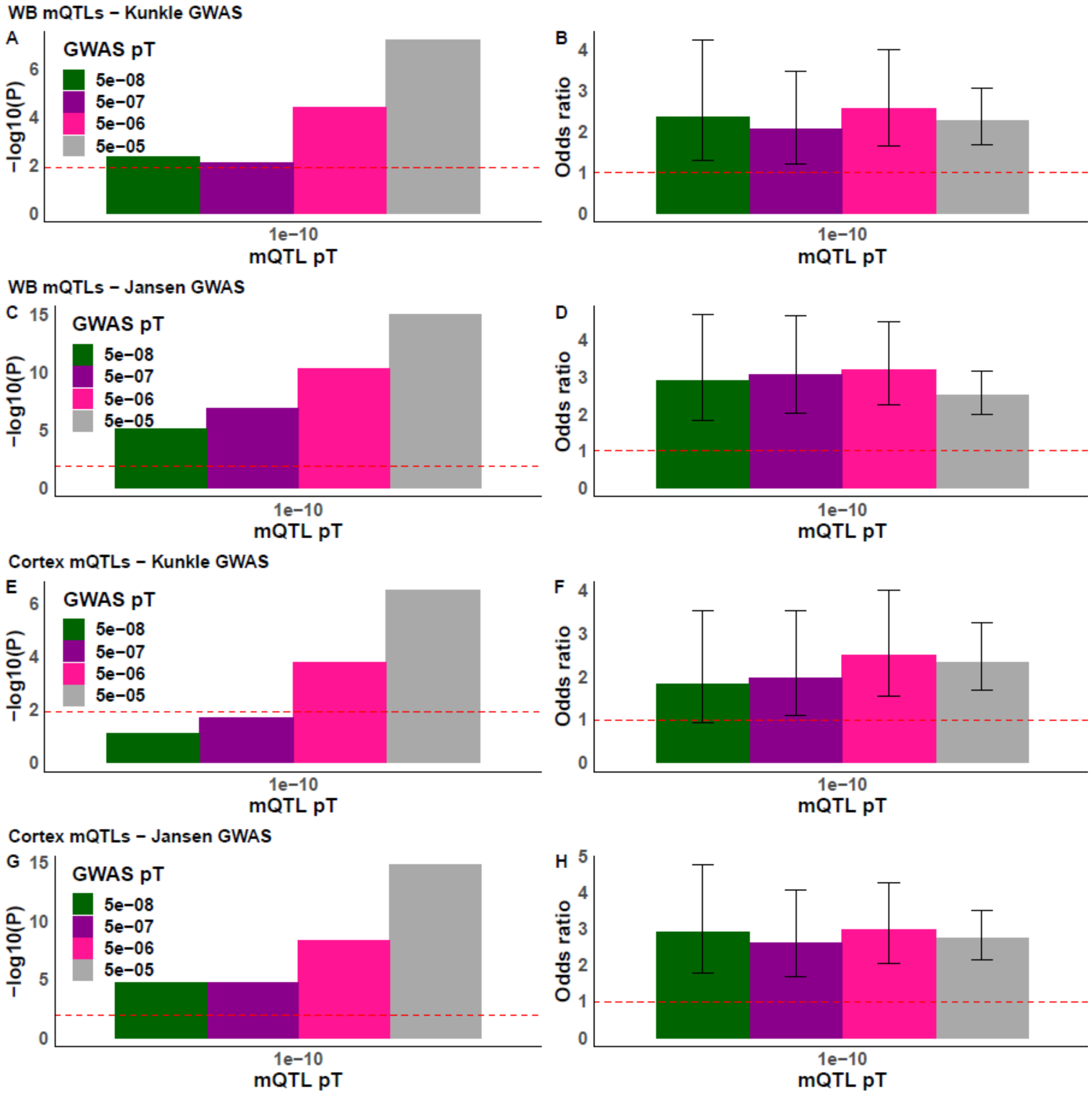


Figure 5.11: Enrichment analysis of whole blood and cortex mQTLs with LOAD GWAS variants. *The left panel (**A, C, E** and **G**) is the significance threshold of the enrichment analysis and the right panel (**B, D, F** and **H**) is the enrichment fold change (odds ratios) with 95% confidence intervals. The x-axis represents the p-value threshold for the mQTLs to be included in the analysis. A p-value < 0.0125 and an odds ratio > 1 = evidence for enrichment which is represented by the horizontal red lines. The colours represent the different p-value threshold of the LOAD GWAS variants to be tested for enrichment. A = whole blood (WB) mQTL significance of the Kunkle GWAS enrichment; B =WB mQTL odds ratio of the Kunkle GWAS enrichment; C = WB mQTL significance of the Jansen GWAS enrichment; D =WB mQTL odds ratio of the Jansen GWAS enrichment; E = cortex mQTL significance of the Kunkle GWAS enrichment; F =cortex mQTL odds ratio of the Kunkle GWAS enrichment; G = cortex mQTL significance of the Jansen GWAS enrichment; and H = cortex mQTL odds ratio of the Jansen GWAS enrichment. pT = p-value threshold.*

Table 5.1: Table of results for the enrichment analysis of whole blood and cortex mQTLs with LOAD GWAS variants. *mQTL/ GWAS = mQTL dataset used in the enrichment analysis / GWAS used in the enrichment analysis. pT = p-value threshold. OR = odds ratio.*

mQTL/ GWAS	pT	OR	P	Beta	SE
WB/ Kunkle	5.00e-05	2.27	6.19E-08	0.82	0.15
WB/ Kunkle	5.00e-06	2.56	3.70E-05	0.94	0.23
WB/ Kunkle	5.00e-07	2.05	7.70E-03	0.72	0.27
WB/ Kunkle	5.00e-08	2.34	4.56E-03	0.85	0.30
WB/ Jansen	5.00e-05	2.52	1.08E-15	0.92	0.12
WB/ Jansen	5.00e-06	3.1	5.29E-11	1.16	0.18
WB/ Jansen	5.00e-07	3.08	1.14E-07	1.12	0.21
WB/ Jansen	5.00e-08	3.11	2.94E-06	1.14	0.24
Cortex/ Kunkle	5.00e-05	2.34	3.16E-07	0.85	0.17
Cortex/ Kunkle	5.00e-06	2.49	1.73E-04	0.91	0.24
Cortex/ Kunkle	5.00e-07	1.98	0.02	0.68	0.30
Cortex/ Kunkle	5.00e-08	1.82	0.08	0.60	0.34
Cortex/ Jansen	5.00e-05	2.76	1.37E-15	1.01	0.13
Cortex/ Jansen	5.00e-06	2.98	4.27E-09	1.09	0.19
Cortex/ Jansen	5.00e-07	2.63	1.86E-05	0.97	0.23
Cortex/ Jansen	5.00e-08	2.93	1.87E-05	1.07	0.25

5.4.7 Bayesian colocalisation

There is a strong correlation structure amongst proximally located SNPs, a concept known as LD (see Chapter 1 section **1.2** for more details). Similarly, evidence suggests there is a correlation structure between DNAm levels at neighbouring DNAm sites (Hannon *at al.*, 2018). I aimed to characterise the relationship between genetic variants which influence DNAm, to identify if there is an underlying regional correlation structure. To statistically test if proximally located DNAm sites are influenced by the same causal variant, I used a Bayesian co-localisation approach – implemented with the *Coloc* package within R – identifying if any pairs of DNAm sites which were located within 250 kb of each other were associated with a shared mQTL variant.

In the whole blood mQTL dataset I tested 3,535,812 pairs of DNAm sites (median distance between sites = 110,493bp; IQR = 47,914 - 178,085bp). The posterior probability for 3,520,781 (99.5%) of these sites provided evidence for a co-localized association within the same genomic region (PP3 + PP4 > 0.99). This is as expected since only *cis* mQTLs were considered. 281,898 of these sites (7.7%) had 'suggestive' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 1). The median distance between pairs of sites with 'suggestive' evidence was 15,119bp (IQR = 1,456 – 54,396bp). 234,460 of these sites (6.6%) had 'convincing' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 5). The median distance between these sites was 12,394bp (IQR = 1,004- 49,110 bp). An example heatmap showing the Bayesian colocalisation relationship on chromosome 11 around the *BRSK2* locus is shown in **Figure 5.12** where multiple DNAm sites along this gene have a common underlying genetic signal. *BRSK2* been shown to be involved in the phosphorylation of tau in AD (Morshed *at al.*, 2020). Of note, these DNAm sites are not contiguous; several of the genetically mediated DNAm sites located within this gene do not share the same mQTL signal.

In the cortex PFC mQTL dataset I tested 524,966 pairs of DNAm sites (median distance between sites = 106,096bp; IQR = 41,324- 176,055bp). The posterior probability for 523,983 (99.8%) of these sites provided evidence for a co-localized association within the same genomic region. Again, this is as we would expect since only *cis* mQTLs were considered. 50,151 of these sites (9.6%) had suggestive evidence to support associations between both DNAm sites with the same causal mQTL variant. The median distance between pairs of sites with 'suggestive' evidence was 2,600bp (IQR = 235 – 25,834), which was closer in comparison to the whole blood, although these may be driven by the power differences between the two tissues. 42,661 of these sites (8.1%) had 'convincing' evidence to support associations between both DNAm sites with the same causal mQTL variant. The median distance between pairs of sites with 'convincing' evidence was 1,351bp (IQR = 197 – 9,714bp), which was about 11 time closer than observed in whole blood data, probably resulting from the reduced power in my cortex data.

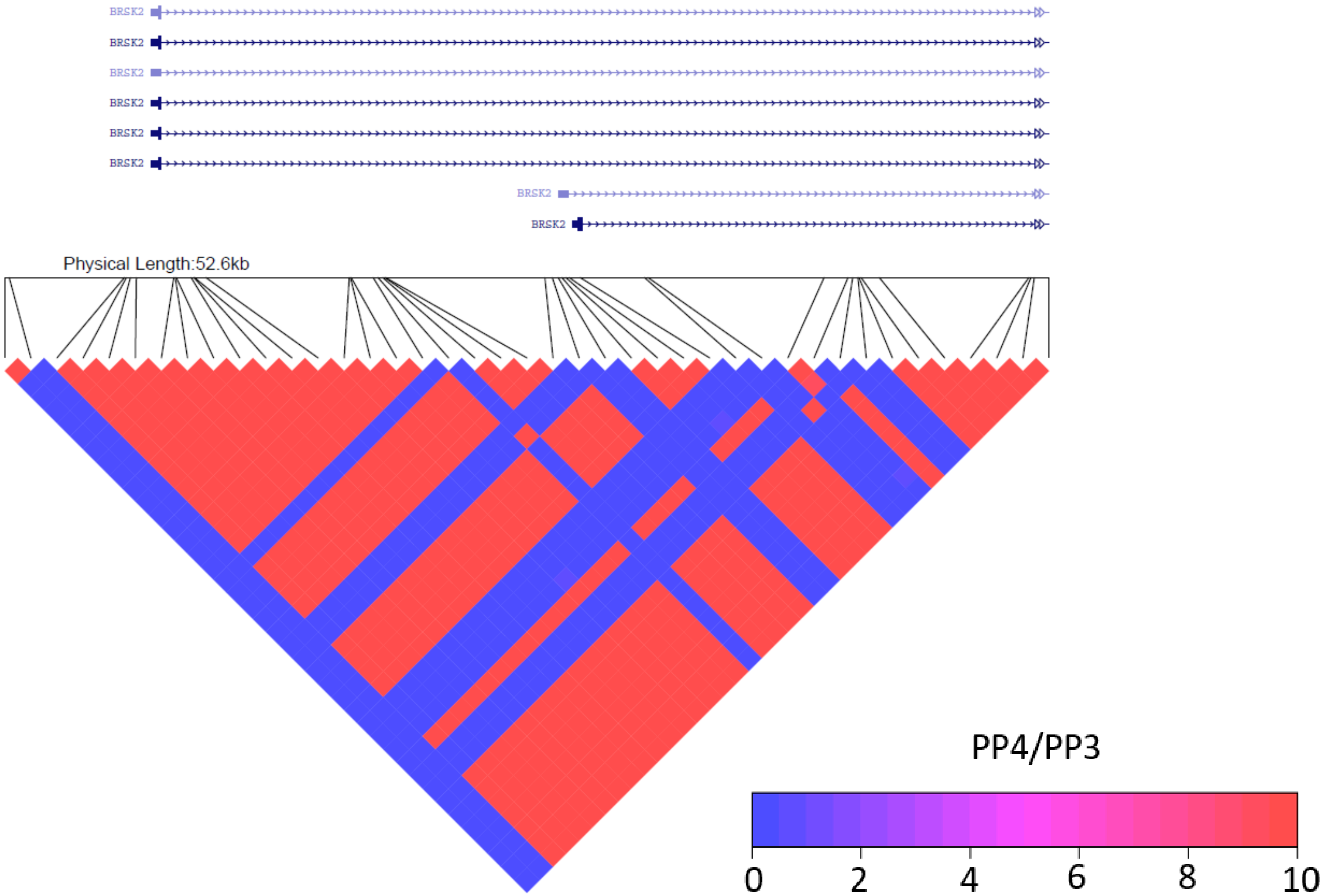


Figure 5.12: Heatmap of the whole blood Bayesian co-localisation results for all pairs of DNA-methylation sites with at least one significant mQTL (p < 1e-10) in a genomic region on chromosome 11 (chr11:1403751– 1456379). *Each square represents the relationship between two DNA-methylation sites (ordered by genomic location). The colour of each square indicates the strength of the evidence for a shared genetic signal (from blue [weak] to red [strong]). This strength is calculated as the ratio of the posterior probabilities that they share the same causal variant (PP4) compared to two distinct causal variants (PP3). The ratio was capped to a maximum value of 10.*

5.4.8 Identifying putative pleiotropic relationships between DNA methylation and Alzheimer's disease in whole blood using SMR

As described earlier (section **5.1**) there has been a focus on utilising QTLs to help refine genetic signals with the goal of prioritising causal genes for diseases. I used SMR to test the 167,854 DNAm sites identified in my whole blood mQTL analysis (see above; **Section 5.4.2**) using the two latest AD GWAS with available summary statistics (Jansen *at al.*, 2019; Kunkle *at al.*, 2019). SMR analysis involves two steps as previously described (see section **5.1**). The methodology can be interpreted as an analysis to test if the effect size of a SNP on AD is mediated by DNAm (Zhu *at al.*, 2016).

5.4.8.1 SMR results using the Kunkle *at al.* GWAS

Using the Kunkle *at al.* GWAS, which used clinically defined cases, I applied SMR and identified 81 associations which passed a Bonferroni significant threshold (p < 3.5e-07) and 26 which passed both the significance threshold and the HEIDI threshold (p > 0.01; see **Table 5.2; Figure 5.13**). Of these 26, 21 (80.8%), were positively associated with LOAD (i.e. hypermethylation is associated with increased susceptibility for developing LOAD at these sites and the *cis*-mQTL and GWAS SNP tested in the SMR analysis had the same direction of effect; see **Figure 5.14A**) and the remaining 5 DNAm sites were negatively associated with LOAD (i.e. hypermethylation is associated with decreased susceptibility for developing LOAD at these sites and the *cis*-mQTL and GWAS SNP tested in the SMR analysis had opposite directions of effect; see **Figure 5.14B**). All of the associations were identified in regions proximal to GWAS peaks, including several along chromosome 11 where 4 distinct GWAS loci have previously been identified annotated to the genes *SPI1, MS4A2, PICALM* and *SORL1* (see **Figure 5.15**).

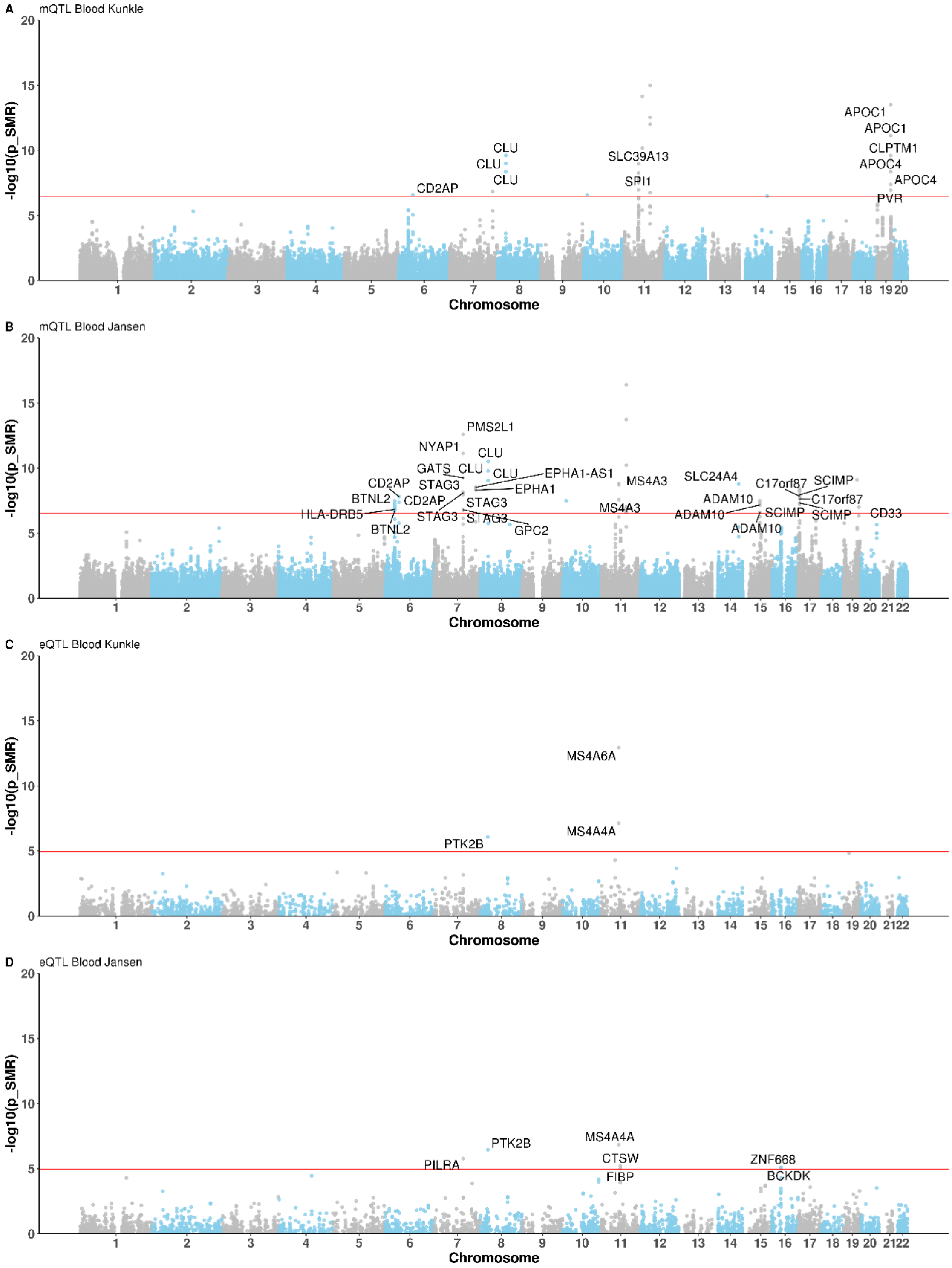


Figure 5.13: Manhattan plots of Summary data-based Mendelian Randomisation (SMR) tests for pleiotropic effects between LOAD and regulatory markers in whole blood. *__(A)__ using DNA methylation quantitative trait loci (mQTL) and the Kunkle et al GWAS summary statistics; __(B)__ using mQTLs and the Jansen et al GWAS summary statistics; __(C)__ using expression quantitated trait loci (eQTLs) and the Kunkle et al GWAS summary statistics; and __(D)__ using eQTLs and the Jansen et al GWAS summary statistics. The x axis is the genomic position, segregated by chromosome. Shown on the y-axis of each plot is the –log10 p-value from the SMR analysis using mQTLs __(A)__ and __(B)__ or eQTLs __(C)__ and __(D)__. Each point represents an SMR test for a particular mQTL/ eQTL site which passed the Bonferroni and HEIDI significant thresholds. The red horizontal line represents the genome-wide multiple testing significance threshold.*

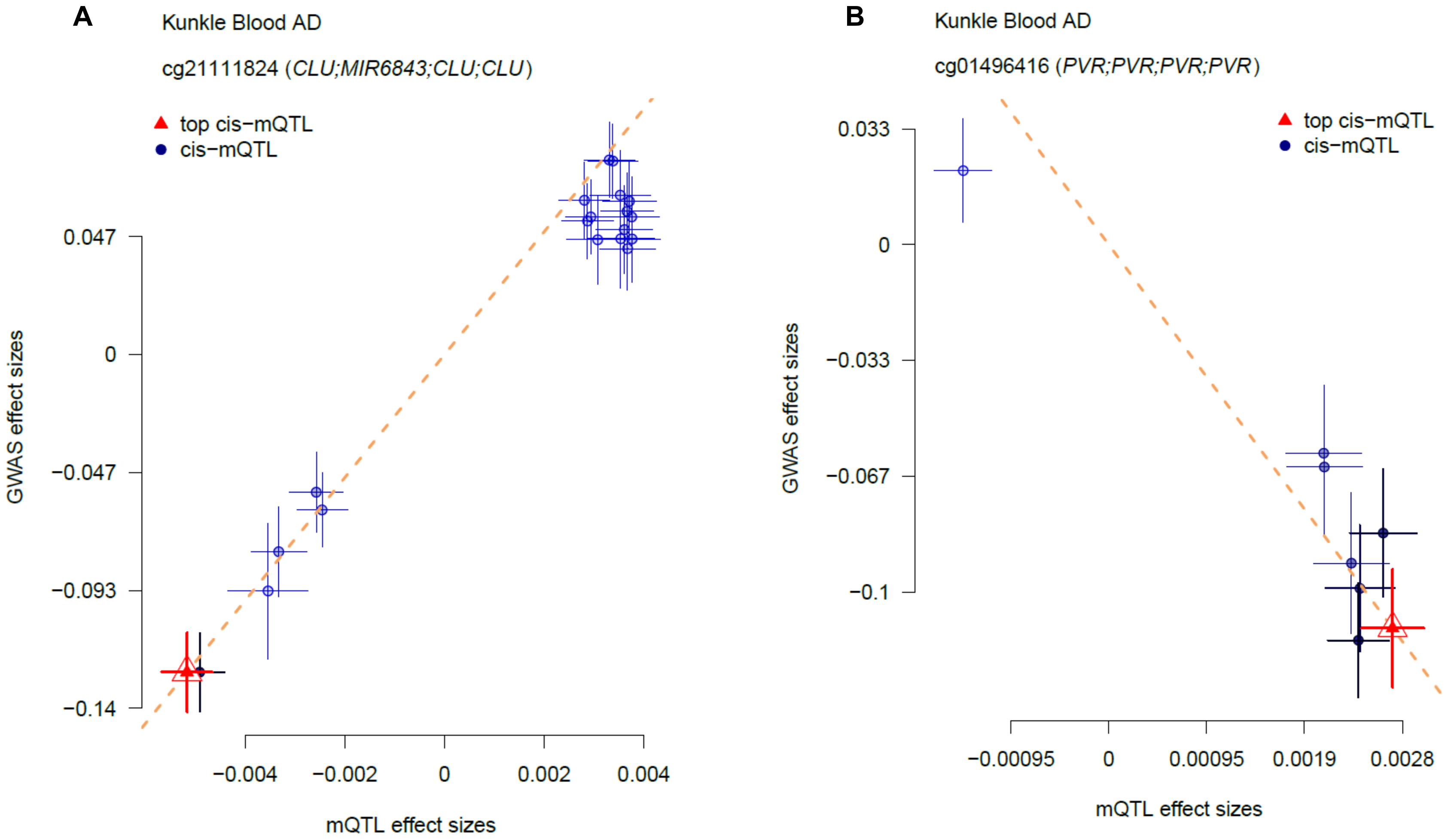


Figure 5.14: Example plots of the effect sizes of SNPs (included in the HEIDI test) from the Kunkle et al. GWAS plotted against those for SNPs from the whole blood mQTL dataset for two significant SMR hits. ***(A)*** *The effect sizes of the SNPs positively correlated for the association between cg21111824 and LOAD (annotated to CLU, chromosome 8).* ***(B)*** *The effect sizes of the SNPs negatively correlated for the association between cg01496416 and LOAD (annotated to PVR, chromosome 19). The colour of the SNPs corresponds to the LD strength (light blue = lower r^2, dark blue = higher r^2) of cis-mQTLs with the top cis-mQTL (red). Error bars are the standard errors of SNP effects. The orange dashed lines represents the estimate of bxy at the top cis-eQTL.*

Table 5.2: Summary data-based Mendelian Randomisation (SMR) test results for pleiotropic effects between LOAD and DNA methylation using the Kunkle at al. GWAS and the whole blood mQTL dataset (N = 26 hits where p < 3.5e-07 and HEIDI > 0.01*).* *CHR: chromosome; BP: base position (hg19); Instrument: top cis-mQTL for DNAm site with smallest p-value for association; Beta SMR: estimate for the effect of DNA methylation on LOAD; P SMR: p-value for SMR association; HEIDI: p-value for the HEIDI test; N SNP HEIDI; number of SNPs included in the HEIDI test.*

DNA site				**GWAS**		**SMR**			**HEIDI**	
Probe ID	**CHR**	**BP**	**Annotated Gene**	**Instrument (CHR:BP)**	**Ref MAF**	**Beta**	**SE**	**P**	**P**	**N SNP**
cg20172563	6	47487173	*CD2AP*	6:47525202	0.27	3.12	0.61	2.66E-07	1.96E-02	20
cg05908241	7	143109367	-	7:143109208	0.17	4.62	0.88	1.44E-07	8.14E-02	20
cg16292768	8	27467783	*CLU*	8:27467821	0.39	12.11	2.06	4.31E-09	1.93E-01	10
cg22217144	8	27468166	*CLU;MIR6843*	8:27467821	0.39	18.11	2.96	9.82E-10	7.34E-02	20
cg21111824	8	27468171	*CLU;MIR6843*	8:27467821	0.39	24.27	3.84	2.52E-10	7.46E-02	20
cg08871934	10	11720283	-	10:11720308	0.36	-1.96	0.38	2.68E-07	8.80E-02	20
cg25206146	11	47383181	*SPI1*	11:47391745	0.32	8.33	1.43	5.59E-09	4.54E-01	20
cg20449816	11	47432366	*SLC39A13*	11:47432034	0.32	3.47	0.57	1.08E-09	1.20E-01	20
cg27552578	11	47621330	-	11:47663049	0.35	6.93	1.25	3.10E-08	2.07E-02	20
cg05585544	11	47624801	-	11:47725306	0.34	8.25	1.55	1.11E-07	4.73E-02	20
cg17688768	11	47628441	-	11:47650138	0.35	1.58	0.28	1.72E-08	1.85E-02	20
cg18512352	11	47633146	-	11:47650138	0.35	4.09	0.73	2.64E-08	1.07E-02	20
cg02521229	11	60019236	-	11:60019161	0.41	0.90	0.12	7.12E-15	1.01E-02	20
cg18684128	11	60033393	-	11:60030559	0.41	7.97	1.22	6.58E-11	1.98E-01	20
cg18959616	11	85814918	-	11:85812210	0.36	2.47	0.35	1.00E-12	2.51E-01	20
cg23423086	11	85856245	-	11:85856187	0.32	-6.72	0.92	2.91E-13	2.46E-01	20
cg24251493	11	85866690	-	11:85868640	0.38	34.50	6.60	1.74E-07	8.83E-01	13
cg04441687	11	85869322	-	11:85868640	0.38	1.47	0.18	1.00E-15	2.28E-01	20
cg23484461	14	92936957	*SLC24A4*	14:92936690	0.20	1.39	0.27	3.29E-07	1.63E-01	20
cg01496416	19	45147715	*PVR*	19:45149235	0.24	-40.41	7.63	1.17E-07	3.43E-02	8
cg15233575	19	45221584	-	19:45233343	0.19	-27.59	4.44	5.15E-10	1.14E-01	3
cg03793277	19	45416910	*APOC1*	19:45416291	0.38	35.17	4.63	3.09E-14	1.96E-02	7
cg23270113	19	45417587	*APOC1*	19:45410444	0.38	-27.91	4.07	7.35E-12	2.94E-01	7

cg27353824	19	45445521	*APOC4*	19:45450033	0.46	19.25	3.51	4.20E-08	3.53E-02	11
cg25017250	19	45445693	*APOC4*	19:45448465	0.36	13.72	2.34	4.37E-09	1.45E-01	11
cg13766031	19	45457706	*CLPTM1*	19:45454759	0.46	17.27	2.73	2.54E-10	1.70E-02	15

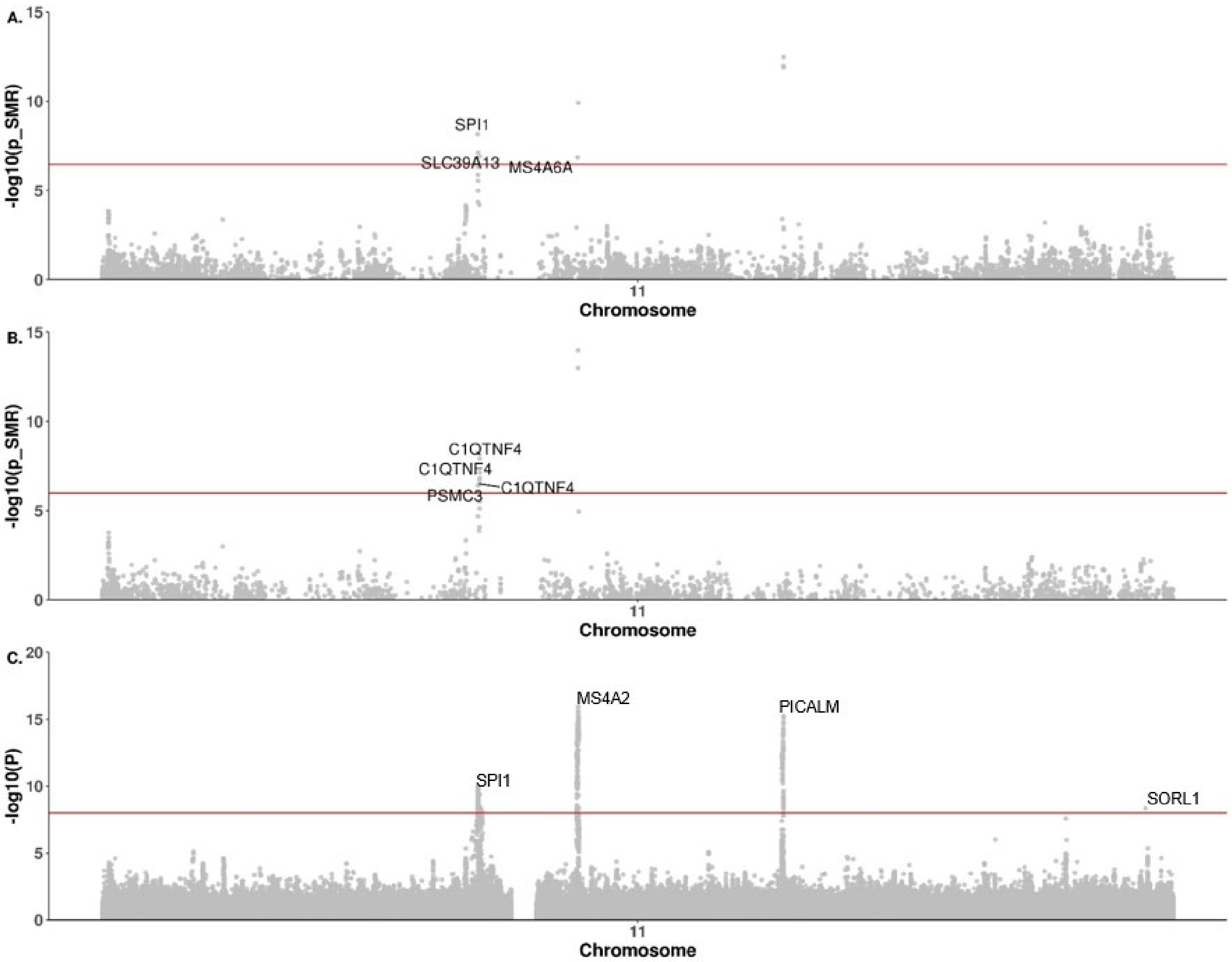


***Figure 5.15: Manhattan plots of Summary data-based Mendelian Randomisation (SMR) tests for pleiotropic effects between LOAD and DNA methylation across chromosome 11, using the Kunkle et al. AD GWAS results in combination with (A) the whole blood mQTL dataset; (B) the brain PFC mQTL dataset; and (C) the Kunkle et al GWAS results.** The x axis is the genomic position, segregated by chromosome. In **(A)** and **(B)** shown on the y-axis of each plot is the –log10 P-value from the SMR analysis using DNA methylation quantitative trait loci (mQTL). Each point represents an SMR test for a particular DNA methylation site. The red horizontal line represents the genome-wide multiple testing significance threshold. In **(C)** shown on the y-axis is the –log P-value from the GWAS. Each point represents a SNP. The red horizontal line represents the genome-wide significance threshold (P<5e-8).*

5.4.8.1.1 Identifying regions in the Kunkle whole blood SMR analysis

Since there is a correlation structure between SNPs and DNAm sites, I aimed to identify regions (defined as probes within 250kb of each other) within my SMR results, as there were several significant results around LOAD GWAS variants – e.g. the same genetic instrument is associated with three different DNAm sites located within the *CLU* gene (see **Table 5.3**). To investigate this, I used the Bayesian colocalisation results to identify if each pair of DNAm probes in *cis* had evidence for colocalisation. I identified 21 pairs of DNAm sites in *cis* across 5 regions within the Kunkle SMR results. All 21 pairs had evidence for a co-localized association within the same genomic region (PP3 + PP4 > 0.99). 13 pairs of sites had 'suggestive' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 1) (see **Table 5.3**). 12 pairs of sites had 'convincing' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 5) (**see Table 5.3**) which suggests these sites are genetically co-regulated by a mutual genetic variant. I identified co-localised sites around *CLU* and several along chromosome 11, including 7 pairs around the *SPI1* GWAS region, a pair around the *MSA4A2* GWAS region and a pair around the *PICALM* GWAS region (see **Table 5.2** and **Figure 5.15**). Although there were several SMR associations identified on chromosome 19 around the *APOE* locus, there was no evidence that these were genetically co-regulated using Bayesian colocalisation. This can also be seen by differing directions of effect from the SMR analysis (see **Table 5.2**) suggesting multiple independent SNPs may be having an effect on LOAD in this region mediated by DNAm at different sites. This goes in line with previous studies showing that multiple independent SNPs in this region have an independent influence on LOAD risk (Cervantes *at al.*, 2011).

Table 5.3: Bayesian Colocalisation results for the Kunkle et al whole blood SMR significant DNA methylation probes where there was evidence for colocalisation with the same causal mQTL variant. *There is evidence for colocalisation for four pairs of the DNA methylation probes which had a significant SMR result. The posterior probability all sites tested provided evidence for a co-localized association within the same genomic region (PP3 + PP4 > 0.99). 13 pairs of sites had 'suggestive' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 1). 12 pairs of sites had 'convincing' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 5). The strength of the evidence is based on co-localisation criteria developed by Guo and colleagues (Guo at al., 2015). PP.H_i.abf = Posterior probability for each of the Bayesian Colocalisation hypotheses (i).*

DNA methylation sites							**Bayesian Colocalisation results**							
Site 1	**CHR**	**BP**	**Annotated Gene**	**Site 2**	**BP**	**Annotated Gene**	**N snps**	**PP.H0.abf**	**PP.H1.abf**	**PP.H2.abf**	**PP.H3.abf**	**PP.H4.abf**	**P3 + P4**	**P4/P3**
cg22217144	8	27468166	*CLU;MIR6843*	cg16292768	27467783	*CLU*	2597	5.35e-29	1.92e-14	1.29e-17	3.65e-03	9.96e-01	1.00	273.28
cg21111824	8	27468171	*CLU;MIR6843*	cg16292768	27467783	*CLU*	2598	1.37e-32	1.55e-14	3.31e-21	2.75e-03	9.97e-01	1.00	362.95
cg21111824	8	27468171	*CLU;MIR6843*	cg22217144	27468166	*CLU;MIR6843*	2600	9.75e-36	1.10e-17	3.51e-21	2.98e-03	9.97e-01	1.00	334.97
cg20449816	11	47432366	*SLC39A13*	cg25206146	47383181	*SPI1*	1342	6.52e-142	3.55e-14	7.34e-129	0.40	0.60	1.00	1.51
cg05585544	11	47624801	-	cg27552578	47621330	-	1454	9.43e-52	1.97e-34	2.28e-19	0.05	0.95	1.00	20.40
cg17688768	11	47628441	-	cg27552578	47621330	-	1449	0	1.30e-34	0	0.03	0.97	1.00	31.92
cg17688768	11	47628441	-	cg05585544	47624801	-	1450	0	1.82e-19	0	0.04	0.96	1.00	25.90
cg18512352	11	47633146	-	cg27552578	47621330	-	1441	3.41e-121	1.10e-34	8.25e-89	0.03	0.97	1.00	38.13
cg18512352	11	47633146	-	cg05585544	47624801	-	1442	7.46e-106	2.40e-19	1.56e-88	0.05	0.95	1.00	19.32
cg18512352	11	47633146	-	cg17688768	47628441	-	1445	0	0	4.24e-89	0.01	0.99	1.00	78.10
cg18684128	11	60033393	-	cg02521229	60019236	-	1884	0	0	8.52e-19	0.08	0.92	1.00	11.27
cg23423086	11	85856245	-	cg18959616	85814918	-	2087	4.02e-56	2.10e-32	1.25e-25	0.06	0.94	1.00	14.53

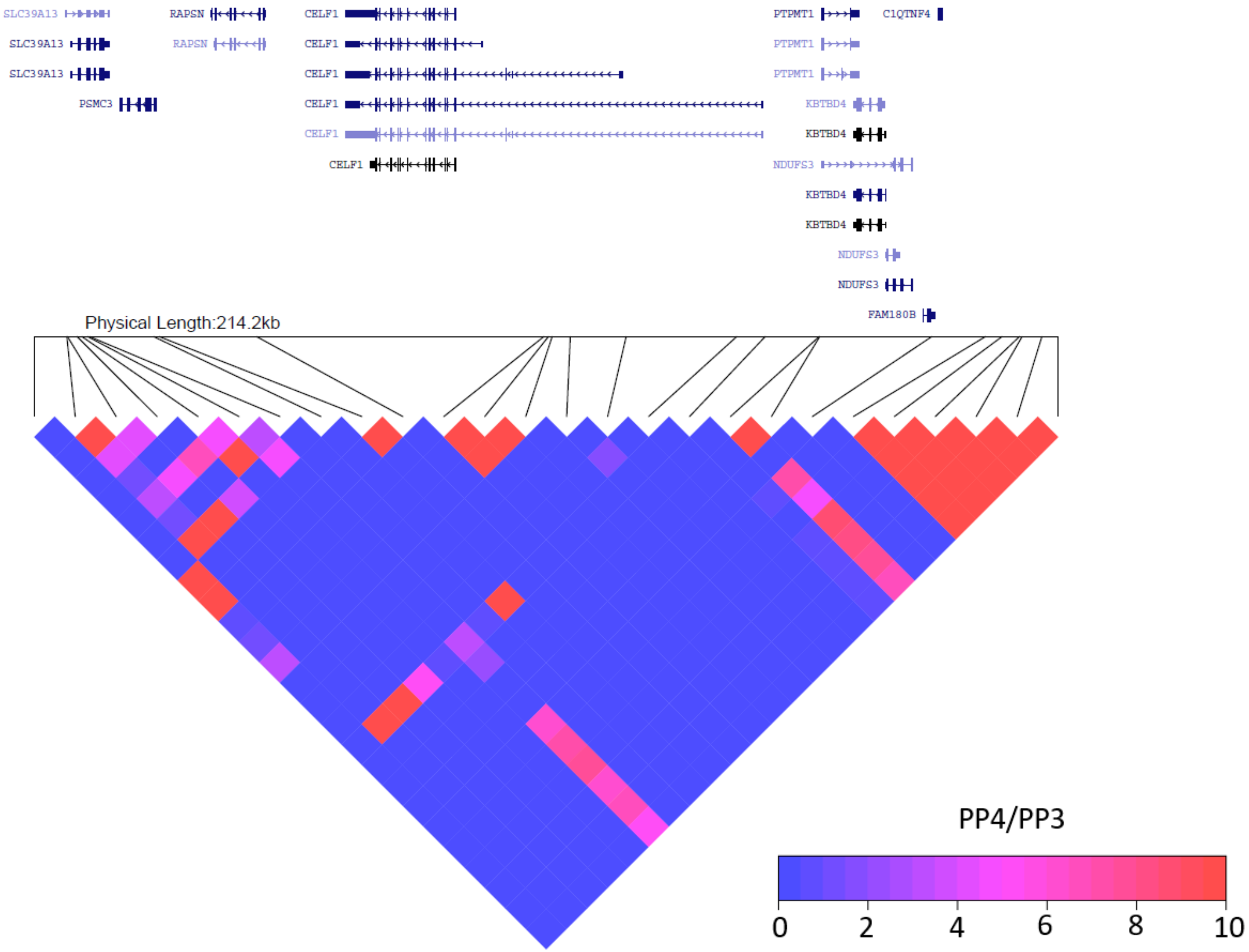


Figure 5.16: Heatmap of Bayesian co-localization results for all pairs of DNA-methylation sites with at least one significant mQTL (p < 1 × 10−10) in a genomic region on chromosome 11 (chr11: 47610098– 48028299). *Each square represents the relationship between two DNA-methylation sites (ordered by genomic location). The colour of each square indicates the strength of the evidence for a shared genetic signal (from blue [weak] to red [strong]). This strength is calculated as the ratio of the posterior probabilities that they share the same causal variant (PP4) compared to two distinct causal variants (PP3). The ratio was capped to a maximum value of 10.*

5.4.8.1.2 The SMR associations are relevant in the context of LOAD

Several of the DNAm sites prioritised by SMR as having pleiotropic associations with SNPs and LOAD are not annotated to a gene within the Illumina manifest. 13 of the DNAm sites do have gene annotations and have previously been implicated in LOAD.

Three of these sites were identified as being a region based on the Bayesian colocalisation analysis (i.e. evidence that these sites are genetically co-regulated): cg16292768, cg22217144, cg21111824 which are located on chromosome 8 and annotated to *CLU* (see **Table 5.3**). These sites were positively associated with LOAD (p=4.31e-09; p=9.82e-10; and p=2.52e-10 (see **Figure 5.14**A), respectively; see **Table 5.2**). *CLU* is an AD risk gene, is involved in brain function during ageing and studies have demonstrated that *CLU* risk carriers have increased rates of cognitive decline (Thambisetty *at al.*, 2013).

Six sites on chromosome 11 were identified as being a region based on the Bayesian colocalisation results (see **Table 5.3**). cg25206146 located on chromosome 11, annotated to *SPI1* - which is the nearest gene to one of the top LOAD GWAS loci (Kunkle, 2019) - was positively associated with LOAD (p= 5.59e-09; see Table 5.2). cg20449816, located on chromosome 11, annotated to *SLC39A13* was positively associated with LOAD (p= 1.08e-09; see **Table 5.2**). This gene is found in the LD block of *SPI1*. cg27552578, cg05585544, cg17688768 and cg18512352 were all positively associated with LOAD (p=3.10e-08; p=1.11e-07; p=1.72e-08; and p=2.64e-08, respectively). These four sites do not have a gene annotation based on the Illumina manifest, however the Bayesian colocalisation results suggest they are genetically co-regulated with DNAm sites associated around the *SPI1* region.

Several sites were not identified as regions within the Bayesian colocalisation analysis but had a significant SMR result including:

- cg20172563 located on chromosome 6 and annotated to *CD2AP* was positively associated with LOAD (p=2.66e-07; see **Table 5.2**). *CD2AP* is an AD risk gene. Evidence suggests *CD2AP* increases AD risk through tau-induced neurotoxicity, Aβ processing, abnormal neurite structure modulation and blood-brain barrier dysfunction (Qing-Qing, Yu-Chao, & Zhi-Ying, 2018).

- cg23484461, located on chromosome 14 and annotated to *SLC24A4* (see **Figure 5.17**) was positively associated with LOAD (p= 3.29e-07; see **Table 5.2**). *SLC24A4* encodes a member of the potassium-dependent sodium/calcium exchanger protein family. *SLC24A4* has previously been associated with AD in GWAS and evidence suggests this gene also plays a role in cognitive ageing (Yu *at al.*, 2015). In addition, differential methylation in the PFC of a DNAm site located within the *SLC24A4* locus has been associated with LOAD pathology (Yu *at al.*, 2015).
- cg01496416, located on chromosome 19, annotated to PVR was negatively associated with LOAD (p=1.17e-07; see **Table 5.2** and **Figure 5.14B**). PVR is a LOAD-risk gene which resides near the *APOE* region. Previous research suggests this gene is potentially mediated by both gene expression and DNAm in the PFC (Marioni *at al.*, 2018).
- cg03793277, cg27353824, and cg25017250, located on chromosome 19 were positively associated with LOAD (p= 3.09e-14; p= 4.20e-08; and p= 4.37e-09, respectively; see **Table 5.2**). DNAm at cg23270113 was negatively associated with LOAD (p= 7.35e-12; see **Table 5.2**). These sites are annotated to the loci *APOC1* and APOC4, which both reside near the *APOE* region and are known are *APOE* cluster genes. *APOC1* is known to facilitate dementia under oxidative stress (Prendecki *at al.*, 2018) and a SNP located in APOC4 has been associated with increased AD risk independently from *APOE* ε4 (Cervantes *at al.*, 2011).

Many of the DNAm sites without a current gene annotation based on the illumina manifest were located close to genes (within 250kb) which have been implicated in LOAD. I identified two regions within these results based on the Bayesian colocalisation results. There was evidence that cg02521229 and cg18684128 were genetically co-regulated (see **Table 5.2**). These sites are located on chromosome 11 near (~50kb away) the *MS4A6A* locus, which is the nearest gene to a leading LOAD GWAS loci, were positively associated with LOAD (p = 7.12e-15; p = 6.58e-11, respectively; see Table 5.1). *MS4A6A* is part of the MS4A locus which contains several genes implicated in immune modulation (Villegas-Llerena, Phillips, Garcia-Reitboeck, Hardy, & Pocock, 2016). *MS4A6A* is highly expressed in microglia and is thought to play a role insoluble *TREM2* production. *MS4A6A* expression in the parietal lobe is associated with more advanced cortex pathology in AD patients (Villegas-Llerena *at al.*, 2016).

There was evidence that cg24251493 and cg04441687 were also genetically co-regulated based on the Bayesian-colocalisation results (see **Table 5.2**). These sites were positively associated with LOAD (p=1.74e-07 and p=1.00e-15, respectively) and are located near the *PICALM* region (85kb away). *PICALM* is a LOAD risk gene identified by GWAS (Jansen *at al.*, 2019; Kunkle *at al.*, 2019).

Several of the unannotated sites were not genetically co-regulated with another SMR significant DNAm site. This includes cg05908241, located on chromosome 7 which is found ~3kb from the *EPHA1* locus (see **Figure 5.18**), which was positively associated with LOAD (p= 3.29e-07; see **Table 5.2**). *EPHA1* is a LOAD GWAS gene and evidence supports a role for the regulation of this gene via eQTLs in whole blood (Liu *at al.*, 2018). Additionally, cg08871934 located on chromosome 10 and is found ~3kb away from the *ECHDC3* of the locus (see **Figure 5.18**), was negatively associated with LOAD (p=2.68e-07). *ECHDC3* is LOAD GWAS gene (Jansen *at al.*, 2019; Kunkle *at al.*, 2019)

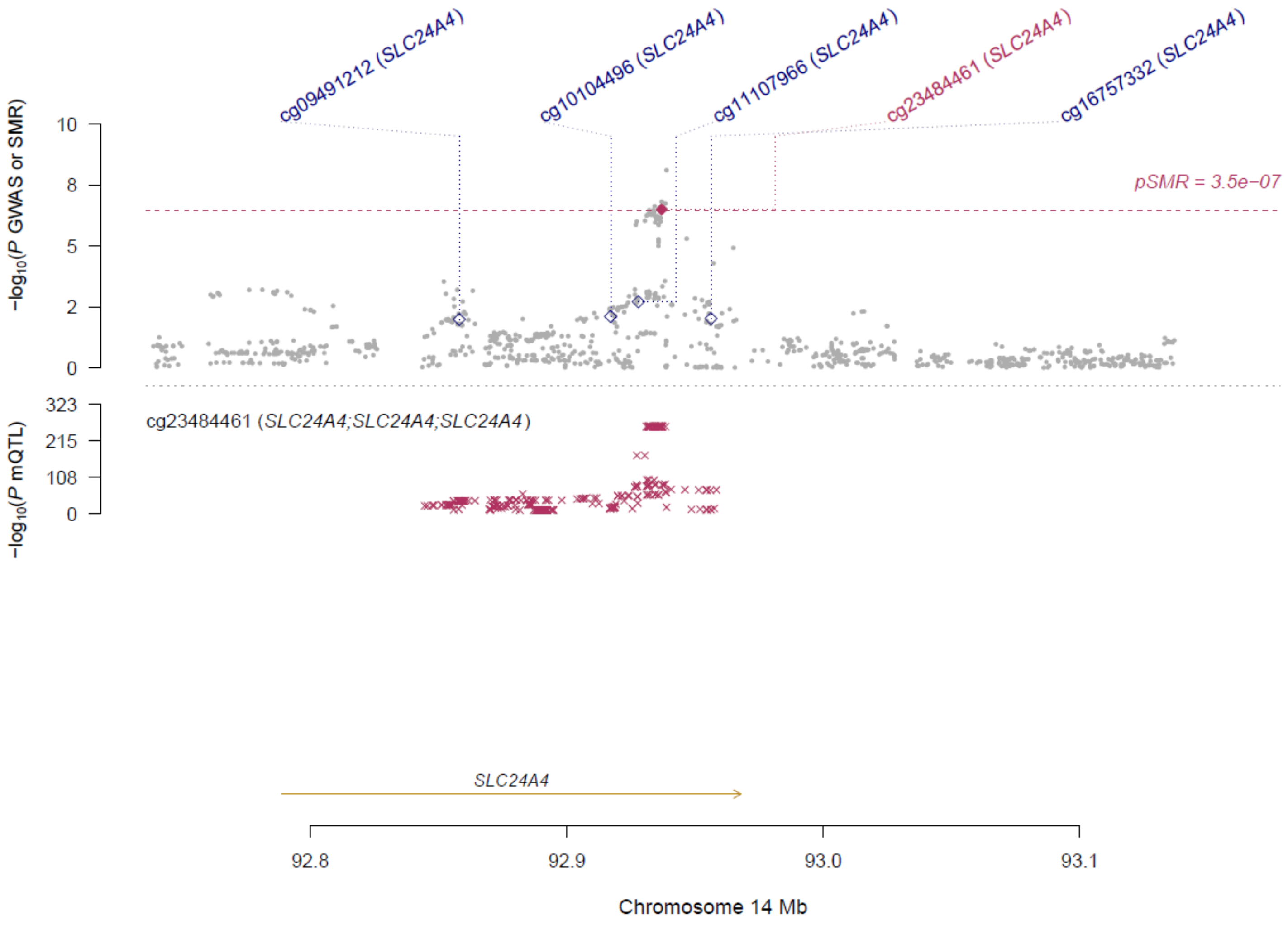


***Figure 5.17: Prioritizing genes at the SLC24A4 GWAS locus using SMR analysis with whole blood mQTLs and the Kunkle et al LOAD GWAS**. Shown are results at the SLC24A4 locus (chromosome 14) for LOAD. Top plot, grey dots represent the P values for SNPs from the Kunkle et al LOAD GWAS, diamonds represent the P values for probes from the SMR test – those filled in and highlighted in red passed the SMR and HEIDI tests. Bottom plots, the mQTL P values of SNPs from the whole blood dataset for the probes around the SLC24A4 locus. The top and bottom plots include all the SNPs available in the region in the GWAS and mQTL summary data, respectively, not just the SNPs common to both datasets.*

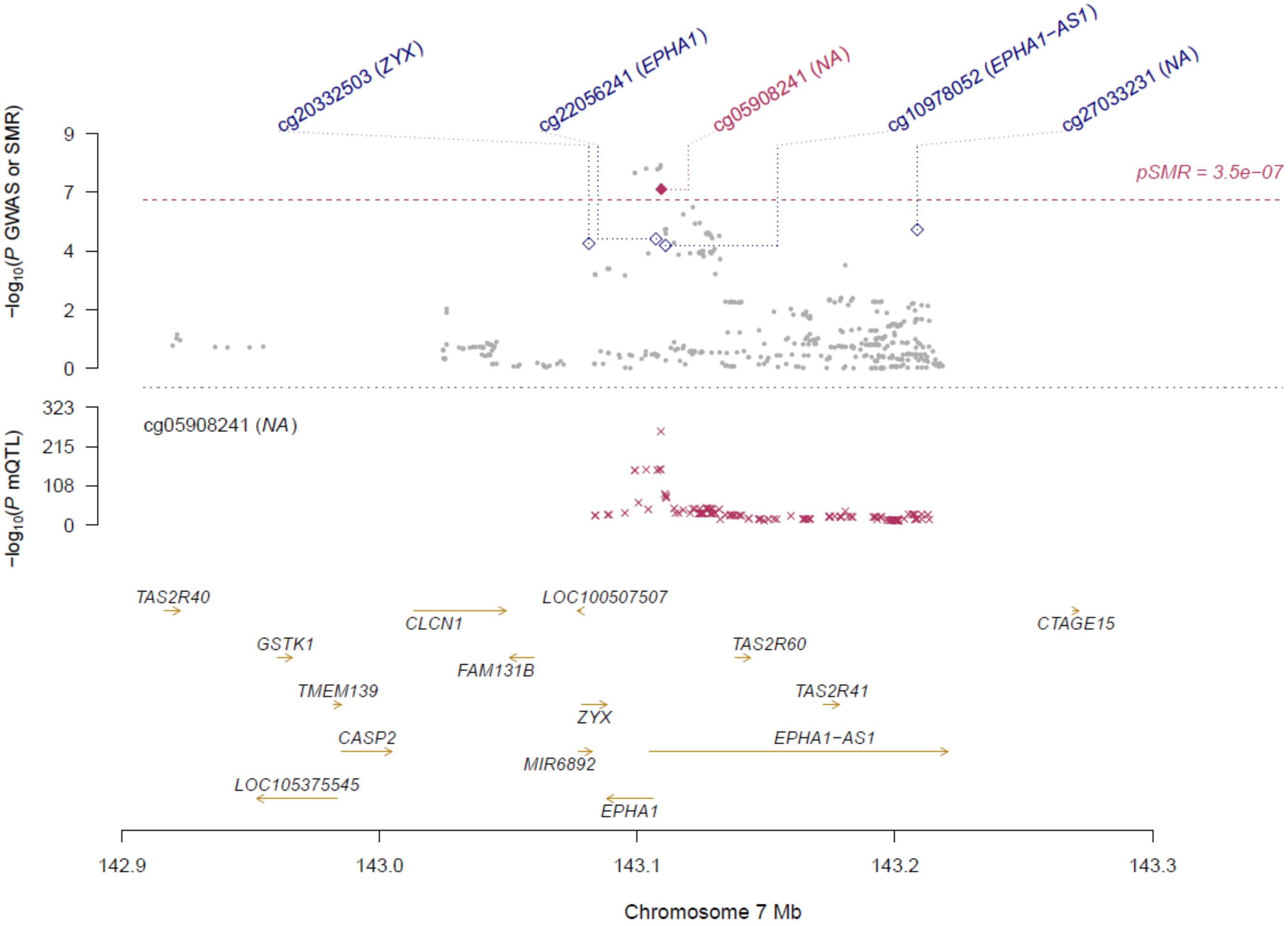


Figure 5.18: Prioritizing genes at the EPHA1 GWAS locus using SMR analysis with whole blood mQTLs and the Kunkle et al LOAD GWAS*. Shown are results at the EPHA1 locus (chromosome 7) for LOAD. Top plot, grey dots represent the P values for SNPs from the Kunkle et al LOAD GWAS, diamonds represent the P values for probes from the SMR test – those filled in and highlighted in red passed the SMR and HEIDI tests. Bottom plots, the mQTL P values of SNPs from the whole blood dataset for the probes around the EPHA1 locus. The top and bottom plots include all the SNPs available in the region in the GWAS and mQTL summary data, respectively, not just the SNPs common to both datasets.*

5.4.8.1.3 Pathway analysis

In order to relate the identified DNAm sites prioritised by SMR to biological functions I ran gene ontology (GO) pathway analysis using the R package methylGSA. I ran the analysis including probes which reached the threshold $p<5e-5$ (92 probes, tagging 43 genes). I identified 287 significant pathways (see **Figure 5.19**) including several lipid related pathways such as cholesterol metabolism and transport. Lipids are involved in *APP* processing and trafficking and influence the formation of amyloid-beta (Aβ) peptides which are involved in AD pathogenesis (Penke *at al.*, 2018). In addition, multiple amyloid and tau related pathways were identified including Aβ binding and tau binding pathways.

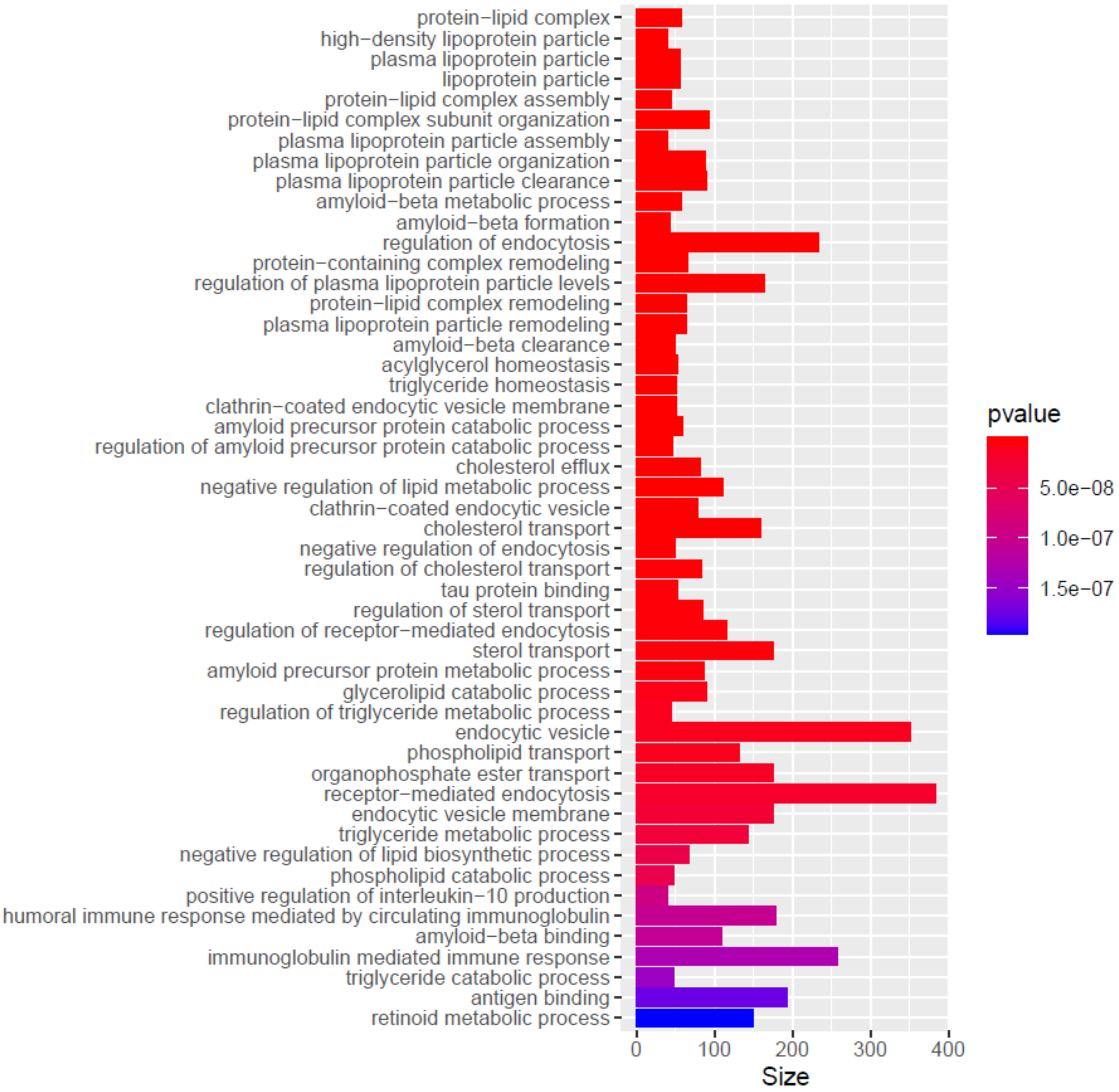


Figure 5.19: Pathway analysis of the whole Blood mQTL SMR results which incorporated the Kunkle et al (2019) GWAS. *Shown are the top 50 significant result from the pathway analysis. The y-axis is the gene ontology pathway. The x axis is size, which represents the number of genes included in the pathway.*

5.4.8.2 SMR results using the Jansen *at al.* GWAS

As the power of the GWAS has a large influence on the ability of the SMR approach to detect pleiotropic associations, I identified more pleiotropic associations when using the Jansen *at al.* GWAS which incorporated both clinically diagnosed AD cases and AD-by-proxy (71,880 cases, 383,378 cognitively normal controls). I applied SMR and identified 106 associations which passed the Bonferroni significant threshold (p < 3.5e-07) and 48 which passed both the significance threshold and the HEIDI threshold (p >

0.01; see **Table 5.4** and **Figure 5.13).** Of these 48, 24 (50%) were positively associated with LOAD. Like for the Kunkle SMR analysis, many of the associations were identified in regions proximal to GWAS peaks, including several along chromosome 11, where 4 distinct GWAS loci have previously been identified which are annotated to the genes *SPI1*, MS4A2, *PICALM* and *SORL1* (see **Figure 5.20**).

Of the 48 significant probes which passed the SMR and HEIDI analysis, 32 (67%) were annotated to genes. Seven of these genes had multiple tagging probes, providing internally consistent - although not necessarily independent - replications of the results. This included DNAm sites annotated to the genes: BTNL2, *CD2AP*, *GPC2*, *STAG3*, *MS4A3*, *ADAM10*, *SCIMP*, C17orf87 and PVR which have all been implicated in LOAD (Bellenguez *at al.*, 2020; Jansen *at al.*, 2019; Kunkle *at al.*, 2019; Rosenthal & Kamboh, 2014).

Table 5.4: Summary data-based Mendelian Randomisation (SMR) test results for pleiotropic effects between LOAD and DNA methylation using the Jansen at al. GWAS and the whole blood mQTL dataset. (N = 48 hits where p < 3.5e-07 and HEIDI > 0.01). *CHR: chromosome; BP: base position (hg19 P); Instrument: top cis-mQTL for DNAm site with smallest p-value for association; Beta SMR: estimate for the effect of DNAm on LOAD; P SMR: p-value for SMR association; HEIDI: p-value for the HEIDI test; N SNP HEIDI; number of SNPs included in the HEIDI test.*

DNA site				**GWAS**		**SMR**			**HEIDI**	
Probe ID	**CHR**	**BP**	**Annotated Gene**	**Instrument (CHR:BP)**	**Ref MAF**	**Beta**	**SE**	**P**	**P**	**N SNP**
cg03100814	6	32367672	*BTNL2*	6:32376746	0.32	-0.91	0.174	1.91E-07	7.73E-02	20
cg14241129	6	32367729	*BTNL2*	6:32368087	0.32	-0.75	0.140	8.28E-08	5.72E-02	20
cg12672189	6	32427868	-	6:32423194	0.16	-0.13	0.023	5.54E-08	8.07E-02	20
cg20946741	6	32428328	-	6:32418657	0.16	-0.22	0.039	3.28E-08	2.19E-02	20
cg10556520	6	32428377	-	6:32423194	0.16	-0.36	0.068	1.06E-07	2.06E-01	20
cg08265274	6	32490444	*HLA-DRB5*	6:32681277	0.36	-0.05	0.010	1.49E-07	1.59E-02	20
cg15710545	6	32578114	-	6:32590331	0.32	0.09	0.017	1.44E-07	3.91E-02	20
cg03149641	6	47444455	*CD2AP*	6:47552180	0.27	0.30	0.054	1.55E-08	3.87E-02	20
cg20172563	6	47487173	*CD2AP*	6:47525202	0.27	0.50	0.092	4.22E-08	9.76E-02	20
cg18090197	7	99769602	*GPC2*	7:99792608	0.26	1.76	0.336	1.60E-07	2.07E-02	20
cg00048759	7	99775422	*STAG3;GPC2*	7:99807146	0.26	-0.82	0.157	1.64E-07	1.01E-01	20
cg10084644	7	99775521	*STAG3;GPC2*	7:99816179	0.26	-0.67	0.117	1.12E-08	3.91E-02	20
cg00553149	7	99775558	*STAG3;GPC2*	7:99807146	0.26	-0.38	0.066	6.98E-09	3.90E-02	20
cg10407106	7	99779719	*STAG3*	7:99784704	0.26	0.79	0.136	7.95E-09	4.28E-02	20
cg17830204	7	99819110	*GATS;PVRIG*	7:99787372	0.26	0.50	0.081	5.59E-10	1.46E-02	20
cg19116668	7	99932089	*PMS2L1*	7:99971834	0.31	0.73	0.100	2.61E-13	4.04E-01	20
cg03579757	7	100091793	*NYAP1*	7:100012579	0.30	0.91	0.133	7.26E-12	3.55E-01	20
cg03760621	7	143104506	*EPHA1-AS1;EPHA1*	7:143104331	0.41	1.05	0.178	3.00E-09	2.10E-02	20
cg18997129	7	143105850	*EPHA1*	7:143104331	0.41	1.77	0.302	4.82E-09	1.74E-02	20
cg16292768	8	27467783	*CLU*	8:27467821	0.39	1.84	0.300	9.27E-10	1.15E-01	10
cg22217144	8	27468166	*CLU;MIR6843*	8:27467821	0.39	2.74	0.429	1.60E-10	4.83E-02	20
cg21111824	8	27468171	*CLU;MIR6843*	8:27467821	0.39	3.68	0.554	3.10E-11	3.35E-02	20
cg08871934	10	11720283	-	10:11720308	0.36	-0.31	0.056	3.21E-08	4.25E-01	20

cg17173423	11	59823993	*MS4A3*	11:59852569	0.28	1.18	0.197	1.89E-09	1.70E-02	20
cg01440285	11	59837091	*MS4A3*	11:59826677	0.31	1.06	0.190	2.66E-08	2.81E-02	20
cg18684128	11	60033393	-	11:60030559	0.41	1.05	0.175	1.64E-09	2.18E-02	20
cg18959616	11	85814918	-	11:85812210	0.36	0.39	0.051	1.84E-14	7.06E-02	20
cg01904978	11	85847072	-	11:85845473	0.39	1.78	0.343	2.32E-07	2.50E-02	20
cg16209351	11	85850230	-	11:85858497	0.47	4.80	0.932	2.56E-07	2.21E-01	20
cg19619504	11	85850481	-	11:85858497	0.47	1.50	0.229	5.98E-11	3.47E-02	20
cg24251493	11	85866690	-	11:85868640	0.38	5.36	1.005	9.94E-08	3.56E-01	13
cg04441687	11	85869322	-	11:85868640	0.38	0.23	0.027	3.97E-17	3.23E-01	20
cg10816016	11	85873480	-	11:85854176	0.44	3.68	0.699	1.35E-07	6.34E-02	20
cg23484461	14	92936957	*SLC24A4*	14:92936690	0.20	0.24	0.039	1.68E-09	2.75E-02	20
cg20288868	15	59042462	*ADAM10*	15:59016315	0.28	-0.95	0.177	6.90E-08	2.71E-01	20
cg15770593	15	59042482	*ADAM10*	15:59042012	0.28	-0.72	0.133	4.74E-08	1.49E-01	20
cg20532450	15	59042728	*ADAM10*	15:59016315	0.28	-0.73	0.141	2.56E-07	8.11E-02	20
cg09265987	15	59050614	-	15:59052210	0.30	-0.47	0.084	3.19E-08	1.33E-01	20
cg18702655	17	5138293	*SCIMP;LOC100130950*	17:5158714	0.12	-0.78	0.136	1.15E-08	6.75E-01	20
cg10520397	17	5138461	*SCIMP;LOC100130950*	17:5155919	0.12	-1.10	0.203	6.03E-08	6.31E-01	20
cg03146371	17	5138469	*SCIMP;LOC100130950*	17:5158714	0.12	-0.98	0.180	4.75E-08	7.38E-01	20
cg21337881	17	5138645	*C17orf87*	17:5155919	0.12	-1.29	0.231	2.25E-08	3.99E-01	20
cg17588003	17	5138696	*C17orf87*	17:5155919	0.12	-0.70	0.124	1.45E-08	5.33E-01	20
cg03167326	17	5158046	-	17:5155919	0.12	-0.29	0.050	3.96E-09	4.17E-01	20
cg12439163	17	5165803	-	17:5155919	0.12	-0.55	0.096	6.53E-09	2.88E-01	20
cg23619370	17	5168138	-	17:5155919	0.12	-0.48	0.082	5.14E-09	1.80E-01	20
cg15233575	19	45221584	-	19:45233343	0.19	-4.11	0.669	7.99E-10	5.04E-02	3
cg02400211	19	51727987	*CD33*	19:51736383	0.33	-0.52	0.099	1.00E-07	1.00E-01	20

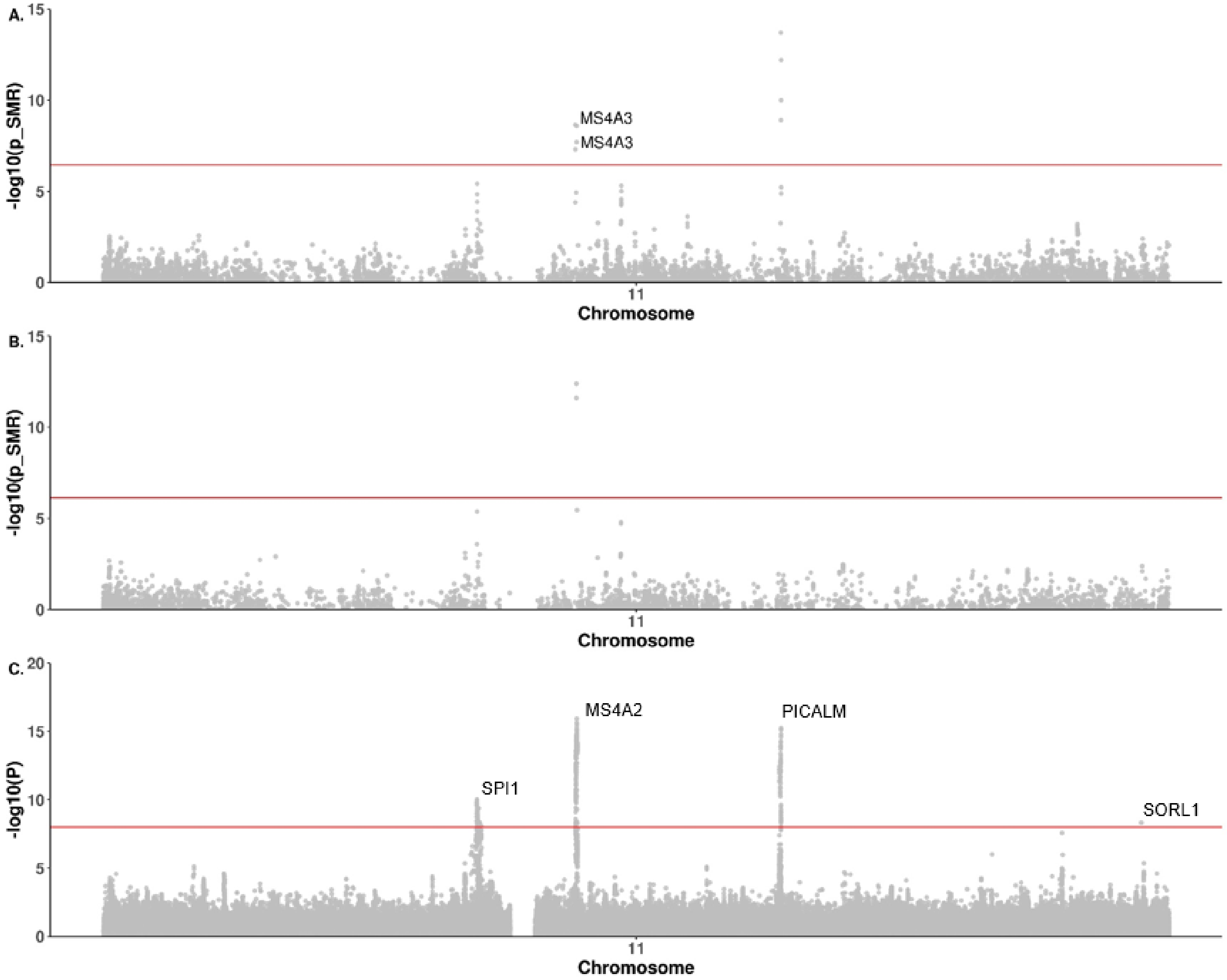


Figure 5.20: Manhattan plots of Summary data-based Mendelian Randomisation (SMR) tests for pleiotropic effects between LOAD and DNA methylation, using the Jansen et al. GWAS along chromosome 11 *in* ***(A)*** *the whole blood mQTL dataset;* ***(B)*** *the brain PFC mQTL dataset; and* ***(C)*** *the Jansen et al GWAS results. The x axis is the genomic position, segregated by chromosome. In* ***(A)*** *and* ***(B)*** *shown on the y-axis of each plot is the –log10 P-value from the SMR analysis using DNA methylation quantitative trait loci (mQTL). Each point represents an SMR test for a particular DNA methylation site. The red horizontal line represents the genome-wide multiple testing significance threshold. In* ***(C)*** *shown on the y-axis is the –log P-value from the GWAS. Each point represents a SNP. The red horizontal line represents the genome-wide significance threshold (P<5e-8).*

5.4.8.2.1 Identifying regions in the Jansen whole blood SMR analysis

I used the Bayesian colocalisation results to identify if each pair of DNAm probes in the same region had evidence for colocalisation, in total examining 82 pairs of sites. All 82 pairs had evidence for a co-localized association within the same genomic region (PP3 + PP4 > 0.99). 57 pairs of DNAm sites had 'suggestive' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 1). 56 pairs of sites had 'convincing' evidence to support associations between both DNAm sites with the same causal variant (see **Table 5.5**). This includes a region of 28 pairs of sites around the *SCIMP* locus, a region of 16 pairs of sites around the *ZCWPW1* locus, a region of 6 pairs of sites around the *ADAM10* locus, a region of three sites at the *CLU* locus, a region of two pairs of sites around the *CD2AP* locus, a pair of sites at the *MS4A3* locus and a pair of sites at the *EPHA1* locus (see **Table 5.5**). This suggests there is genetic co-regulation within multiple regions and therefore we can use these data to aid the interpretation of the SMR results by grouping results. Several of these regions were also identified in the Kunkle et al analysis, including sites annotated to *CLU* and *MS4A3* (for a more extensive comparison between analyses see **5.3.10**).

Table 5.5: Bayesian Colocalisation results for the Jansen et al whole blood SMR significant DNA methylation probes. *There is evidence for colocalisation for several of the DNA methylation probes which had a significant SMR result. The posterior probability for all pairs of sites tested provided evidence for a co-localized association within the same genomic region (PP3 + PP4 > 0.99). 55 pairs of sites had 'suggestive' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 1). 54 pairs of sites had 'convincing' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 5). The strength of the evidence is based on co-localisation criteria developed by Guo and colleagues (Guo at al., 2015). PP.*H_i*.abf = Posterior probability for each of the Bayesian Colocalisation hypotheses (i).*

DNA methylation sites							**Bayesian Colocalisation results**							
Site 1	**CHR**	**BP**	**Gene**	**Site 2**	**BP**	**Gene**	**N snps**	**PP.H0.abf**	**PP.H1.abf**	**PP.H2.abf**	**PP.H3.abf**	**PP.H4.abf**	**P3 + P4**	**P4/P3**
cg20172563	6	47487173	*CD2AP*	cg03149641	47444455	*CD2AP*	2277	9.60E-155	2.51E-94	1.91E-62	0.0490	0.951	1	19.40
cg20946741	6	32428328	-	cg12672189	32427868		8138	1.12E-133	3.50E-81	3.90E-54	1.21E-01	8.79E-01	1	7.26
cg18997129	7	143105850	*EPHA1*	cg03760621	143104506	*EPHA1-AS1*	628	4.10E-28	2.75E-19	3.77E-11	2.43E-02	9.76E-01	1	40.07
cg17830204	7	99819110	GATS	cg18090197	99769602	*GPC2*	994	1.22E-71	5.37E-09	2.03E-64	8.83E-02	9.12E-01	1	10.33
cg17830204	7	99819110	GATS	cg00048759	99775422	*STAG3*	998	1.47E-71	6.44E-09	2.21E-64	9.61E-02	9.04E-01	1	9.41
cg17830204	7	99819110	GATS	cg10084644	99775521	*STAG3*	998	4.43E-85	1.94E-22	2.76E-64	1.20E-01	8.80E-01	1	7.33
cg17830204	7	99819110	GATS	cg00553149	99775558	*STAG3*	998	1.58E-108	6.91E-46	3.24E-64	1.41E-01	8.59E-01	1	6.08
cg17830204	7	99819110	GATS	cg10407106	99779719	*STAG3*	1003	4.46E-82	1.96E-19	1.21E-64	5.22E-02	9.48E-01	1	18.15
cg03579757	7	100091793	*NYAP1*	cg19116668	99932089	*PMS2L1*	916	2.24E-53	7.17E-30	5.21E-25	1.66E-01	8.34E-01	1	5.02
cg00048759	7	99775422	*STAG3*	cg18090197	99769602	*GPC2*	1037	4.32E-16	6.50E-09	7.17E-09	1.07E-01	8.93E-01	1	8.34
cg10084644	7	99775521	*STAG3*	cg18090197	99769602	*GPC2*	1037	1.03E-29	6.40E-09	1.71E-22	1.05E-01	8.95E-01	1	8.49
cg10084644	7	99775521	*STAG3*	cg00048759	99775422	*STAG3*	1041	7.48E-30	4.66E-09	1.13E-22	6.91E-02	9.31E-01	1	13.46
cg00553149	7	99775558	*STAG3*	cg18090197	99769602	*GPC2*	1037	2.98E-53	6.12E-09	4.95E-46	1.01E-01	8.99E-01	1	8.92
cg00553149	7	99775558	*STAG3*	cg00048759	99775422	*STAG3*	1041	2.31E-53	4.76E-09	3.48E-46	7.06E-02	9.29E-01	1	13.16
cg00553149	7	99775558	*STAG3*	cg10084644	99775521	*STAG3*	1041	4.15E-67	8.54E-23	2.58E-46	5.22E-02	9.48E-01	1	18.15
cg10407106	7	99779719	*STAG3*	cg18090197	99769602	*GPC2*	1037	2.09E-26	5.68E-09	3.47E-19	9.35E-02	9.06E-01	1	9.69
cg10407106	7	99779719	*STAG3*	cg00048759	99775422	*STAG3*	1041	2.10E-26	5.71E-09	3.16E-19	8.51E-02	9.15E-01	1	10.76
cg10407106	7	99779719	*STAG3*	cg10084644	99775521	*STAG3*	1041	4.73E-40	1.29E-22	2.95E-19	7.92E-02	9.21E-01	1	11.63
cg10407106	7	99779719	*STAG3*	cg00553149	99775558	*STAG3*	1041	1.42E-63	3.86E-46	2.92E-19	7.85E-02	9.22E-01	1	11.74

cg22217144	8	27468166	*CLU*;MIR6843	cg16292768	27467783	*CLU*	2597	5.35E-29	1.92E-14	1.29E-17	3.65E-03	9.96E-01	1	273.28
cg21111824	8	27468171	*CLU*;MIR6843	cg16292768	27467783	*CLU*	2598	1.37E-32	1.55E-14	3.31E-21	2.75E-03	9.97E-01	1	362.95
cg21111824	8	27468171	*CLU*;MIR6843	cg22217144	27468166	*CLU;MIR6843*	2600	9.75E-36	1.10E-17	3.51E-21	2.98E-03	9.97E-01	1	334.97
cg01440285	11	59837091	*MS4A3*	cg17173423	59823993	*MS4A3*	1775	4.79E-24	2.47E-20	3.01E-05	1.54E-01	8.46E-01	1	5.48
cg15770593	15	59042482	*ADAM10*	cg20288868	59042462	*ADAM10*	1803	2.97E-125	6.48E-28	5.25E-99	1.14E-01	8.86E-01	1	7.78
cg20532450	15	59042728	*ADAM10*	cg20288868	59042462	*ADAM10*	1803	1.25E-40	1.79E-28	2.22E-14	3.06E-02	9.69E-01	1	31.65
cg20532450	15	59042728	*ADAM10*	cg15770593	59042482	*ADAM10*	1803	2.22E-112	3.16E-100	4.85E-15	5.92E-03	9.94E-01	1	167.97
cg09265987	15	59050614	-	cg20288868	59042462	*ADAM10*	1777	2.47E-51	1.62E-27	4.36E-25	2.86E-01	7.14E-01	1	2.5
cg09265987	15	59050614	-	cg15770593	59042482	*ADAM10*	1777	3.16E-123	2.08E-99	6.90E-26	4.44E-02	9.56E-01	1	21.51
cg09265987	15	59050614	-	cg20532450	59042728	*ADAM10*	1777	4.33E-38	2.84E-14	6.17E-26	3.96E-02	9.60E-01	1	24.27
cg21337881	17	5138645	C17orf87	cg18702655	5138293	*SCIMP*	2337	2.28E-48	4.77E-29	4.22E-22	7.83E-03	9.92E-01	1	126.74
cg21337881	17	5138645	C17orf87	cg10520397	5138461	*SCIMP*	2337	1.78E-32	3.71E-13	1.21E-21	2.44E-02	9.76E-01	1	40.01
cg21337881	17	5138645	C17orf87	cg03146371	5138469	*SCIMP*	2337	4.30E-32	8.98E-13	5.30E-22	1.01E-02	9.90E-01	1	98.17
cg17588003	17	5138696	C17orf87	cg18702655	5138293	*SCIMP*	2337	5.66E-57	9.46E-29	1.05E-30	1.65E-02	9.83E-01	1	59.49
cg17588003	17	5138696	C17orf87	cg10520397	5138461	*SCIMP*	2337	2.45E-41	4.09E-13	1.67E-30	2.70E-02	9.73E-01	1	36.1
cg17588003	17	5138696	C17orf87	cg03146371	5138469	*SCIMP*	2337	4.50E-41	7.52E-13	5.55E-31	8.29E-03	9.92E-01	1	119.69
cg17588003	17	5138696	C17orf87	cg21337881	5138645	*C17orf87*	2338	3.06E-50	5.11E-22	6.39E-31	9.69E-03	9.90E-01	1	102.18
cg10520397	17	5138461	*SCIMP*	cg18702655	5138293	*SCIMP*	2339	1.71E-39	1.17E-28	3.17E-13	2.07E-02	9.79E-01	1	47.4
cg03146371	17	5138469	*SCIMP*	cg18702655	5138293	*SCIMP*	2339	4.79E-39	5.91E-29	8.86E-13	9.94E-03	9.90E-01	1	99.57
cg03146371	17	5138469	*SCIMP*	cg10520397	5138461	*SCIMP*	2339	3.48E-23	4.29E-13	2.37E-12	2.83E-02	9.72E-01	1	34.32
cg03167326	17	5158046	-	cg18702655	5138293	*SCIMP*	2310	8.70E-155	3.68E-29	1.61E-128	5.81E-03	9.94E-01	1	171.06
cg03167326	17	5158046	-	cg10520397	5138461	*SCIMP*	2310	6.65E-139	2.81E-13	4.54E-128	1.82E-02	9.82E-01	1	53.91
cg03167326	17	5158046	-	cg03146371	5138469	*SCIMP*	2310	2.39E-138	1.01E-12	2.95E-128	1.15E-02	9.88E-01	1	85.91
cg03167326	17	5158046	-	cg21337881	5138645	*C17orf87*	2311	7.55E-148	3.19E-22	1.58E-128	5.68E-03	9.94E-01	1	175.1
cg03167326	17	5158046	-	cg17588003	5138696	*C17orf87*	2312	1.43E-156	6.04E-31	2.39E-128	9.11E-03	9.91E-01	1	108.81
cg12439163	17	5165803	-	cg18702655	5138293	*SCIMP*	2304	7.30E-100	5.27E-29	1.35E-73	8.77E-03	9.91E-01	1	113.04

cg12439163	17	5165803	-	cg10520397	5138461	*SCIMP*	2304	4.21E-84	3.04E-13	2.88E-73	1.98E-02	9.80E-01	1	49.49
cg12439163	17	5165803	-	cg03146371	5138469	*SCIMP*	2304	1.22E-83	8.82E-13	1.50E-73	9.89E-03	9.90E-01	1	100.16
cg12439163	17	5165803	-	cg21337881	5138645	*C17orf87*	2305	4.67E-93	3.38E-22	9.77E-74	6.06E-03	9.94E-01	1	163.91
cg12439163	17	5165803	-	cg17588003	5138696	*C17orf87*	2306	7.30E-102	5.27E-31	1.22E-73	7.83E-03	9.92E-01	1	126.75
cg12439163	17	5165803	-	cg03167326	5158046		2345	1.34E-199	9.68E-129	5.66E-74	3.10E-03	9.97E-01	1	322.01
cg23619370	17	5168138	-	cg18702655	5138293	*SCIMP*	2302	9.91E-101	6.55E-29	1.83E-74	1.11E-02	9.89E-01	1	88.82
cg23619370	17	5168138	-	cg10520397	5138461	*SCIMP*	2302	5.21E-85	3.44E-13	3.56E-74	2.25E-02	9.77E-01	1	43.38
cg23619370	17	5168138	-	cg03146371	5138469	*SCIMP*	2302	1.18E-84	7.82E-13	1.46E-74	8.65E-03	9.91E-01	1	114.56
cg23619370	17	5168138	-	cg21337881	5138645	*C17orf87*	2303	6.13E-94	4.05E-22	1.28E-74	7.47E-03	9.93E-01	1	132.88
cg23619370	17	5168138	-	cg17588003	5138696	*C17orf87*	2304	7.10E-103	4.69E-31	1.19E-74	6.85E-03	9.93E-01	1	144.92
cg23619370	17	5168138	-	cg03167326	5158046		2343	2.04E-200	1.35E-128	8.62E-75	4.70E-03	9.95E-01	1	211.76
cg23619370	17	5168138	-	cg12439163	5165803		2351	1.11E-145	7.32E-74	8.01E-75	4.30E-03	9.96E-01	1	231.74

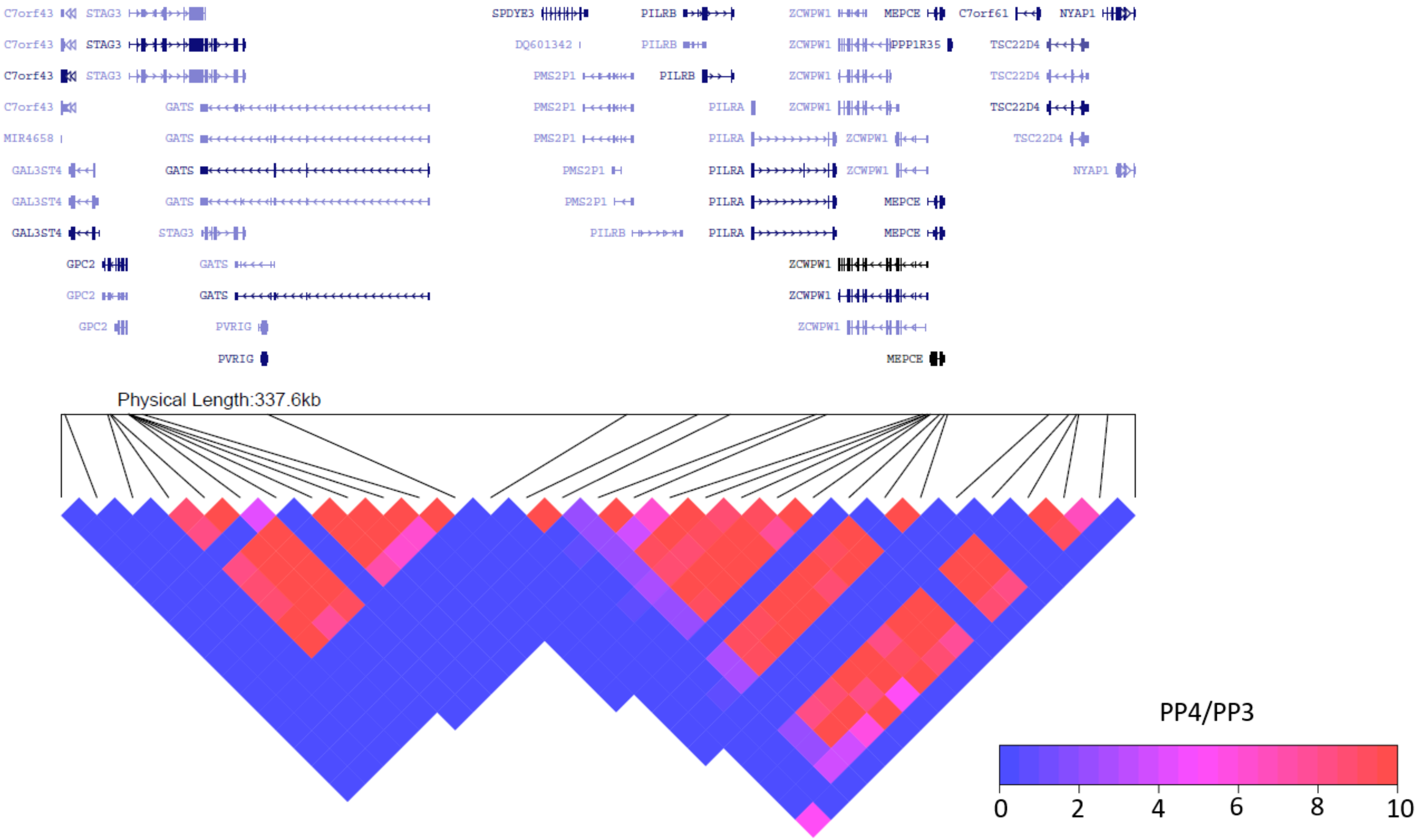


Figure 5.21: Heatmap of Bayesian co-localisation results for all pairs of DNA-methylation sites with at least one significant mQTL (p < 1e-10) in a genomic region on chromosome 7 (chr7: 99754227– 100091793). *Each square represents the relationship between two DNA-methylation sites (ordered by genomic location). The colour of each square indicates the strength of the evidence for a shared genetic signal (from blue [weak] to red [strong]). This strength is calculated as the ratio of the posterior probabilities that they share the same causal variant (PP4) compared to two distinct causal variants (PP3). The ratio was capped to a maximum value of 10.White indicates pairs of DNA-methylation sites that were not tested for co-localisation.*

5.4.8.2.2 The SMR associations are relevant in the context of LOAD

Of the DNAm sites which were prioritised by SMR as having a pleiotropic relationship between a SNP and LOAD several are annotated to genes which have previously been implicated in LOAD. This includes a region (see **Table 5.5**) of eight sites located around the *ZCWPW1*/*NYAP1* GWAS locus. Of note, several of the DNAm sites are annotated to alternative genes residing near this locus (-200kb - + 120kb away) (see **Table 5.4**). One site included in this region is cg03579757, which is annotated to *NYAP1* and was positively associated with LOAD (p= 7.26e-12; see **Table 5.3**). *NYAP1* has been identified as a LOAD-risk factor through GWAS and has previously been nominated as a candidate eQTL gene (Kikuchi *at al.*, 2019). *NYAP1* regulates neuronal morphogenesis and is upregulated in the AD hippocampus (Kikuchi *at al.*, 2019). Several of the other DNAm sites within this region were annotated to *STAG3*, which has also been associated with LOAD. Of note, not all sites shared the same direction of effect, suggesting multiple independent SNPs could be having an effect on LOAD in this region mediated by DNAm at different sites.

I identified a region of four sites around the *ADAM10* locus located on chromosome 15. Within this region cg20288868, cg15770593 and cg20532450 were annotated to *ADAM10* and were negatively associated with LOAD (p=6.90e-08; p=4.74e-08; and p=2.56e-07, respectively; see **Table 5.4**). One additional site in this region - cg09265987 - was not annotated to a gene based on the Illumina manifest however it was also negatively associated with LOAD (p=3.19e-08). The unannotated DNAm site is located ~30kb from *ADAM10*. *ADAM10* has been identified as a LOAD risk factor through GWAS (Kunkle, 2019) and has been identified as the primary α-secretase in the process of amyloid-beta (Aβ) protein precursor cleavage and plays a role in reducing the generation of the Aβ peptides.

A region of two sites located around the *EPHA1* locus located on chromosome 7 was identified, with cg03760621 and cg18997129 being positively associated with LOAD (p=3.00e-09; p=4.82e-09, respectively; see **Table 5.4** and **Figure 5.22**). *EPHA1* is a LOAD risk gene and evidence supports a role for the regulation of this gene via eQTLs in whole blood (Liu *at al.*, 2018)

A region of two sites around the *MSA43* locus located on chromosome 11 was identified, with cg17173423 and cg01440285 being were positively associated with

LOAD (p=1.89e-09; and p=2.66e-08, respectively; see **Table 5.4**). *MS4A3* is part of the MS4A locus. *MS4A3* gene expression varies between tissues but is limited to cells that have functions related to the immune response and haematopoietic cells (Naj *at al.*, 2011).

Several DNAm sites around the HLA region were identified as having an effect on LOAD (see **Table 5.4**). However, there was no evidence of genetic co-regulation in the Bayesian colocalisation analysis suggesting several independent genetic variants may have an effect on LOAD in this region which are mediated by DNAm at different sites. One site identified as having a pleiotropic relationship with load was cg08265274 – located on chromosome 6 and annotated to *HLA-DRB5* – which was negatively associated with LOAD (p= 1.49e-07; see **Table 5.4**). The expression of *HLA-DRB5* in microglia is positively associated with measures of AD pathology (Villegas-Llerena *at al.*, 2016).

Several sites were not identified as regions within the Bayesian colocalisation analysis but had a significant SMR result including cg02400211 – located on chromosome 19, annotated to *CD33* (see **Figure 5.23**), a risk factor for LOAD – and was negatively associated with LOAD (p= 1.00e-07; see **Table 5.4**). *CD33* is a sialic acid-binding immunoglobulin-like lectin that regulates innate immunity. It is expressed in microglia and exhibits increased expression in AD and numbers of *CD33* immunoreactive microglia positively correlate with Aβ plaque burden. *CD33* has also been shown to inhibit microglial uptake of Aβ (Griciuc *at al.*, 2013).

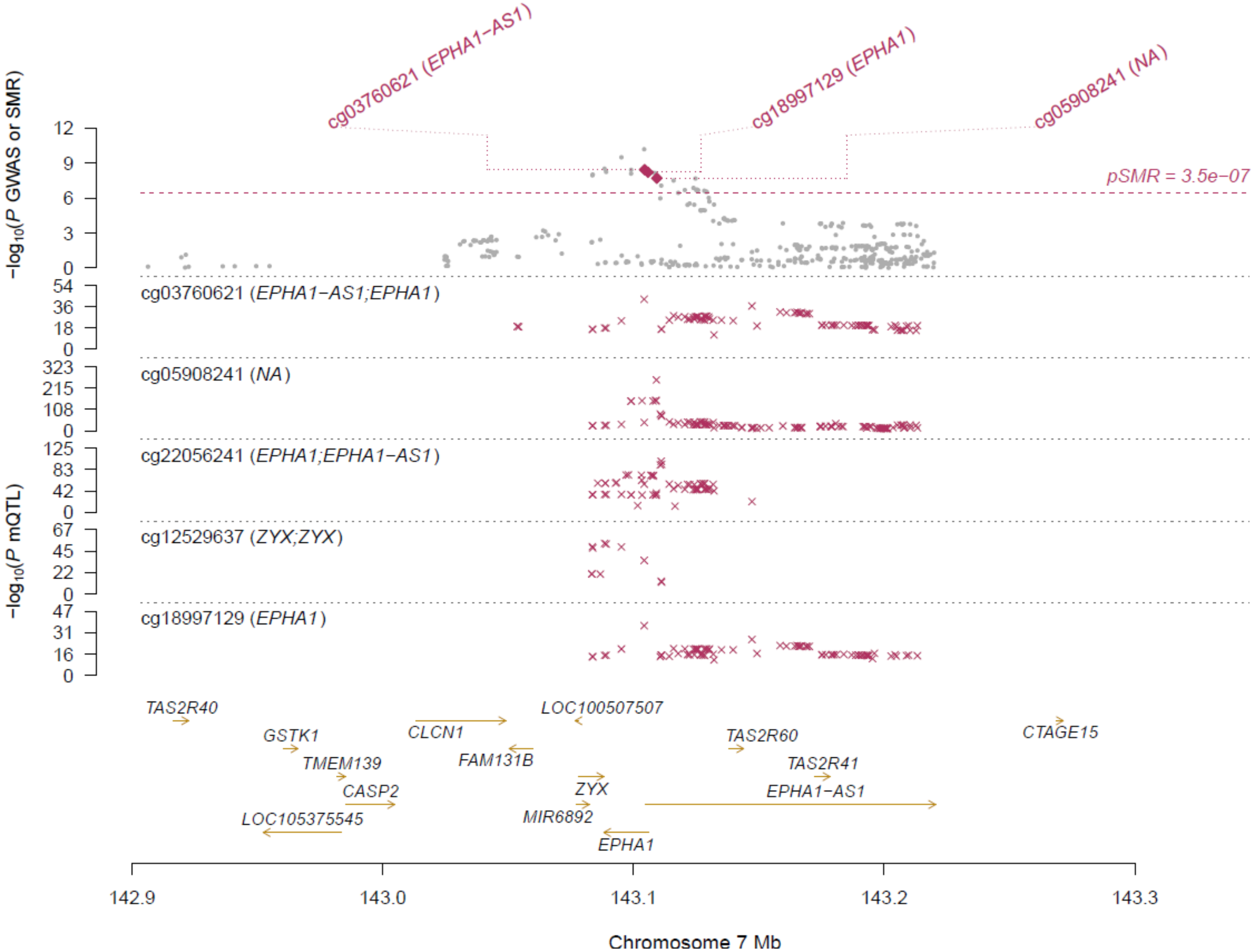


***Figure 5.22 Prioritizing genes at the EPHA1 GWAS locus using SMR analysis with whole blood mQTLs and the Jansen et al LOAD GWAS**. Shown are results at the EPHA1 locus (chromosome 7) for LOAD. Top plot, grey dots represent the P values for SNPs from the Jansen et al LOAD GWAS, diamonds represent the P values for probes from the SMR test – those filled in and highlighted in red passed the SMR and HEIDI tests. Bottom plots, the mQTL P values of SNPs from the whole blood dataset for the probes around the EPHA1 locus. The top and bottom plots include all the SNPs available in the region in the GWAS and mQTL summary data, respectively, not just the SNPs common to both datasets.*

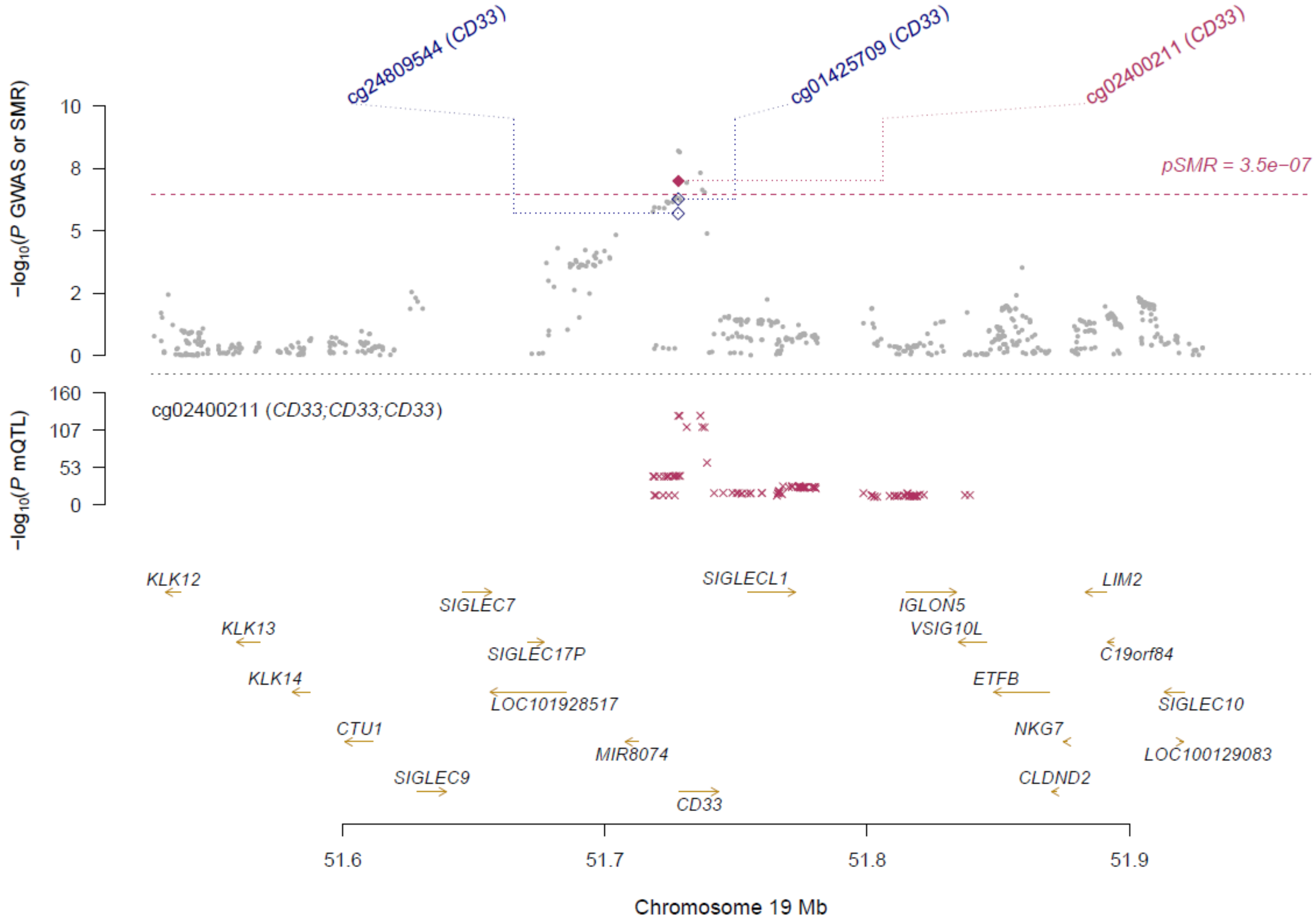


Figure 5.23: Prioritizing genes at the CD33 GWAS locus using SMR analysis with whole blood mQTLs and the Jansen et al LOAD GWAS. *Shown are results at the CD33 locus (chromosome 19) for LOAD. Top plot, grey dots represent the P values for SNPs from the Jansen et al LOAD GWAS, diamonds represent the P values for probes from the SMR test – those filled in and highlighted in red passed the SMR and HEIDI tests. Bottom plots, the mQTL P values of SNPs from the whole blood dataset for the probes around the CD33 locus. The top and bottom plots include all the SNPs available in the region in the GWAS and mQTL summary data, respectively, not just the SNPs common to both datasets.*

5.4.8.2.3 Pathway analysis

In order to relate the DNAm sites prioritised by SMR to biological functions I ran GO pathway analysis using the R package methylGSA. I ran the analysis including probes which reached the threshold p<5e-5 (172 probes, tagging 74 genes). I identified 148 significant pathways (see **Figure 5.24**) including pathways related to cellular processes in the cortex such as oligodendrocyte development, microglial and glial cell activation, gliogenesis and myelination. Neuronal and glial cell interactions are vital for synaptic homeostasis and research suggest these processes are disrupted in AD cases (Nordengen *at al.*, 2019).

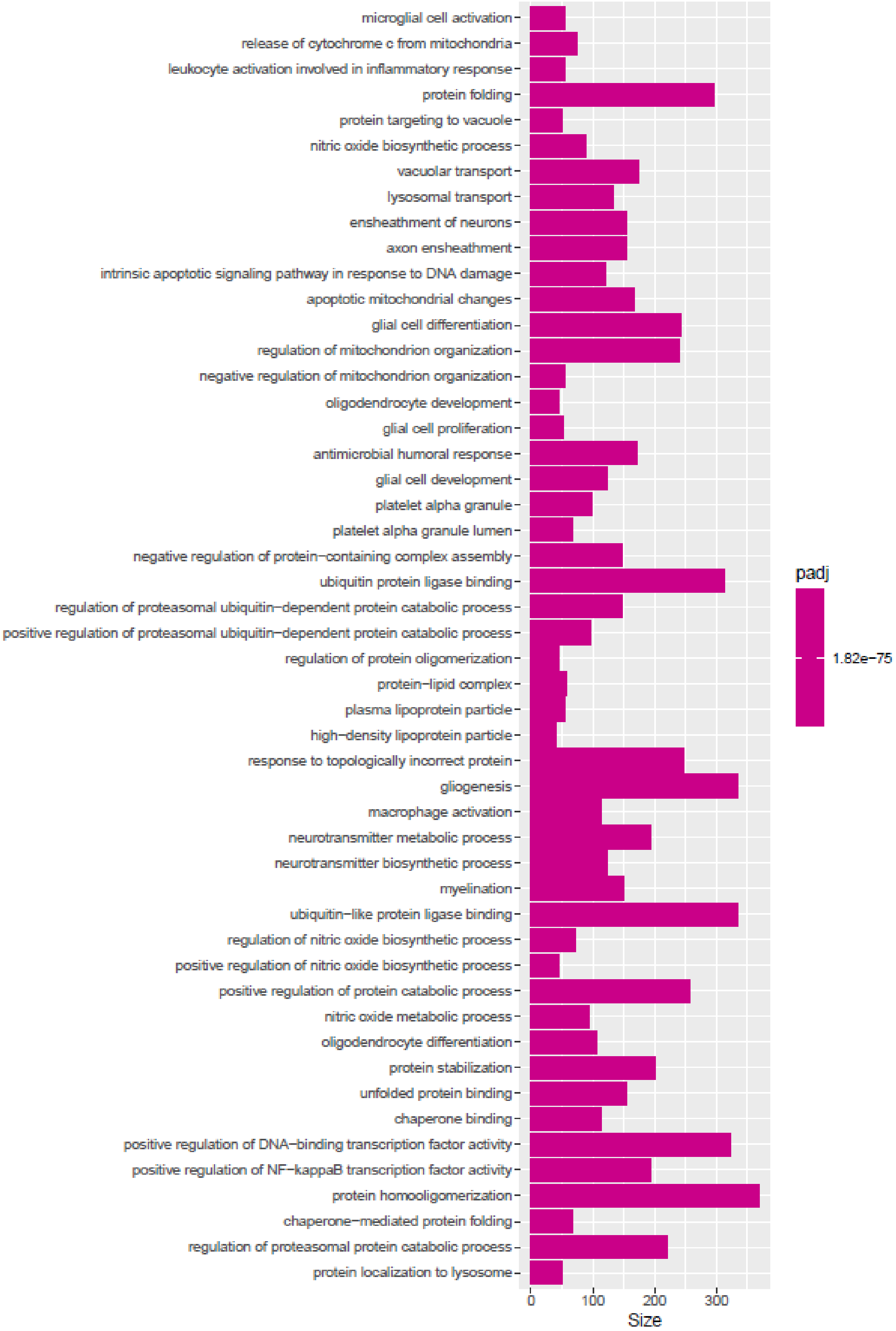


Figure 5.24: Pathway analysis of the whole Blood mQTL SMR results which incorporated the Jansen et al (2019) GWAS. *Shown are the top 50 significant result from the pathway analysis. The y-axis represents the gene ontology biological processes. The x axis is size, which represents the number of genes included in the pathway.*

5.4.9 Identifying putative pleiotropic relationships between DNA methylation and Alzheimer’s disease in the prefrontal cortex using SMR

I used SMR to test the 42,926 DNAm sites identified in my cortex PFC mQTL analysis (see above; section **5.4.3**) with the two latest AD GWAS with publicly available summary statistics.

5.4.9.1 SMR results using the Kunkle et al GWAS

Using the Kunkle *at al.* GWAS I applied SMR and identified 19 associations which passed the Bonferroni significant threshold ($p < 1.15e\text{-}06$) and 10 which passed both the significance threshold and the HEIDI threshold ($p > 0.01$; see **Table 5.6** and **Figure 5.25**). Of these, five (50%) were positively associated with LOAD. Many of the associations were identified in regions proximal to GWAS peaks (see **Figure 5.15(B)**), including several along chromosome 11.

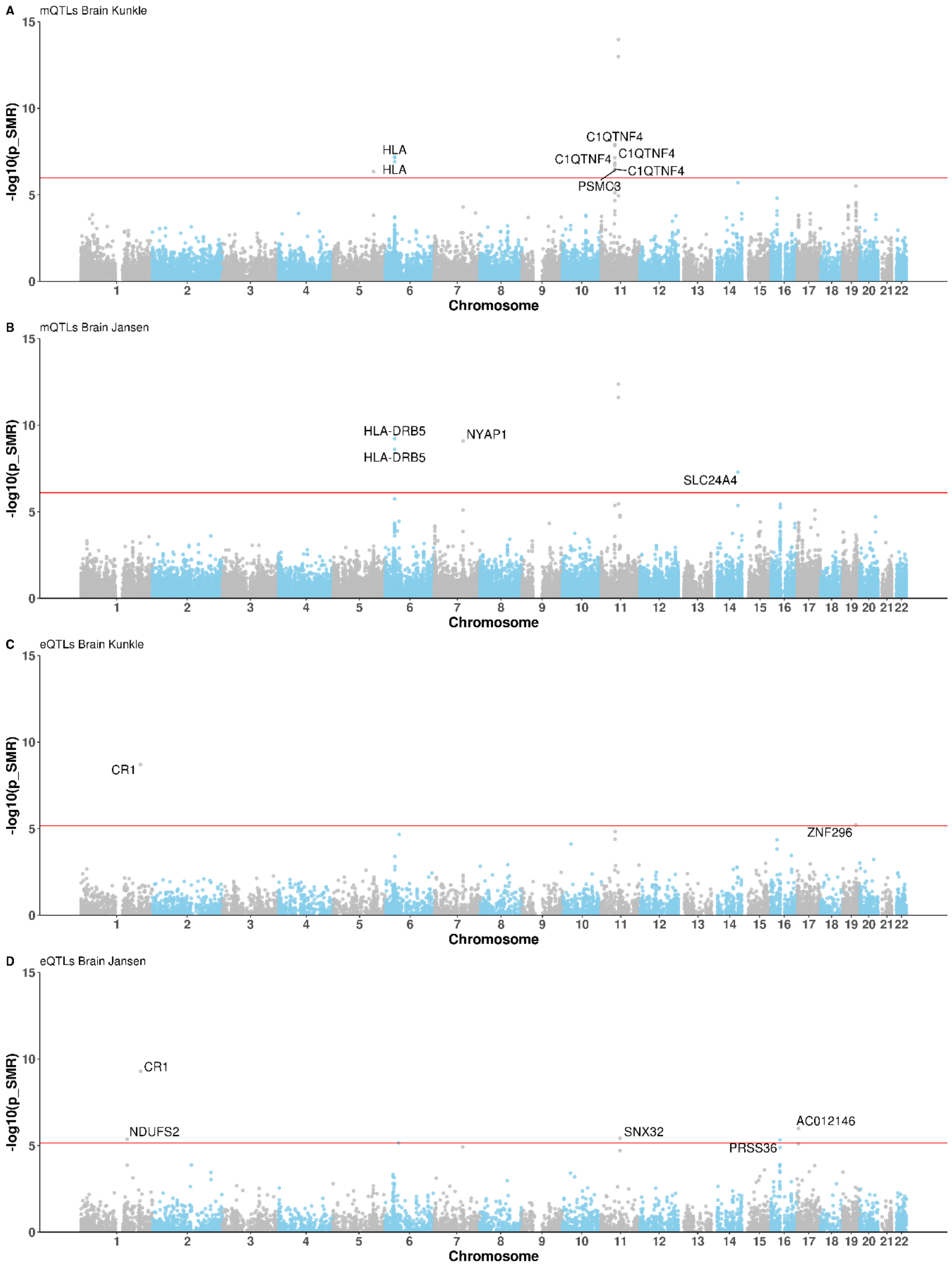


Figure 5.25: Manhattan plots of Summary data-based Mendelian Randomisation (SMR) tests for pleiotropic effects between LOAD and regulatory markers in cortex. *(**A**) using DNA methylation quantitative trait loci (mQTL) and the Kunkle et al GWAS summary statistics; (**B**) using mQTLs and the Jansen et al GWAS summary statistics; (**C**) using expression quantitated trait loci (eQTLs) and the Kunkle et al GWAS summary statistics; and (**D**) using eQTLs and the Jansen et al GWAS summary statistics. The x axis is the genomic position, segregated by chromosome. Shown on the y-axis of each plot is the –log10 p-value from the SMR analysis using mQTLs (**A**) and (**B**) or eQTLs (**C**) and (**D**). Each point represents an SMR test for a particular mQTL/ eQTL site which passed the Bonferroni and HEIDI significant thresholds. The red horizontal line represents the genome-wide multiple testing significance threshold.*

Table 5.6: Summary data-based Mendelian Randomisation (SMR) test results for pleiotropic effects between LOAD and DNA methylation using the Kunkle at al. GWAS and the cortex mQTL dataset. (N = 10 hits where p < 1.15e-06 and HEIDI > 0.01). *CHR: chromosome; BP: base position (hg19 P); Instrument: top cis-mQTL for DNAm site with smallest p-value for association; Beta SMR: estimate for the effect of DNAm on LOAD; P SMR: p-value for SMR association; HEIDI: p-value for the HEIDI test; N SNP HEIDI; number of SNPs included in the HEIDI test.*

DNAm site				**GWAS**		**SMR**			**HEIDI**	
Probe ID	**CHR**	**BP**	**Gene**	**Instrument (CHR:BP)**	**Ref MAF**	**Beta**	**SE**	**P**	**P**	**N SNP**
cg13076785	6	32520916	*HLA-DRB6*	6:32561656	0.43	-0.61	0.11	6.91E-08	4.36E-01	20
cg15710545	6	32578114	-	6:32576894	0.42	1.08	0.20	6.43E-08	2.12E-01	17
cg20307385	11	47447363	*PSMC3*	11:47462140	0.40	5.86	1.15	3.89E-07	4.00E-02	20
cg09507712	11	47616693	*C1QTNF4*	11:47606483	0.37	-1.73	0.30	1.24E-08	4.50E-02	20
cg07409245	11	47616751	*C1QTNF4*	11:47687147	0.37	-1.51	0.29	1.97E-07	1.60E-02	20
cg15575356	11	47616757	*C1QTNF4*	11:47687147	0.37	-1.86	0.36	3.30E-07	1.90E-02	20
cg27051260	11	47616825	*C1QTNF4*	11:47606483	0.37	-2.02	0.36	1.44E-08	1.20E-02	20
cg17688768	11	47628441	-	11:47550835	0.37	2.94	0.56	1.46E-07	1.60E-02	20
cg24977308	11	47636548	-	11:47486885	0.37	2.33	0.43	7.19E-08	8.40E-02	20
cg02521229	11	60019236	-	11:60020112	0.41	0.81	0.11	1.06E-14	5.10E-02	20

5.4.9.1.1 Identifying regions in the Kunkle cortex SMR analysis

I used the Bayesian colocalisation results to identify if each pair of DNAm sites in the same region (defined as probes within 250kb of each other) had evidence for colocalisation. In total I looked at 16 pairs of sites, and all demonstrated evidence for a co-localized association within the same genomic region (PP3 + PP4 > 0.99). 10 pairs of DNAm sites had 'suggestive' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 1) (see **Table 5.7**), all of which also met the criteria for 'convincing' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 5) (see **Table 5.7**). The DNAm sites in the region are all located on chromosome 11 around the *SPI1* LOAD GWAS locus, although they are annotated either to *C1QTNF4* (which is located ~200kb from *SPI1*) or do not have a gene annotation. These results suggest that these DNAm sites are genetically co-regulated and it is likely they have a shared causal variant associated with DNA methylation level. Therefore, I considered this as a region to aid the interpretation of the SMR results.

Table 5.7: Bayesian Colocalisation results for the Kunkle et al cortex SMR significant DNA methylation probes. *There is evidence for colocalisation for several of the DNA methylation probes which had a significant SMR result. The posterior probability for all pairs of sites tested provided evidence for a co-localized association within the same genomic region (PP3 + PP4 > 0.99). 10 pairs of sites had 'suggestive' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 1). 10 pairs of sites had 'convincing' evidence to support associations between both DNAm sites with the same causal mQTL variant (PP3 + PP4 > 0.99 & PP4/PP3 > 5). The strength of the evidence is based on co-localisation criteria developed by Guo and colleagues (Guo at al., 2015). PP.H_i.abf = Posterior probability for each of the Bayesian Colocalisation hypotheses (i).*

DNA methylation sites							**Bayesian Colocalisation results**							
Site 1	**CHR**	**BP**	**Annotated Gene**	**Site 2**	**BP**	**Annotated Gene**	**N SNPs**	**PP.H0.abf**	**PP.H1.abf**	**PP.H2.abf**	**PP.H3.abf**	**PP.H4.abf**	**P3 + P4**	**P4/P3**
cg07409245	11	47616751	*C1QTNF4*	cg09507712	47616693	*C1QTNF4*	1364	1.69e-77	4.89e-43	2.52e-36	7.23e-02	9.28e-01	1	12.84
cg15575356	11	47616757	*C1QTNF4*	cg09507712	47616693	*C1QTNF4*	1364	4.23e-69	4.18e-43	6.33e-28	6.15e-02	9.38e-01	1	15.25
cg15575356	11	47616757	*C1QTNF4*	cg07409245	47616751	*C1QTNF4*	1364	9.73e-63	9.60e-37	2.82e-28	2.69e-02	9.73e-01	1	36.16
cg17688768	11	47628441	-	cg09507712	47616693	*C1QTNF4*	1350	9.15e-71	3.43e-43	1.37e-29	5.03e-02	9.50e-01	1	18.86
cg17688768	11	47628441	-	cg07409245	47616751	*C1QTNF4*	1350	5.72e-64	2.14e-36	1.66e-29	6.13e-02	9.39e-01	1	15.32
cg17688768	11	47628441	-	cg15575356	47616757	*C1QTNF4*	1350	1.48e-55	5.54e-28	1.46e-29	5.37e-02	9.46e-01	1	17.61
cg24977308	11	47636548	-	cg09507712	47616693	*C1QTNF4*	1341	1.81e-77	3.48e-43	2.70e-36	5.11e-02	9.49e-01	1	18.56
cg24977308	11	47636548	-	cg07409245	47616751	*C1QTNF4*	1341	7.59e-71	1.46e-36	2.21e-36	4.15e-02	9.59e-01	1	23.10
cg24977308	11	47636548	-	cg15575356	47616757	*C1QTNF4*	1341	2.05e-62	3.94e-28	2.02e-36	3.79e-02	9.62e-01	1	25.36
cg24977308	11	47636548	-	cg17688768	47628441	-	1358	4.39e-64	8.44e-30	1.64e-36	3.07e-02	9.69e-01	1	31.60

5.4.9.1.2 The SMR associations are relevant in the context of LOAD

Of the DNAm sites which were prioritised by SMR as having a pleiotropic relationship between a SNP and LOAD, six are annotated to genes (three unique genes) which have all previously been implicated in LOAD. This includes a region (region identified using Bayesian colocalisation; see **Table 5.7**) of six sites located around the *SPI1* LOAD GWAS locus located on chromosome 11, although several of the DNAm sites are annotated to alternative genes residing near this locus (see **Table 5.4**). Four of the sites in this region - cg09507712, cg07409245, cg15575356 and cg27051260 – located on chromosome 11 and annotated to the gene *C1QTNF4* – were negatively associated with LOAD (p=1.24e-08; p=1.97e-07; p=3.30e-07; p= 1.44e-08, respectively; see **Table 5.6**). *C1QTNF4* is located in the same LD block as *SPI1*. However, AD genetic studies have found 8 independent variants in LD within the *SPI1* region which are also eQTLs for *C1QTNF4* (Rosenthal & Kamboh, 2014). *SPI1* may be acting in combination with or serving as a proxy for other genes around this region that mediate AD risk. This is further supported by the fact two of the other sites in this region - cg17688768 and cg24977308 – are positively associated with LOAD (p=1.46e-07 and p=7.19e-08), suggesting that there may be more than one genetic variant acting within the region which are mediated by DNAm at different sites.

Several sites were not identified as regions within the Bayesian colocalisation analysis but had a significant SMR result including cg13076785 – located on chromosome 6, annotated to gene *HLA-DRB6* (see **Figure 5.26**), a LOAD risk gene involved in immune response – which was negatively associated with LOAD (p = 6.91e-08; see **Table 5.6**). It resides in in the LD block of *HLA-DRB1*. The expression of *HLA-DRB1* in microglia is positively associated with measures of AD pathology (Villegas-Llerena *at al.*, 2016). Additionally, cg20307385, located on chromosome 11, annotated to gene *PSMC3*, was positively associated with LOAD (p = 3.89e-07; see **Table 5.6**). *PSMC3* is located in the LD block of *SPI1*, a LOAD risk gene, however there was no evidence that this site genetically co-regulated with the sites above, providing further evidence that there may be several SNPs which have an effect on LOAD independently and are mediated by DNAm are different sites.

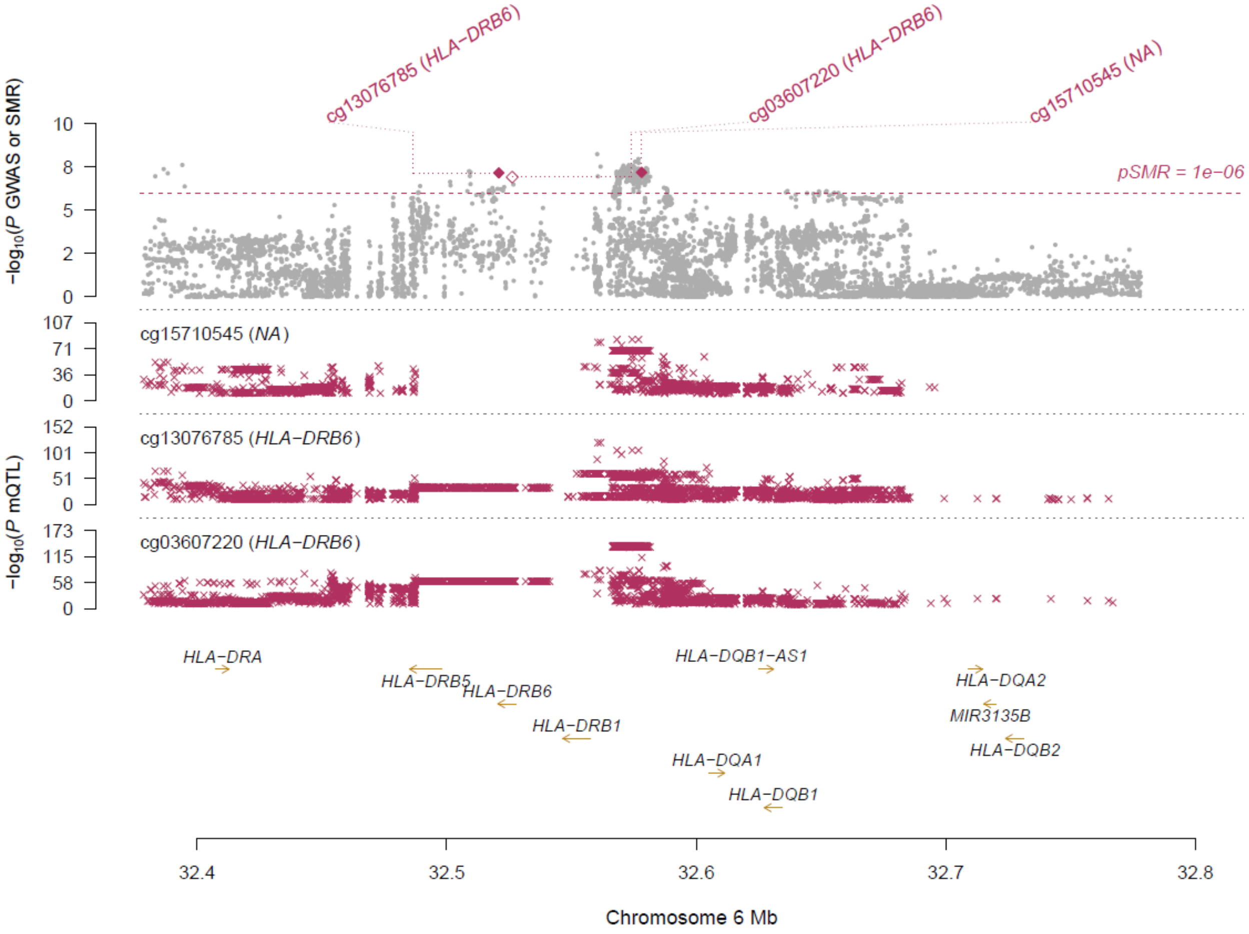


Figure 5.26: Prioritizing genes at the HLA GWAS locus using SMR analysis with brain prefrontal cortex mQTLs and the Kunkle et al LOAD GWAS*. Shown are results at the HLA locus (chromosome 6) for LOAD. Top plot, grey dots represent the P values for SNPs from the Jansen et al LOAD GWAS, diamonds represent the P values for probes from the SMR test – those filled in and highlighted in red passed the SMR and HEIDI tests. Bottom plots, the mQTL P values of SNPs from the whole blood dataset for the probes around the HLA locus. The top and bottom plots include all the SNPs available in the region in the GWAS and mQTL summary data, respectively, not just the SNPs common to both datasets.*

5.4.9.1.3 Pathway analysis

In order to relate the DNAm sites prioritised by SMR to biological functions I ran GO pathway analysis using the R package methylGSA. I ran the analysis including probes which reached the threshold p<5e-5 (22 probes, tagging 11 genes). I identified 18 significant pathways (see **Figure 5.27**) including pathways related to Aβ processing including metabolic processes and Aβ formation.

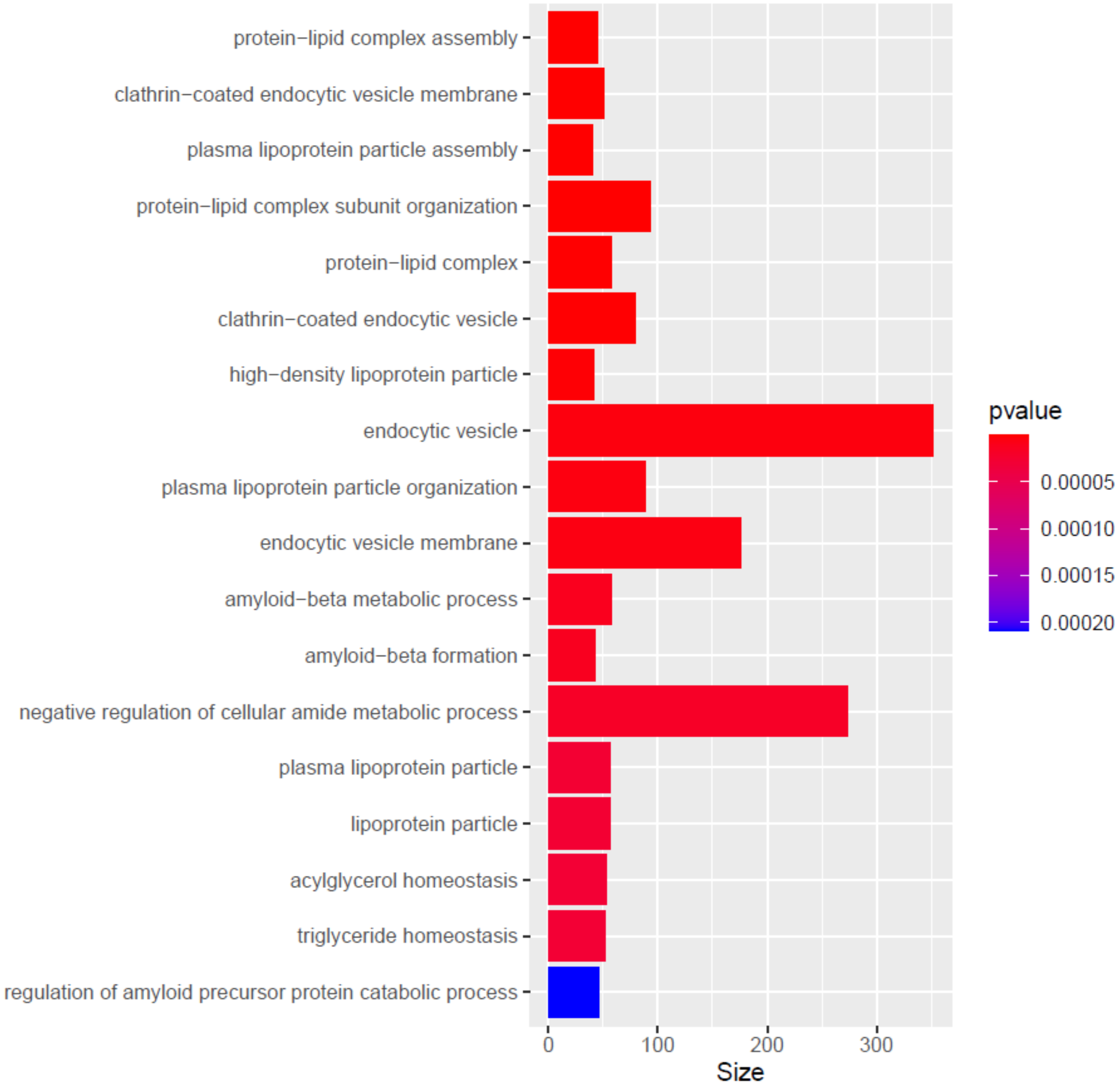


Figure 5.27: Pathway analysis of the cortex mQTL SMR results which incorporated the Kunkle et al (2019) GWAS. *Shown are the significant result from the pathway analysis (n=18). The y-axis represents the gene ontology biological processes. The x axis is size, which represents the number of genes included in the pathway.*

5.4.9.2 SMR results using the Jansen et al GWAS

Using the Jansen *at al.* GWAS I applied SMR and identified 22 associations which passed the Bonferroni significant threshold (p < 1.15e-06) and four which passed both the significance threshold and the HEIDI threshold (p > 0.01; see **Table 5.8** and **Figure 5.25**). Of these four, three (75%) were positively associated with LOAD. Of the DNAm sites which were prioritised by SMR as having a pleiotropic relationship between a SNP and LOAD, all are annotated to genes which have previously been implicated in LOAD:

- cg08265274, located on chromosome 6, annotated to gene *HLA-DRB5* was negatively associated with LOAD (p=5.97e-10; see **Table 5.8**). *HLA-DRB5* resides in the LD block of *HLA-DRB1*. The expression of *HLA-DRB1* in microglia is positively associated with measures of AD pathology (Villegas-Llerena *at al.*, 2016).
- cg03579757, located on chromosome 7 annotated to gene *NYAP1* was positively associated with LOAD (p=8.19e-10; see **Table 5.8**). *NYAP1* has been identified as a LOAD-risk factor through GWAS (Kunkle, 2019). *NYAP1* regulates neuronal morphogenesis and is upregulated in the AD hippocampus (Kikuchi *at al.*, 2019).
- cg23484461, located on chromosome 14, annotated to gene *SLC24A4* was positively associated with LOAD (5.13e-08; see **Table 5.8**). *SLC24A4* has previously been identified as an AD GWAS gene and evidence suggests this gene plays a role in neural development (Yu *at al.*, 2015). In addition, differential methylation in the PFC of a DNAm site located within the *SLC24A4* locus has been associated with LOAD pathology (Yu *at al.*, 2015).

These sites were all located on different chromosomes and therefore there was no evidence that they are genetically co-regulated with each other.

Table 5.8: Summary data-based Mendelian Randomisation (SMR) test results for pleiotropic effects between LOAD and DNA methylation using the Jansen at al. GWAS and the cortex mQTL dataset. (N = 4 hits where p < and HEIDI > 0.01). *CHR: chromosome; BP: base position (hg19 P Instrument: top cis-mQTL for DNAm site with smallest p-value for association; Beta SMR: estimate for the effect of DNAm on LOAD; P SMR: p-value for SMR association; HEIDI: p-value for the HEIDI test; N SNP HEIDI; number of SNPs included in the HEIDI test.*

DNAm site				**GWAS**		**SMR**			**HEIDI**	
Probe ID	**CHR**	**BP**	**Annotated Gene**	**Instrument (CHR:BP)**	**Ref MAF**	**Beta**	**SE**	**P**	**P**	**N SNP**
cg08265274	6	32490444	*HLA-DRB5*	6:32573415	0.19	-0.11	0.02	2.51E-09	2.30E-02	20
cg03579757	7	100091793	*NYAP1*	7:100012579	0.19	-0.10	0.02	5.97E-10	1.50E-02	20
cg02521229	11	60019236	-	11:60020112	0.28	0.39	0.06	8.19E-10	4.27E-01	20
cg23484461	14	92936957	*SLC24A4*	14:92937293	0.41	-0.16	0.02	2.47E-12	1.21E-01	20

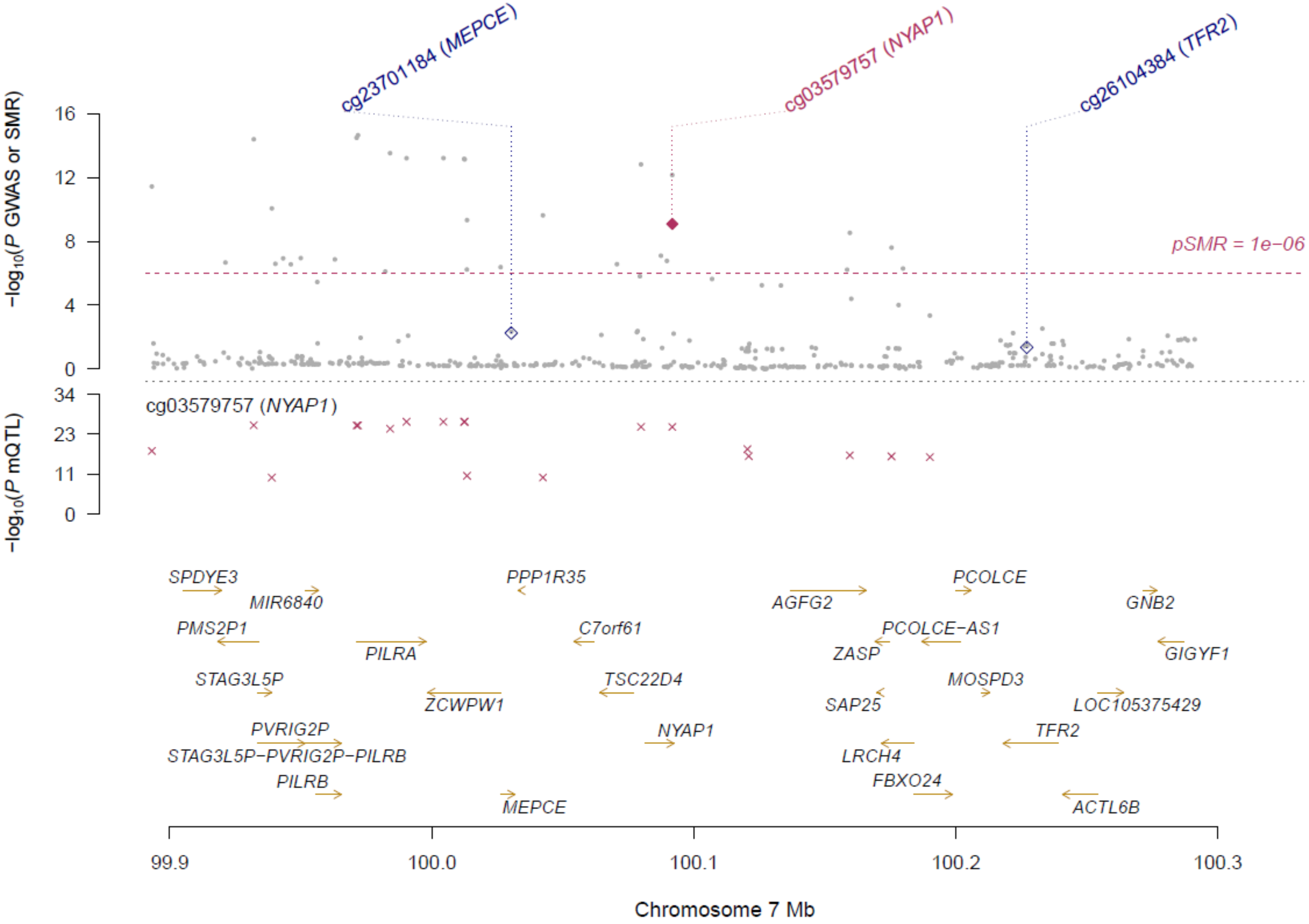


Figure 5.28 Prioritizing genes at the NYAP1 GWAS locus using SMR analysis with brain prefrontal cortex mQTLs and the Jansen et al LOAD GWAS*. Shown are results at the NYAP1 locus (chromosome 7) for LOAD. Top plot, grey dots represent the P values for SNPs from the Jansen et al LOAD GWAS, diamonds represent the P values for probes from the SMR test – those filled in and highlighted in red passed the SMR and HEIDI tests. Bottom plots, the mQTL P values of SNPs from the whole blood dataset for the probes around the NYAP1 locus. The top and bottom plots include all the SNPs available in the region in the GWAS and mQTL summary data, respectively, not just the SNPs common to both datasets.*

5.4.9.2.1 Pathway analysis

In order to relate the DNAm sites prioritised by SMR to biological functions I ran GO pathway analysis using the R package methylGSA including genes annotated to DNAm sites reaching the threshold p<5e-5 (24 probes, tagging 17 genes). I identified 40 significant pathways (see **Figure 5.29**) including pathways relating to ion processing such as calcium transport. The calcium hypothesis of AD states that dysregulation to mechanisms which regulate calcium homeostasis may play a role in neuronal dysfunction in AD (Alzheimer's Association Calcium Hypothesis Workgroup, 2017). Interestingly, the pathways identified from the SMR analysis incorporating cortex mQTLS are more related towards neurotransmitters, whereas the whole blood analysis was had more pathways related to lipid processing which suggests there are some tissue specific differences.

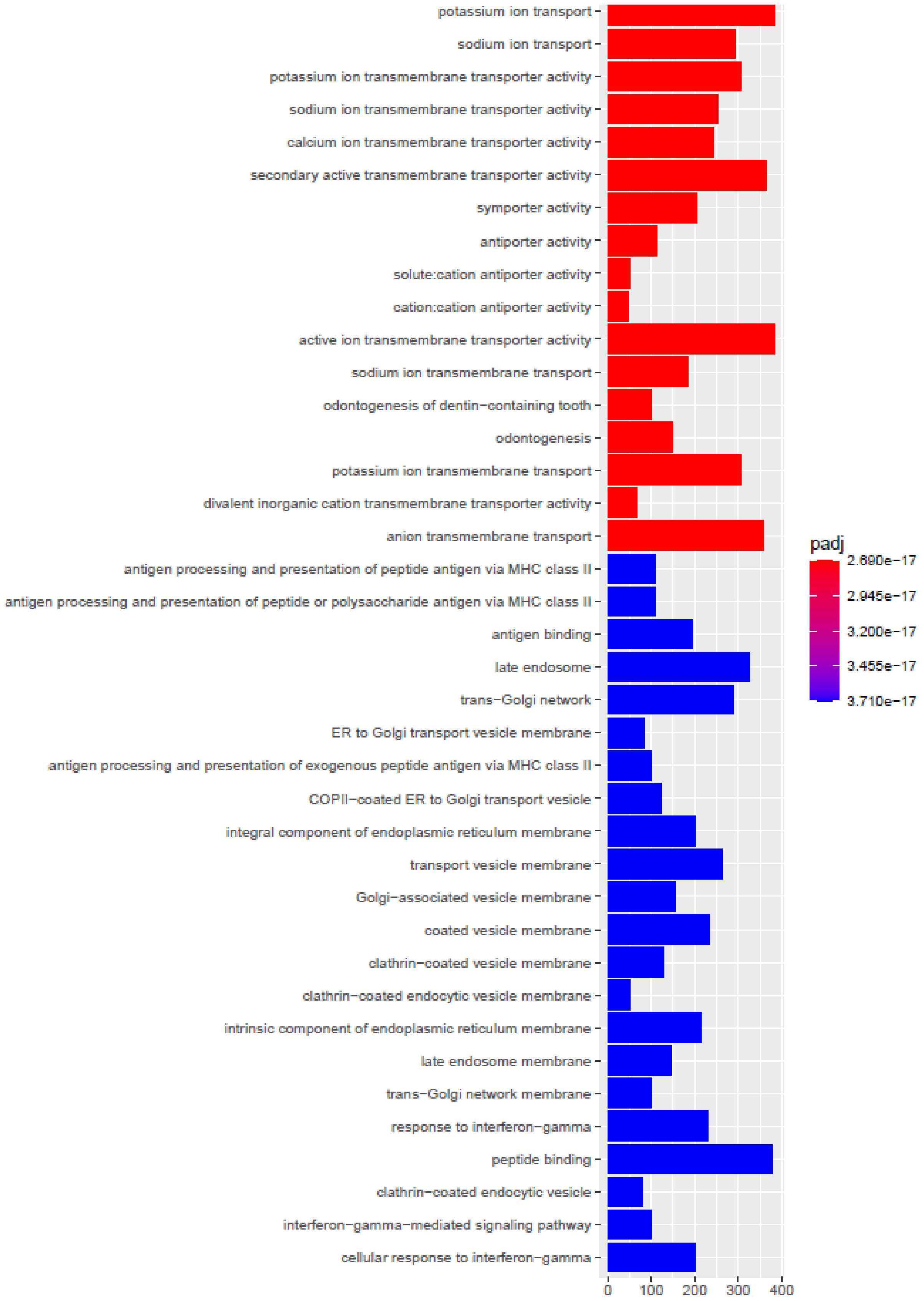


Figure 5.29: Pathway analysis of the cortex mQTL SMR results which incorporated the Jansen et al (2019) GWAS. *Shown are the significant result from the pathway analysis (n=40). The y-axis represents the gene ontology biological processes. The x axis is size, which represents the number of genes included in the pathway.*

5.4.10 Comparing the whole blood and cortex mQTL SMR results

I compared the whole blood and cortex mQTL SMR results. First, I looked for consistency in the direction of effect of the SMR beta values, including SMR results which reached nominal significance (p SMR<0.05), since more stringent thresholds resulted in too few sites for comparison across tissues. The effects were generally concordant across both tissues and both LOAD GWAS (see **Figure 5.30**), which was confirmed by highly significant (Bonferroni p < 0.05/24 = 0.002) sign test p-values for all comparisons (see **Figure 5.30**). These results suggest the mQTLs included in the analyses are generally consistent across tissues. However, the direction was not consistent for all DNAm sites. This may represent heterogeneity between tissues and the summary statistics. Of note, since a more relaxed p-value threshold was used we cannot make definitive conclusions regarding this as it is likely there are some false positive results included in this analysis.

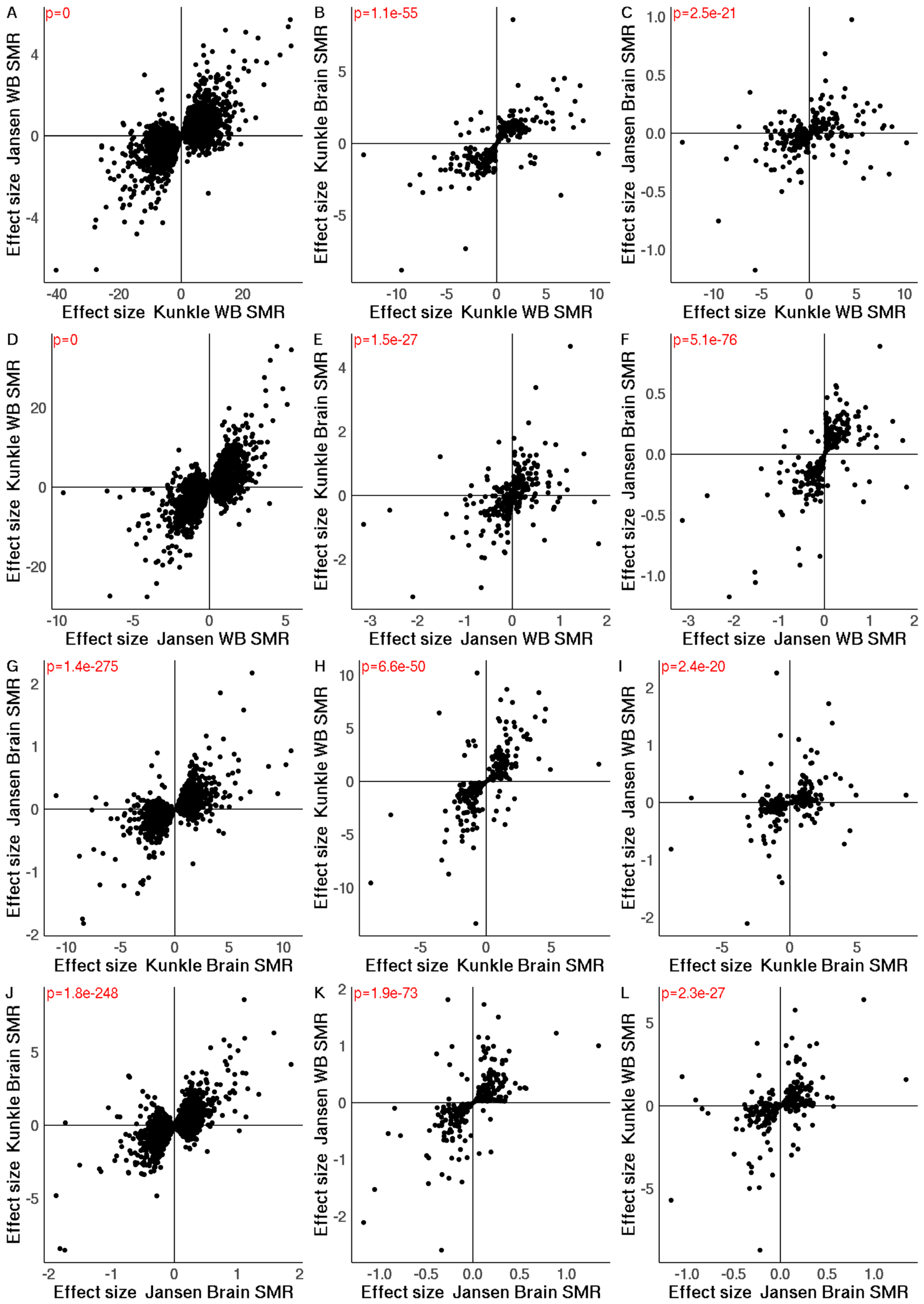


Figure 5.30: The direction of effect is generally consistent across all mQTL SMR analyses when considering nominally significant probes (p SMR < 0.05). *p= binomial sign test p-value.*

I explored if there was any overlap across the Bonferroni significant whole blood and cortex mQTL SMR results. I included results generated both with the Kunkle *at al.* (2019) and Jansen *at al.* (2019) GWAS summary statistics so I could evaluate the differences between the summary statistics as well as between tissues. There were multiple overlapping sites between the SMR results (see **Figure 5.31**) suggesting there are correlations across tissues.

When comparing the Kunkle and Jansen blood mQTL results there was an overlap of 11 sites (see **Table 5.9**), four of which were on chromosome 11 and three on chromosome 8 annotated to *CLU*. When comparing the Kunkle and Jansen cortex mQTL results there was an overlap of one site - cg02521229 (chromosome 11) – this site has no proximal gene annotate but is located 60kb from the MS4A region. This site also overlapped with the Kunkle blood mQTL results.

Of note, many of the DNAm SMR associations identified in the blood mQTL SMR analysis were not tested in the cortex mQTL SMR analysis; fewer mQTLs were identified in the cortex cohort due to the smaller sample size. A Venn diagram showing the overlapping probes when limited to probes tested across all mQTL SMR analyses is shown in **Figure 5.32**. Six SMR significant probes overlapped across blood and cortex results with. In addition to cg02521229, when looking at comparisons between the Kunkle *at al.* blood and cortex SMR there was an overlap of an additional probe (see **Table 5.10**), located on chromosome 11 near the *SPI1* gene. When looking at comparisons between the Jansen *at al.* blood and cortex SMR results there were three overlapping probes (see **Table 5.11**) which were located on chromosomes 6, 7 and 14 and annotated to the genes *HLA-DRB5*, *NYAP1* and *SLC24A4*, respectively. One of these three DNAm probes - cg23484461, annotated to *SLC24A4* – was also found in the Kunkle *at al.* blood analysis. cg15710545 (chromosome 6) overlapped between the Kunkle cortex and the Jansen blood SMR analyses.

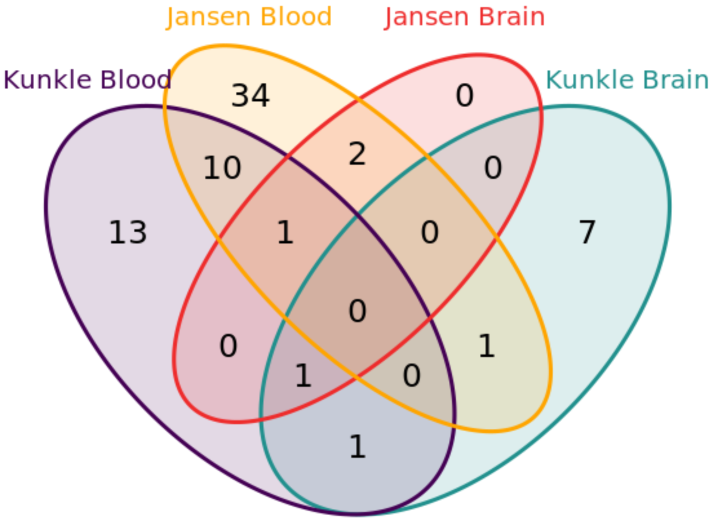


Figure 5.31: Overlap of the Summary data-based Mendelian Randomisation (SMR) test results for pleiotropic effects between LOAD and DNA methylation in whole blood and cortex mQTL datasets using both the Kunkle et al GWAS (2019) and the Jansen et al GWAS (2019).

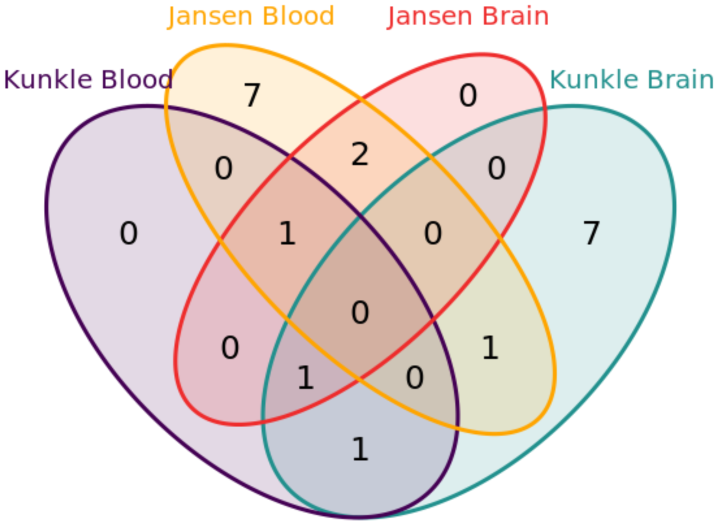


Figure 5.32: Overlap of the Summary data-based Mendelian Randomisation (SMR) test results for pleiotropic effects between LOAD and DNA methylation in whole blood and brain mQTL datasets using both the Kunkle et al GWAS (2019) and the Jansen et al GWAS (2019), limited to probes tested across all analyses.

Table 5.9: Overlapping DNA methylation sites from the results between 2 whole blood mQTL SMR analysis: (1) using the Kunkle et al GWAS and (2) the Jansen et al GWAS. *DNAm: DNA methylation; CHR: chromosome; BP: position (hg19).*

DNAm Probe ID	CHR	BP	Annotated Gene
cg20172563	6	47487173	*CD2AP*
cg16292768	8	27467783	*CLU*
cg22217144	8	27468166	*CLU*
cg21111824	8	27468171	*CLU*
cg08871934	10	11720283	-
cg18684128	11	60033393	-
cg18959616	11	85814918	-
cg24251493	11	85866690	-
cg04441687	11	85869322	-
cg23484461	14	92936957	*SLC24A4*
cg15233575	19	45221584	-

Table 5.10: Overlapping DNA methylation sites from the results between the blood and cortex mQTL SMR analysis using the Kunkle et al GWAS. *DNAm: DNA methylation; CHR: chromosome; BP: position (hg19).*

DNAm Probe ID	CHR	BP	Annotated Gene
cg02521229	11	60019236	-
cg17688768	11	47628441	-

Table 5.11: Overlapping DNA methylation sites from the results between the blood and cortex mQTL SMR analysis using the Jansen et al GWAS. *DNAm: DNA methylation; CHR: chromosome; BP: position (hg19).*

DNAm Probe ID	CHR	BP	Annotated Gene
cg08265274	6	32490444	*HLA-DRB5*
cg03579757	7	100091793	*NYAP1*
cg23484461	14	92936957	*SLC24A4*

Eight DNAm probes were uniquely identified in the Kunkle cortex mQTL SMR analysis (see **Table 5.12**). One of these sites was annotated to *HLA-DRB6*. The remaining seven sites were all located in the same region (~200kb long) around chromosome 11 and annotated to the genes *PSMC3* and *C1QTNF4*, genes which are located adjacent to the *SPI1* GWAS locus. Expression of several genes around this locus are highly correlated with one another and have been associated with AD status (Karch *at al.*, 2016). *SPI1* may be acting in combination with or serving as a proxy for other genes around this region that mediate AD risk. These results indicate there may be tissue specific regulation of some of these genes.

Table 5.12: Unique DNA methylation sites identified in the cortex mQTL SMR analysis using the Kunkle et al (2019) GWAS. *These sites were uniquely identified in this analysis, but they were tested in all four mQTL SMR analyses.* ***.*** *DNAm: DNA methylation; CHR: chromosome; BP: position (hg19).*

DNAm Probe ID	CHR	BP	Gene
cg13076785	6	32520916	*HLA-DRB6*
cg20307385	11	47447363	*PSMC3*
cg09507712	11	47616693	*C1QTNF4*
cg07409245	11	47616751	*C1QTNF4*
cg15575356	11	47616757	*C1QTNF4*
cg27051260	11	47616825	*C1QTNF4*
cg17688768	11	47628441	-
cg24977308	11	47636548	-

Seven DNAm probes were uniquely identified in the Jansen blood mQTL SMR analysis (see **Table 5.13**). Three were located around the same region on chromosome 6 (within 509 bases) and are located near the HLA region. Three Sites were identified around the same region (within 10kb) on chromosome 7. These sites are annotated to *GPC2* and *STAG3*. A recent study investigating the relationship between cardiovascular risk and AD identified a novel association with AD (by conditioning on cardiovascular risk factors) on chromosome 7 tagging GATS, STAG4 or PIVRIG which was not in LD with the other GWAS SNP in this region (Broce *at al.*, 2019). Our results support a role for these genes in the pathogenesis of AD in whole blood.

Table 5.13: Unique DNA methylation sites identified in the whole blood mQTL SMR analysis using the Jansen et al (2019) GWAS. *These sites were uniquely identified in this analysis but they were tested in all four mQTL SMR analyses.* ***.*** *DNAm: DNA methylation; CHR: chromosome; BP: position (hg19).*

DNAm Probe ID	CHR	BP	Gene
cg12672189	6	32427868	-
cg20946741	6	32428328	-
cg10556520	6	32428377	-
cg18090197	7	99769602	*GPC2*
cg00553149	7	99775558	*STAG3*
cg10407106	7	99779719	*STAG3*
cg09265987	15	59050614	-

No DNAm probes were uniquely identified in the Jansen cortex mQTL SMR analysis.

5.4.11 Identifying putative pleiotropic relationships between gene expression and Alzheimer’s disease in whole blood using SMR

I used SMR with a publicly availably whole blood eQTL data generated by (Westra *at al.*, 2015), which included 4,179 genes with the two latest AD GWAS with publicly available summary statistics.

5.4.11.1 SMR results using the Kunkle et al GWAS

Using the Kunkle *at al.* GWAS I applied SMR and identified eight associations which passed the Bonferroni significant threshold (p < 8.38 e-06) and three which passed both the significance threshold and the HEIDI threshold (p > 0.01; see **Table 5.15** and **Figure 5.13**). All three sites (100%) were positively associated with LOAD (i.e. increased gene expression was associated with LOAD). Of the genes which were prioritised by SMR as having a pleiotropic relationship between a SNP and LOAD all three are annotated to genes which have previously been implicated in LOAD:

- ILMN_2330966 – located on chromosome 8, annotated to the gene *PTK2B* (see **Figure 5.33**), was positively associated with LOAD (p=8.55e-07; see **Table 5.15**). *PTK2B* is a LOAD risk gene which localizes specifically to neurons in adult cortex (Salazar *at al.*, 2019). *PTK2B* is directly implicated in a neuronal Aβ signalling pathway and can cause impaired synaptic anatomy and function (Salazar *at al.*, 2019).
- ILMN_1721035 located on chromosome 11, annotated to the gene *MS4A6A* (see **Figure 5.34**) was positively associated with LOAD (p=1.18e-13; see **Table 5.15**). *MS4A6A* is the nearest gene to one of the top LOAD GWAS loci. For more details on the MS4A see **5.3.9.4.2**.
- LMN_2370336 - located on chromosome 11, annotated to the gene *MS4A4A* (see **Figure 5.34**) was positively associated with LOAD (p=7.51e-08; see **Table 5.15**). Interestingly, in a recent study a significant LOAD GWAS SNP (rs1582763; 11:60021948) was associated with both the expression of *MS4A4A* and *MS4A6A* (Deming *at al.*, 2019), suggesting both of these genes likely play a role in the aetiology of LOAD.

Even at a more relaxed threshold of p<1e-04 there were too few genes identified in this analysis to run GO analysis.

Table 5.14: Summary data-based Mendelian Randomisation (SMR) test results for pleiotropic effects between LOAD and gene expression using the Kunkle at al. GWAS and the blood eQTL dataset. (N = 3 hits where p < and HEIDI > 0.01). *CHR: chromosome; BP: base position (hg19 P); Instrument: top cis-eQTL for gene with smallest p-value for association; Beta SMR: estimate for the effect of gene expression on LOAD; P SMR: p-value for SMR association; HEIDI: p-value for the HEIDI test; N SNP HEIDI; number of SNPs included in the HEIDI test.*

Expression site				**GWAS**		**SMR**			**HEIDI**	
Probe ID	**CHR**	**BP**	**Annotated Gene**	**Instrument (CHR:BP)**	**Ref MAF**	**Beta**	**SE**	**P**	**P**	**N SNP**
ILMN_2330966	8	27316679	*PTK2B*	8:27227554	0.31	0.33	0.07	8.55E-07	4.29E-02	6
ILMN_1721035	11	59940569	*MS4A6A*	11:59945745	0.41	0.24	0.03	1.18E-13	4.58E-02	20
ILMN_2370336	11	60075868	*MS4A4A*	11:60099225	0.36	0.76	0.14	7.51E-08	6.03E-01	19

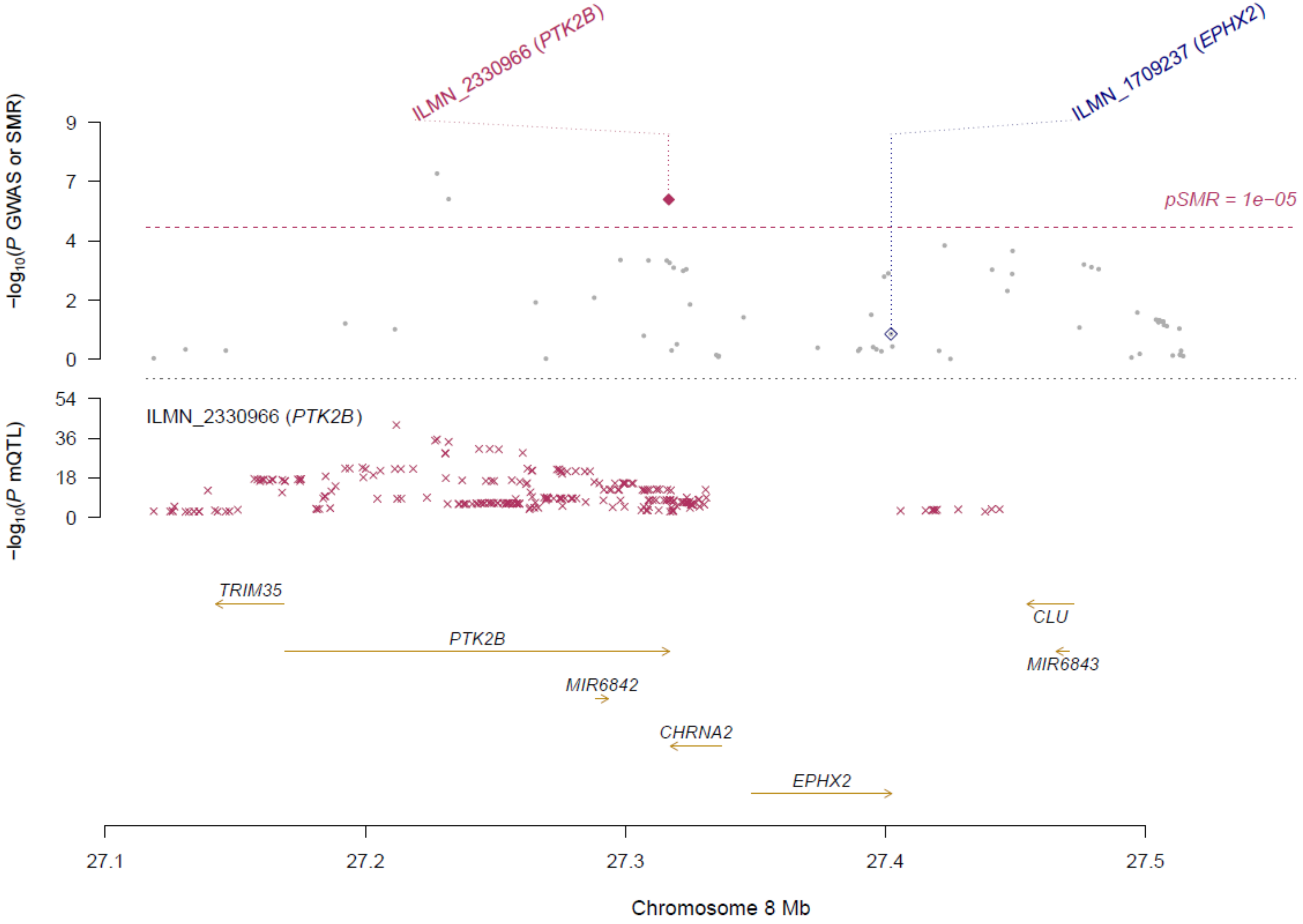


***Figure 5.33: Prioritizing genes at the PTK2B GWAS locus using SMR analysis with whole blood eQTLs and the Kunkle et al LOAD GWAS**. Shown are results at the **PTK2B** locus (chromosome 8) for LOAD. Top plot, grey dots represent the P values for SNPs from the Kunkle et al LOAD GWAS, diamonds represent the P values for probes from the SMR test – those filled in and highlighted in red passed the SMR and HEIDI tests. Bottom plots, the eQTL P values of SNPs from the whole blood dataset for the probes around the PTK2B locus. The top and bottom plots include all the SNPs available in the region in the GWAS and mQTL summary data, respectively, not just the SNPs common to both datasets.*

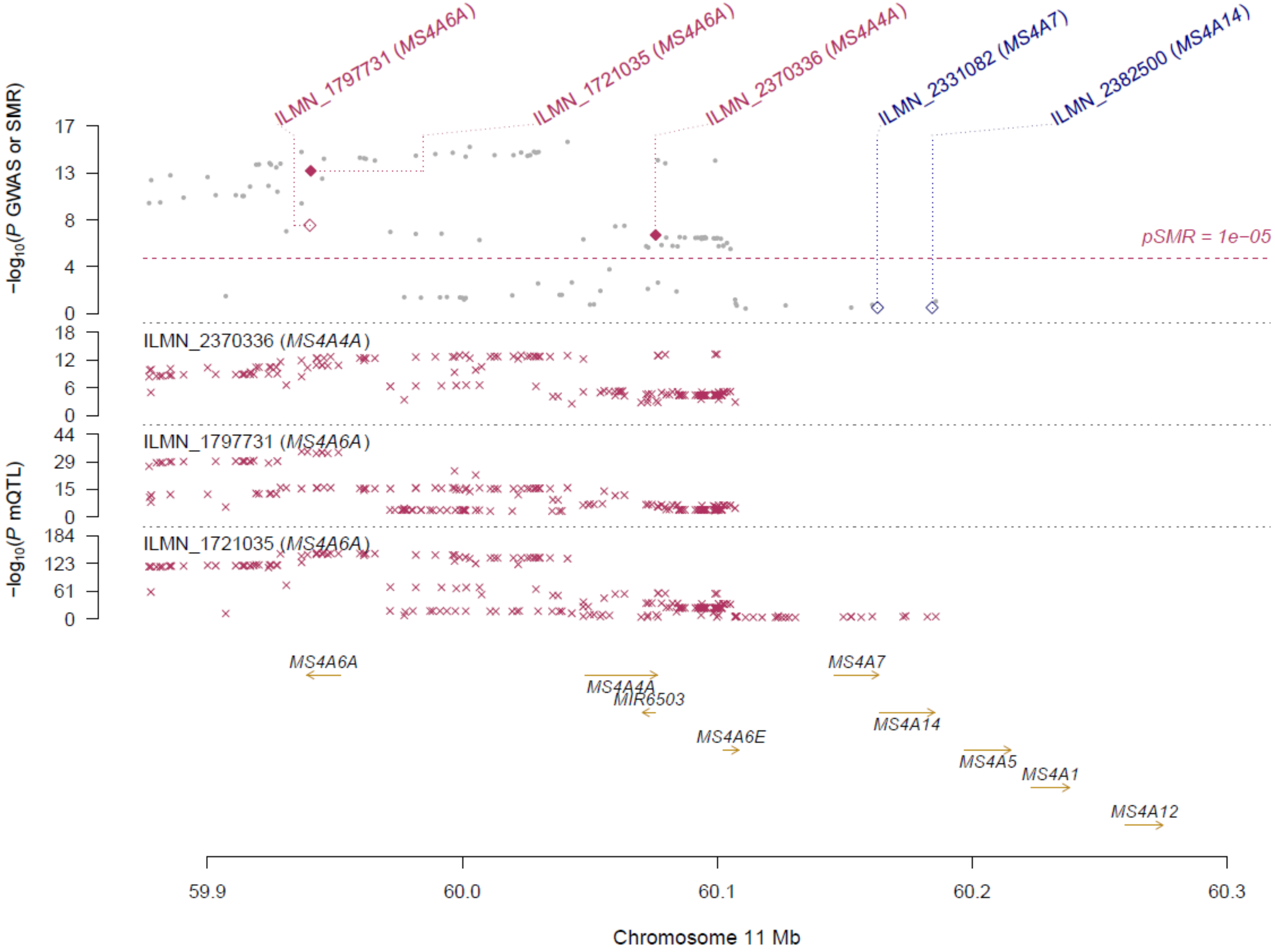


***Figure 5.34: Prioritizing genes at the MS4A GWAS locus using SMR analysis with whole blood eQTLs and the Kunkle et al LOAD GWAS**. Shown are results at the MS4A locus (chromosome 8) for LOAD. Top plot, grey dots represent the P values for SNPs from the Kunkle et al LOAD GWAS, diamonds represent the P values for probes from the SMR test – those filled in and highlighted in red passed the SMR and HEIDI tests. Bottom plots, the eQTL P values of SNPs from the whole blood dataset for the probes around the MS4A locus. The top and bottom plots include all the SNPs available in the region in the GWAS and mQTL summary data, respectively, not just the SNPs common to both datasets.*

5.4.11.2 SMR results using the Jansen et al GWAS

Using the Jansen *at al.* GWAS I applied SMR and identified 11 associations which passed the Bonferroni significant threshold (p < 8.38 e-06) and seven which passed both the significance threshold and the HEIDI threshold (p > 0.01; see **Table 5.16** and **Figure 5.13**). Six (85.7%) were positively associated with LOAD. Of the gene expression probes which were prioritised by SMR as having a pleiotropic relationship between a SNP and LOAD, several are annotated to genes which have previously been implicated in LOAD:

- ILMN_1729915 - located on chromosome 7, annotated to *PILRA* was positively associated with LOAD (p=1.68e-06; see **Table 5.16**). *PILRA* has been linked to LOAD via GWAS and in addition whole exome sequencing supports a role for this gene in LOAD (Patel *at al.*, 2018).
- ILMN_2330966 - located on chromosome 8 *PTK2B* was positively associated with LOAD (p=3.55e-07; see **Table 5.16)**. *PTK2B* is a LOAD risk gene which localizes specifically to neurons in adult cortex (Salazar *at al.*, 2019). *PTK2B* is directly implicated in a neuronal Aβ signalling pathway and can cause impaired synaptic anatomy and function (Salazar *at al.*, 2019).
- ILMN_2370336 - located on chromosome 11 annotated to *MS4A4A* was positively associated with LOAD (p=1.43e-07; see **Table 5.16)**. *MS4A4A* is the nearest gene to one of the top LOAD GWAS loci. It is part of the MS4A locus (see **5.4.8.1.2** for more details on the MS4A locus).
- ILMN_1657797 - located on chromosome 11 annotated to *FIBP* was negatively associated with LOAD (p=9.89e-06; see **Table 5.16**). *FIBP* is differentially expressed in the hippocampus and evidence suggests it may play a role in AD pathogenesis (Zhang *at al.*, 2015).
- ILMN_1693394 - located on chromosome 16 annotated to *BCKDK* was positively associated with LOAD (p=8.21e-06; see **Table 5.16)**. *BCKDK* is an AD risk loci with involved in the regulation of the valine, leucine, and isoleucine catabolic pathways.

The analysis also highlighted novel genes, which have not been previously been implicated in LOAD including:

- ILMN_1794364 - located on chromosome 11 annotated to *CTSW* was positively associated with LOAD (p=6.28e-06; see **Table 5.16)**. *CTSW* is a lysosomal cathepsin; a proteolytic enzyme. *CTSW* is found in CD8+ T cells and natural killer cells. The expression of *CTSW* in cytotoxic T-lymphocytes suggests it is involved in regulation of T-cell cytolytic activity (Linnevers, Smeekens, & Brömme, 1997). Differentiated CD8+ T-cells lacking cytolytic activity have identified in white matter of AD cases (Smolders *at al.*, 2013).
- ILMN_1739236 - located on chromosome 16 annotated to *ZNF668* was positively associated with LOAD (p=7.81e-06; see **Table 5.16)**. *ZNF668* is a zinc finger protein and although this specific gene has not been implicated in AD, zinc finger genes are known to play a role in the development of AD.

Even at a more relaxed threshold of p<1e-04 there were too few genes identified in this analysis to run GO analysis.

Table 5.15: Summary data-based Mendelian Randomisation (SMR) test results for pleiotropic effects between LOAD and gene expression using the Jansen at al. GWAS and the blood eQTL dataset. (N = 7 hits where p < and HEIDI > 0.01). *CHR: chromosome; BP: base position (hg19 P); Instrument: top cis-eQTL for gene with smallest p-value for association; Beta SMR: estimate for the effect of gene expression on LOAD; P SMR: p-value for SMR association; HEIDI: p-value for the HEIDI test; N SNP HEIDI; number of SNPs included in the HEIDI test.*

Expression site				**GWAS**		**SMR**			**HEIDI**	
Probe ID	**CHR**	**BP**	**Annotated Gene**	**Instrument (CHR:BP)**	**Ref MAF**	**Beta**	**SE**	**P**	**P**	**N SNP**
ILMN_1729915	7	99997440	*PILRA*	7:99971313	0.32	0.14	0.03	1.68E-06	8.90E-01	9
ILMN_2330966	8	27316679	*PTK2B*	8:27227554	0.31	0.05	9.60E-03	3.55E-07	1.11E-01	7
ILMN_2370336	11	60075868	*MS4A4A*	11:60099225	0.36	0.11	0.02	1.43E-07	4.65E-01	19
ILMN_1794364	11	65651156	*CTSW*	11:65646557	0.19	0.01	2.64E-03	6.28E-06	6.44E-01	20
ILMN_1657797	11	65651435	*FIBP*	11:65655393	0.19	-0.03	6.78E-03	9.89E-06	2.04E-01	18
ILMN_1739236	16	31072282	*ZNF668*	16:31048079	0.39	0.02	5.53E-03	7.81E-06	1.75E-02	19
ILMN_1693394	16	31123837	*BCKDK*	16:31141993	0.38	0.04	8.30E-03	8.21E-06	2.89E-01	20

5.4.12 Identifying putative pleiotropic relationships between gene expression and Alzheimer's disease in the cortex using SMR

I used SMR utilising publicly availably cortex eQTL data which included 7,370 expression probes (annotated to 7345 genes; a meta-analysis of 3 cortex eQTL datasets) with the two latest AD GWAS with publicly available summary statistics.

5.4.12.1 SMR results using the Kunkle et al GWAS

Using the Kunkle *at al.* GWAS I applied SMR, identifying seven associations which passed the Bonferroni significant threshold (p < 6.78e-06) and two which passed both the significance threshold and the HEIDI threshold (p > 0.01; see **Table 5.17** and **Figure 5.25**). Both are annotated to genes which have previously been implicated in LOAD:

- ILMN_1711611 - located on chromosome 1, annotated to gene *CR1* was positively associated with LOAD (p = 1.98e-09; see **Table 5.17**). *CR1* encodes a type-I transmembrane glycoprotein, has been identified as a risk gene for LOAD. *CR1* is linked to elevated cerebrospinal fluid Aβ42 levels in AD patients (Cruchaga *at al.*, 2013) and influences the severity of vascular amyloid deposition (Biffi *at al.*, 2012).
- ILMN_1693242 - located on chromosome 19 annotated to gene *ZNF296* was negatively associated with LOAD (p = 5.96e-06; see **Table 5.17**). *ZNF296* is a zinc finger protein found around the *APOE* region.

Even at a more relaxed threshold of p<1e-04 there were too few genes to run GO analysis.

Table 5.16: SMR test results for pleiotropic effects between LOAD and gene expression using the Kunkle at al. GWAS and the cortex eQTL dataset. (N = 2 hits where p < 6.78e-06 and HEIDI > 0.01). *CHR: chromosome; BP: base position (hg19 P); Instrument: top cis-eQTL for gene with smallest p-value for association; Beta SMR: estimate for effect of gene expression on LOAD; P SMR: p-value for SMR association; HEIDI: p-value for the HEIDI test; N SNP HEIDI; number of SNPs included in the HEIDI test.*

Expression site				**GWAS**		**SMR**			**HEIDI**	
Probe ID	**CHR**	**BP**	**Annotated Gene**	**Instrument (CHR:BP)**	**Ref MAF**	**Beta**	**SE**	**P**	**P**	**N SNP**
ILMN_1711611	1	207741742	*CR1*	1:207750568	0.82	0.26	0.04	1.98E-09	3.73e-01	15
ILMN_1693242	19	45577302	*ZNF296*	19:45584692	0.83	-0.32	0.07	5.96E-06	1.05e-02	19

5.4.12.2 SMR results using the Jansen et al GWAS

Using the Jansen *at al.* GWAS I applied SMR and identified nine associations which passed the Bonferroni significant threshold (p < 6.78e-06) and five which passed both the significance threshold and the HEIDI threshold (p > 0.01; see **Table 5.18** and **Figure 5.25**). Of these five, one (20%) was positively associated with LOAD. Of the gene expression sites which were prioritised by SMR as having a pleiotropic relationship between a SNP and LOAD all five are annotated to genes which have previously been implicated in LOAD.

- ILMN_1789342 – located on chromosome 1, annotated to gene *NDUFS2* was negatively associated with LOAD (p= 4.28e-06; see **Table 5.18**). *NDUFS2* is a mitochondrial gene which is differentially expressed in *APP*/PS1 transgenic mice (Lachén-Montes *at al.*, 2016).
- ILMN_1711611 – located on chromosome 1, annotated to gene *CR1* was positively associated with LOAD (p= 5.17e-10; see **Table 5.18**). For more details in *CR1* see **5.4.13.2.**
- ILMN_1683737 – located on chromosome 11, annotated to gene *SNX32* was negatively associated with LOAD (p= 3.81e-06; see **Table 5.18**). Previous research found a protein QTL (pQTL) at *SNX32* co-localised with Alzheimer disease (Kibinge, Relton, Gaunt, & Richardson, 2020).
- ILMN_3239881 – located on chromosome 16, annotated to gene *PRSS36* was negatively associated with LOAD (p= 4.70e-06; see **Table 5.18**). *PRSS36* is an AD GWAS gene which has been implicated via eQTL association in the hippocampus, a cortex region highly affected early in AD pathogenesis (Jansen *at al.*, 2019).
- ILMN_3179620 – located on chromosome 17, annotated to *AC012146.7*, was negatively associated with LOAD (p=1.03E-06; see **Table 5.18**). *AC012146.7* is located 120kb from the LOAD risk gene *SCIMP* which has been implicated in innate immunity and immune response (Jansen *at al.*, 2019).

Even at a more relaxed threshold of p< 1e-04 there were too few genes to run GO analysis.

Table 5.17: Summary data-based Mendelian Randomisation (SMR) test results for pleiotropic effects between LOAD and gene expression using the Jansen at al. GWAS and the cortex eQTL dataset. (N = 5 hits where p < 6.78e-06 and HEIDI > 0.01). *CHR: chromosome; BP: base position (hg19 P); Instrument: top cis-eQTL for gene with smallest p-value for association; Beta SMR: estimate for effect of gene expression on LOAD; P SMR: p-value for SMR association; HEIDI: p-value for the HEIDI test; N SNP HEIDI; number of SNPs included in the HEIDI test.*

Expression site				**GWAS**		**SMR**			**HEIDI**	
Probe ID	**CHR**	**BP**	**Annotated Gene**	**Instrument (CHR:BP)**	**Ref MAF**	**Beta**	**SE**	**P**	**P**	**N SNP**
ILMN_1789342	1	161175539	*NDUFS2*	1:161186313	0.69	-0.03	7.31E-03	4.28E-06	2.48E-01	20
ILMN_1711611	1	207741742	*CR1*	1:207750568	0.82	0.04	6.76E-03	5.17E-10	2.30E-02	15
ILMN_1683737	11	65612739	*SNX32*	11:65601560	0.85	-0.01	2.10E-03	3.81E-06	2.90E-02	20
ILMN_3239881	16	31155830	*PRSS36*	16:31154146	0.72	-0.02	3.64E-03	4.70E-06	1.07E-01	20
ILMN_3179620	17	5016531	*AC012146.7*	17:5014212	0.90	-0.02	5.04E-03	1.03E-06	1.57E-01	20

5.4.13 Comparing the whole blood and cortex eQTL SMR results

I looked for consistency in the direction of effect across the expression SMR analyses, including all SMR results which reached nominal significance (p SMR<0.05). There was concordance across the two LOAD GWAS summary statistics within the same tissue as highlighted by the positive correlations and highly significant (Bonferroni p < 0.05/24 = 0.002) sign test p-values (see **Figure 5.35 A, D, G and J**). Several of the eQTL SMR analyses were characterised by the same direction of effect including the Kunkle whole blood and Jansen cortex (sign test p=0.0018; see **Figure 5.35C**), Jansen whole blood and Jansen cortex (sign test p=6.9e-06; see **Figure 5.35F**) and the Jansen cortex and the Jansen whole blood (sign test p=9.5e-05; see **Figure 5.35K**). However, there was evidence of heterogeneity between tissues. For example, when comparing the Kunkle cortex to the Jansen whole blood eQTL SMR analysis there was no concordance between the effect sizes (p=0.24; see **Figure 5.35I**). These results suggest that although there is evidence for homogeneity for some SMR-nominated loci across tissues, there was also evidence of heterogeneity inferring there may be tissue specific transcriptomic differences.

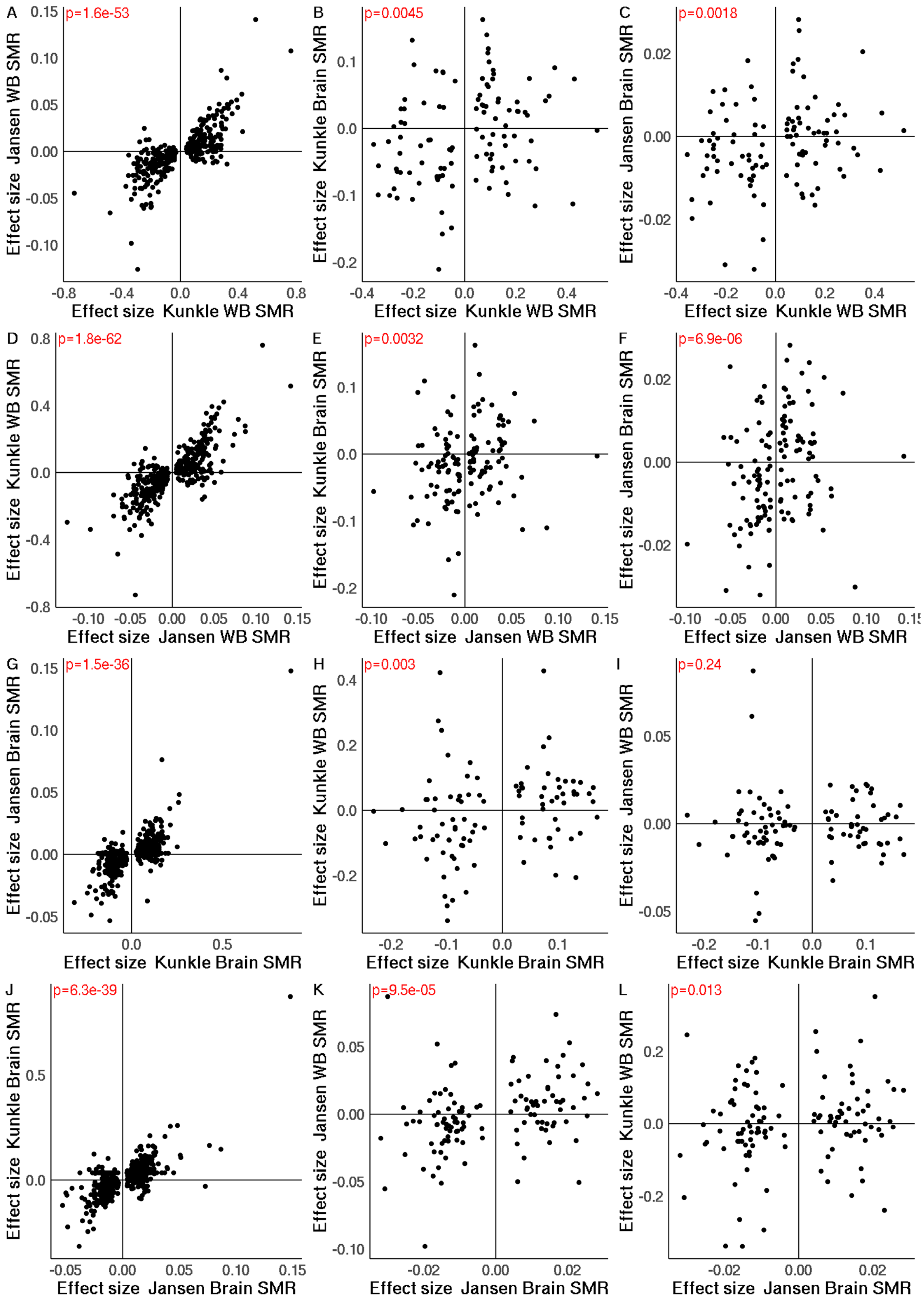


Figure 5.35: Comparisons of the direction across all eQTL SMR analyses when considering nominally significant probes (p SMR < 0.05). *p= binomial sign test p-value.*

I next looked at the overlapping Bonferroni significant expression sites and genes identified from the eQTL SMR analyses, finding three overlapping sites across the analyses (see **Figure 5.36**). When comparing the Kunkle and Jansen blood eQTL results there was an overlap of 2 sites (see **Table 5.19**), one was located on chromosome 8 (annotated to *PTK2B*) and the other was located on chromosome 11 (annotated to *MS4A4A*). When comparing the Kunkle and Jansen cortex eQTL results there was an overlap of one site (see **Table 5.20**), which was located on chromosome 1 (annotated to *CR1*). The replication of these results across the two AD GWAS provides internally consistent replication of results and suggests these genes have a role in LOAD pathogenesis and could be targeted for functional experiments. There was no overlap in the sites or genes identified when looking across the blood and cortex eQTL SMR results, however not all expression sites were tested in each analysis.

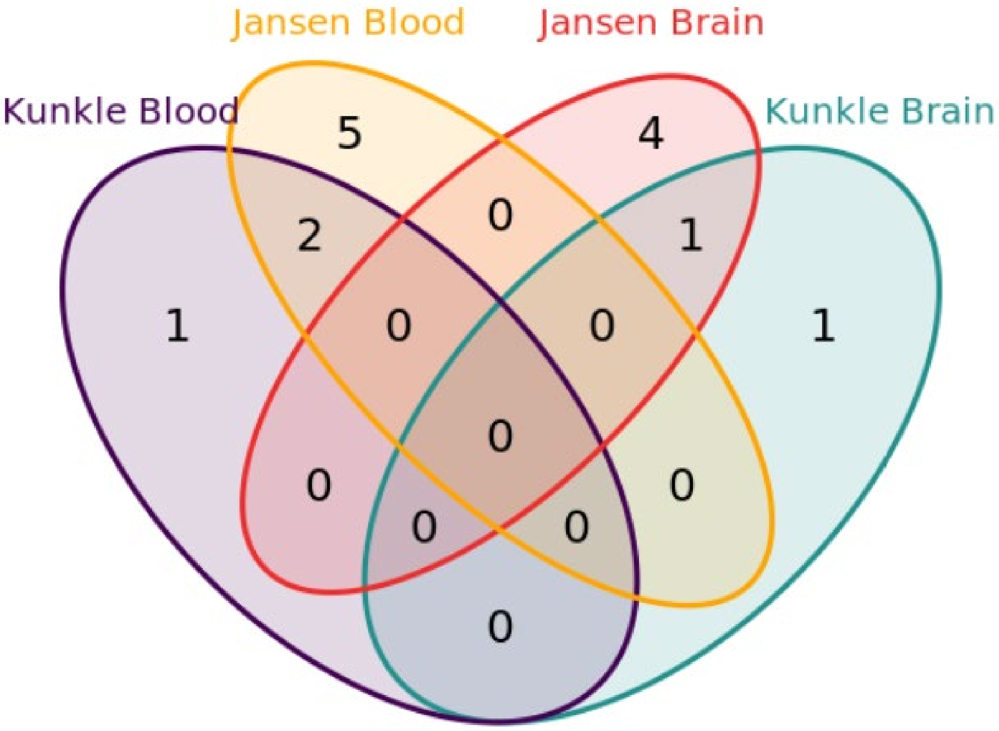


Figure 5.36: Overlap of the Summary data-based Mendelian Randomisation (SMR) test results for pleiotropic effects between LOAD and gene expression in whole blood and cortex mQTL datasets using both the Kunkle et al GWAS (2019) and the Jansen et al GWAS (2019).

Table 5.18: Overlapping expression sites from the results between 2 cortex eQTL SMR analysis: (1) using the Kunkle et al GWAS and (2) the Jansen et al GWAS.*DNAm: DNA methylation; CHR: chromosome; BP: position (hg19).*

DNAm Probe D	CHR	BP	Gene
ILMN_2330966	8	27316679	*PTK2B*
ILMN_2370336	11	60075868	*MS4A4A*

Table 5.19: Overlapping expression sites from the results between 2 cortex eQTL SMR analysis: (1) using the Kunkle et al GWAS and (2) the Jansen et al GWAS. *DNAm: DNA methylation; CHR: chromosome; BP: position (hg19).*

DNAm Probe D	CHR	BP	Gene
ILMN_1711611	1	207741742	*CR1*

5.4.14 Identifying putative pleiotropic relationships between DNA methylation and gene expression in whole blood using SMR

There is evidence indicating that DNAm is directly associated with levels of gene expression, although the mechanisms involved are not well characterised. Generally DNAm has been associated with transcriptional repression which is thought to occur in DNAm dense promoter regions as a consequence of reduced transcription factor binding (Ng & Adrian, 1999). However, DNAm in the gene body has been shown to be a marker of transcription with a potential role in alternative splicing. To investigate associations between DNAm and gene expression in whole blood I used the SMR tool. I included sites based on 3 inclusion criteria: (1) DNAm sites which had a significant mQTL (P<1e-10); (2) expression sites which had a significant eQTL (P<5e-08); and (3) there was a common genetic variant tested within 500 kilobases of the gene expression probe and DNAm site. In whole blood tested I 119,352 pairs between 61,098 DNAm sites and 5,512 gene expression probes (annotated to 4,780 unique genes). I identified 105,389 pairs which had a significant SMR result (Bonferroni p= 4.47e-07). I identified 26,978 pairs, comprised of 20,598 DNAm sites and 4,624 expression sites which had a significant SMR result and which passed the HEIDI test (HEIDI p > 0.01). DNAm sites which have a pleiotropic association with expression are enriched in gene bodies (OR=1.83; p=2.2e-308) and 200bp from the transcription start site (OR=1.78; p=7.06e-173) but are less represented in intergenic regions (OR=0.95; p= 3.24e-04). 12,505 DNAm sites (46%) have positive associations with gene expression. In situations where DNAm is negatively associated with gene expression (54%), there was evidence that these sites are enriched in promoter and enhancer regions including transcription start sites (200bp from transcription start site OR=1.55, p= 1.61E-30; 1500bp from transcription start site, OR=1.72, p= 1.97E-103), 5'UTR (OR=1.54; p= 2.13E-51), 3'UTR (OR=1.40; p= 6.75E-10) and 1stExon (OR=1.52; p= 2.60E-16). This goes in line with previous research suggesting that promoter DNAm generally represses gene expression (Tate & Bird, 1993).

I used these results in combination with the SMR analysis for each molecular marker to identify pleiotropic relationships between DNAm, expression and LOAD (i.e. to test if DNAm, a transcript and LOAD are associated because of a shared causal variant). In addition, I used this information as a means of annotating DNAm sites to genes where there was no such annotation on the Illumina manifest.

5.4.14.1 SMR results using the Kunkle et al GWAS

To identify pleiotropic relationships between DNAm, expression and LOAD (i.e. to test if DNAm, a transcript and LOAD are associated because of a shared causal variant), I utilised the mQTL-eQTL, DNAm-LOAD and expression-LOAD results, identifying situations where the association signals were consistent across the three analyses at a locus. First, I tested these relationships using the whole blood SMR results which were generated using the Kunkle *at al.* GWAS. I included SMR significant mQTL results but relaxed the threshold of the SMR expression results (eQTL p SMR < 1e-4) since an mQTL-eQTL relationship would strengthen the hypothesis of a gene being involved in the pathogenesis of LOAD. By incorporating the whole blood mQTL-eQTL analysis with the whole blood Kunkle *at al.* SMR mQTL and eQTL analyses, I identified nine pleiotropic relationships where DNAm, a transcript and LOAD have a pleiotropic association with a shared variant (see **Table 5.21**). Of these nine, three (27%) DNAm sites had a positive association with gene expression (i.e. hypermethylation is associated with gene expression).

Where there is no gene annotation based on proximity e.g. for cg27552578, cg05585544, cg17688768 and cg18512352, the gene annotation of the associated expression site can be used, and in this case these sites were negatively associated with the gene expression probe ILMN_1686516, which resides in the locus of *CELF1*/CUGBP1 (p=4.45e-08; p=1.35e-07; p= 1.03e-08; and p=1.63e-08, respectively). The four DNAm sites were positively associated with LOAD in the DNAm-LOAD analysis (see **Table 5.2**), whereas the expression site was negatively associated with LOAD in the expression-LOAD analysis (p= 5.18e-05). Therefore, increased DNAm and decreased expression is associated with LOAD which is further supported by the mQTL-eQTL results, where DNAm is associated with reduced gene expression (see **Table 5.21**). Additionally, these four DNAm sites were identified to be genetically co-regulated with each other using Bayesian colocalisation; there is evidence that DNAm at these sites are driven by the same genetic variant and are correlated with one another (see **Table 5.3**). In addition, the sites cg25206146 (annotated to *SPI1* based on the illumina manifest), and cg20449816 (annotated to *SLC39A13* based on the illumina manifest) were significantly associated with the same expression site (p=7.68e-07; p= 8.15e-08, respectively). These results suggest the most functionally relevant gene around this region based on these data is

CELF1/CUGBP1. These DNAm sites cover a broad genomic region (~250kb), although all DNAm sites are negatively associated with expression of the same gene expression probe (ILMN_1686516). The *CELF1* locus is located adjacent to the *SPI1* gene and different GWAS have prioritised either *SPI1* or *CELF1* as the top gene to a GWAS loci (Gjoneska *at al.*, 2015; Jansen *at al.*, 2019; Karch *at al.*, 2016). Expression of several genes within the *CELF1* locus are highly correlated with one another and have been associated with AD status (Karch *at al.*, 2016).

In addition, the mQTL-expression-LOAD results suggest that the most relevant genes around the MS4A locus are *MS4A4A* and *MS4A6A* (see **Table 5.21** and **Figure 5.37**). Two of the unannotated probes in this region (cg02521229 and cg18684128) were positively associated with expression of ILMN_2370336 which is annotated to *MS4A4A* (2.01e-13; p=5.31e-10, respectively). cg18684128 was also significantly associated with a second expression site - ILMN_1721035 (p=1.50e-25) - which is annotated to *MS4A6A* (see **Figure 5.37**). Additionally, the two DNAm sites were identified to be genetically co-regulated with each using Bayesian colocalisation (see **Table 5.3**). Both the DNAm site and expression probe were independently positively associated with LOAD. These results highlight the complex relationship between SNPs, DNAm, gene expression and LOAD.

Overall, when considering pleiotropic relationships between DNAm, gene expression and LOAD, the genes *CELF1*/CUGBP1, *MS4A4A* and *MS4A6A* were prioritised by my analyses. These results suggest these genes may be the most functionally relevant within these regions based on our data and could be potential targets for future functional studies.

Table 5.20: SMR test results for pleiotropic effects between DNAm and gene expression in whole blood based on the Kunkle et al QTL SMR results. *Shown are the SMR results for pleiotropic effects between DNAm and gene expression, where the DNAm sites and expression sites were previously identified as having a pleiotropic association with LOAD when incorporating the Kunkle at al. GWAS, mQTLs and DNAm in SMR analysis. CHR: chromosome; BP: base position (hg19 P); Instrument: top cis-mQTL for gene with smallest p-value for association; Beta SMR: causal estimate for effect of each unit increase in DNA methylation on gene expression; P SMR: p-value for SMR association; HEIDI: p-value for the HEIDI test; N SNP HEIDI; number of SNPs included in the HEIDI test.*

DNAm Site				**Expression Site**				**GWAS**		**SMR**			**HEIDI**	
Probe ID	**CHR**	**BP**	**Annotated Gene**	**Probe ID**	**CHR**	**BP**	**Annotated Gene**	**Instrument (CHR:BP)**	**Ref MAF**	**Beta**	**SE**	**P**	**P**	**N SNP**
cg25206146	11	47383181	*SPI1*	ILMN_1686516	11	47490374	*CELF1*;CUGBP1	11:47391039	0.32	-9.26	1.87	7.68E-07	1.09E-01	13
cg20449816	11	47432366	*SLC39A13*	ILMN_1686516	11	47490374	*CELF1*;CUGBP1	11:47432034	0.32	-4.07	0.76	8.15E-08	1.69E-01	13
cg27552578	11	47621330	-	ILMN_1686516	11	47490374	*CELF1*;CUGBP1	11:47663049	0.35	-9.19	1.68	4.45E-08	2.48E-01	14
cg05585544	11	47624801	-	ILMN_1686516	11	47490374	*CELF1*;CUGBP1	11:47716324	0.34	-10.88	2.06	1.35E-07	3.25E-01	14
cg17688768	11	47628441	-	ILMN_1686516	11	47490374	*CELF1*;CUGBP1	11:47648042	0.35	-2.16	0.38	1.03E-08	2.12E-01	14
cg18512352	11	47633146	-	ILMN_1686516	11	47490374	*CELF1*;CUGBP1	11:47648042	0.35	-5.58	0.99	1.63E-08	8.00E-02	14
cg02521229	11	60019236	-	ILMN_2370336	11	60075868	*MS4A4A*	11:60013857	0.42	1.15	0.16	2.01E-13	2.30E-02	20
cg18684128	11	60033393	-	ILMN_1721035	11	59940569	*MS4A6A*	11:60030457	0.41	32.09	3.07	1.50E-25	4.00E-02	20
cg18684128	11	60033393	-	ILMN_2370336	11	60075868	*MS4A4A*	11:60030457	0.41	10.03	1.61	5.31E-10	2.78E-01	20

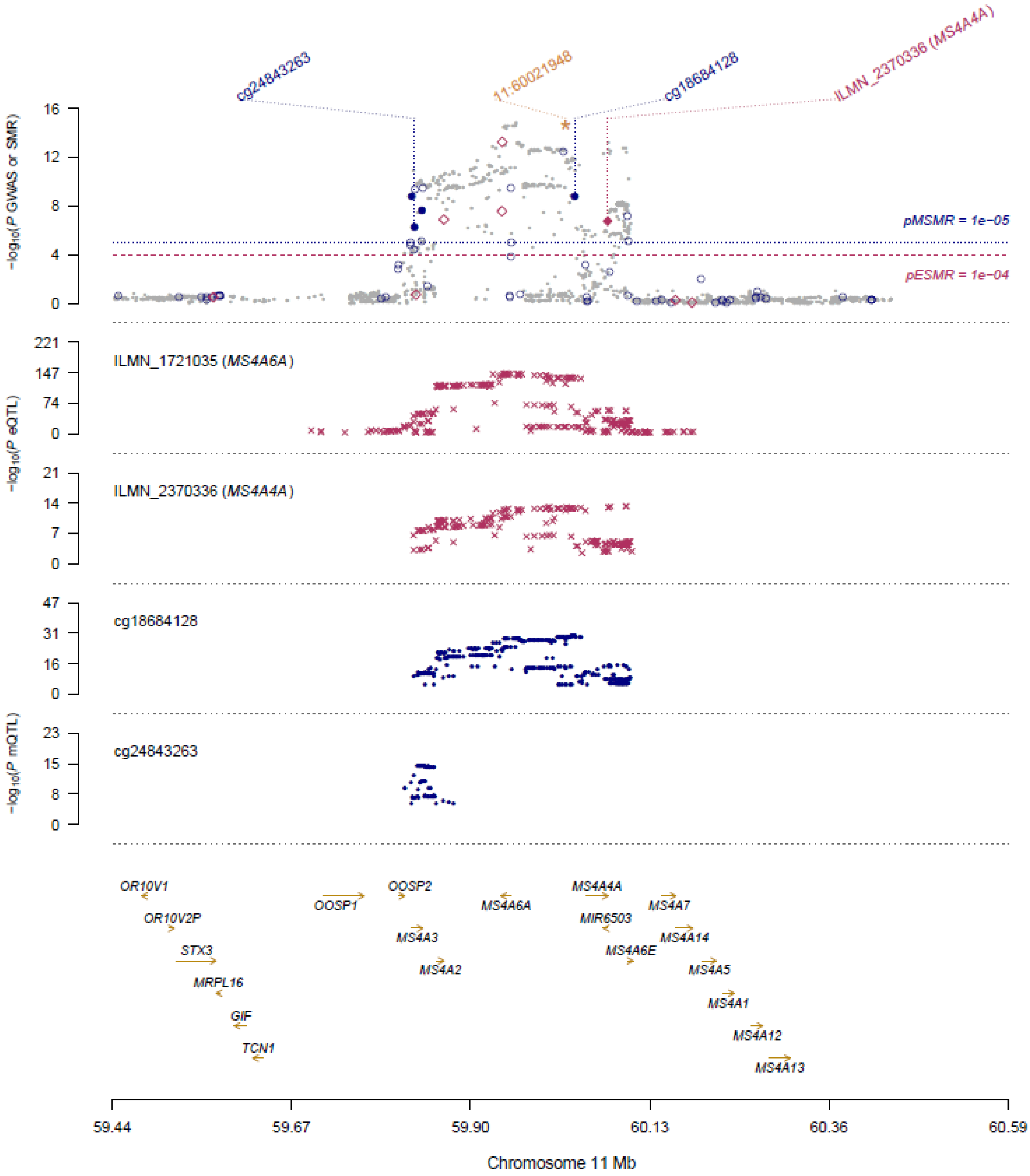


Figure 5.37: Results of SNP and SMR associations across mQTL, eQTL and GWAS in the whole blood datasets using the Kunkle et al GWAS. *The top plot shows −log10 (P) of SNPs from the Kunkle et al LOAD GWAS. The red diamonds and blue circles represent −log10 (P) from SMR tests for associations of gene expression and DNAm probes with LOAD, respectively. The solid diamonds and circles are the probes not rejected by the HEIDI test. The second plot shows −log10(P) of the SNP association for gene expression probes ILMN_1721035 (tagging MS4A6A) and ILMN_2370336 (tagging MS4A6A from the Westera eQTL study. The third plot shows −log10 (P) of the SNP associations for DNAm probe cg18684128 from the mQTL study. The yellow star indicates the previously reported causal variant rs1582763 (11:60021948, annotated to gene MS4A4A).*

5.4.14.2 SMR results using the Jansen et al GWAS

To identify pleiotropic relationships between DNAm, expression and LOAD, I applied the same method as described above (section **5.4.15**). I identified nine pleiotropic associations where DNAm, a transcript and LOAD were associated because of a shared causal variant (see **Table 5.22**). Of these nine, six (67%) DNAm sites had a positive association with gene expression – i.e. increased methylation was associated with increased expression.

There was one probe which was not annotated from the SMR mQTL analysis - cg18684128 – and using the mQTL-eQTL SMR results, this probe was positively associated with expression of ILMN_2370336 which is annotated to *MS4A4A* and goes in line with the Kunkle results.

Each of the DNAm probes prioritised by SMR in the mQTL-eQTL analysis are located on chromosome 7 within a 320kb region of one another, and were associated with expression of the same probe (ILMN_1729915) annotated to *PILRA* (see **Figure 5.38**). It is possible to refine the annotations based on the expression data as opposed to using the proximally located genes. Although these results suggest a role for *PILRA* in the development of LOAD, several of the genes prioritised from the mQTL SMR analysis were not tested in the eQTL SMR analysis (*GPC2*, *STAG3*, *NYAP1*, and *PMS2L1*). Therefore it may not be the only functionally relevant gene in this region.

Overall, these results prioritised the genes *PILRA* and *MS4A4A* for further investigation in LOAD.

Table 5.21: SMR test results for pleiotropic effects between DNAm and gene expression in whole blood based on the Jansen et al SMR results. *Shown are the SMR results for pleiotropic effects between DNAm and gene expression, where the DNAm sites and expression sites were previously identified as having a pleiotropic association with LOAD when incorporating the Jansen at al. GWAS, mQTLs and DNAm in SMR analysis. CHR: chromosome; BP: base position (hg19); Instrument: top cis-mQTL for gene with smallest p-value for association; Beta SMR: causal estimate for effect of each unit increase in DNA methylation on gene expression; P SMR: p-value for SMR association; HEIDI: p-value for the HEIDI test; N SNP HEIDI; number of SNPs included in the HEIDI test.*

DNAm Site				Expression Site				GWAS		SMR			HEIDI	
Probe ID	**CHR**	**BP**	**Annotated Gene**	**Probe ID**	**CHR**	**BP**	**Annotated Gene**	**Instrument (CHR:BP)**	**Ref MAF**	**Beta**	**SE**	**P**	**P**	**N SNP**
cg18090197	7	99769602	*GPC2*	ILMN_1729915	7	99997440	*PILRA*	7:99792608	0.26	13.70	3.02	5.72E-06	4.40E-02	6
cg00048759	7	99775422	*STAG3*	ILMN_1729915	7	99997440	*PILRA*	7:99807146	0.26	-6.36	1.41	6.13E-06	2.31E-01	6
cg10084644	7	99775521	*STAG3*	ILMN_1729915	7	99997440	*PILRA*	7:99815247	0.26	-4.98	1.05	2.36E-06	5.80E-02	6
cg00553149	7	99775558	*STAG3*	ILMN_1729915	7	99997440	*PILRA*	7:99807146	0.26	-2.94	0.61	1.14E-06	1.90E-01	6
cg10407106	7	99779719	*STAG3*	ILMN_1729915	7	99997440	*PILRA*	7:99784704	0.26	5.69	1.21	2.39E-06	1.67E-01	6
cg17830204	7	99819110	GATS	ILMN_1729915	7	99997440	*PILRA*	7:99787372	0.26	3.66	0.74	8.44E-07	8.80E-02	6
cg19116668	7	99932089	*PMS2L1*	ILMN_1729915	7	99997440	*PILRA*	7:99971834	0.31	4.56	0.91	4.76E-07	5.30E-02	6
cg03579757	7	100091793	*NYAP1*	ILMN_1729915	7	99997440	*PILRA*	7:100004446	0.30	6.13	1.19	2.37E-07	8.30E-02	6
cg18684128	11	60033393	-	ILMN_2370336	11	60075868	*MS4A4A*	11:60030457	0.41	10.00	1.61	5.31E-10	2.78E-01	20

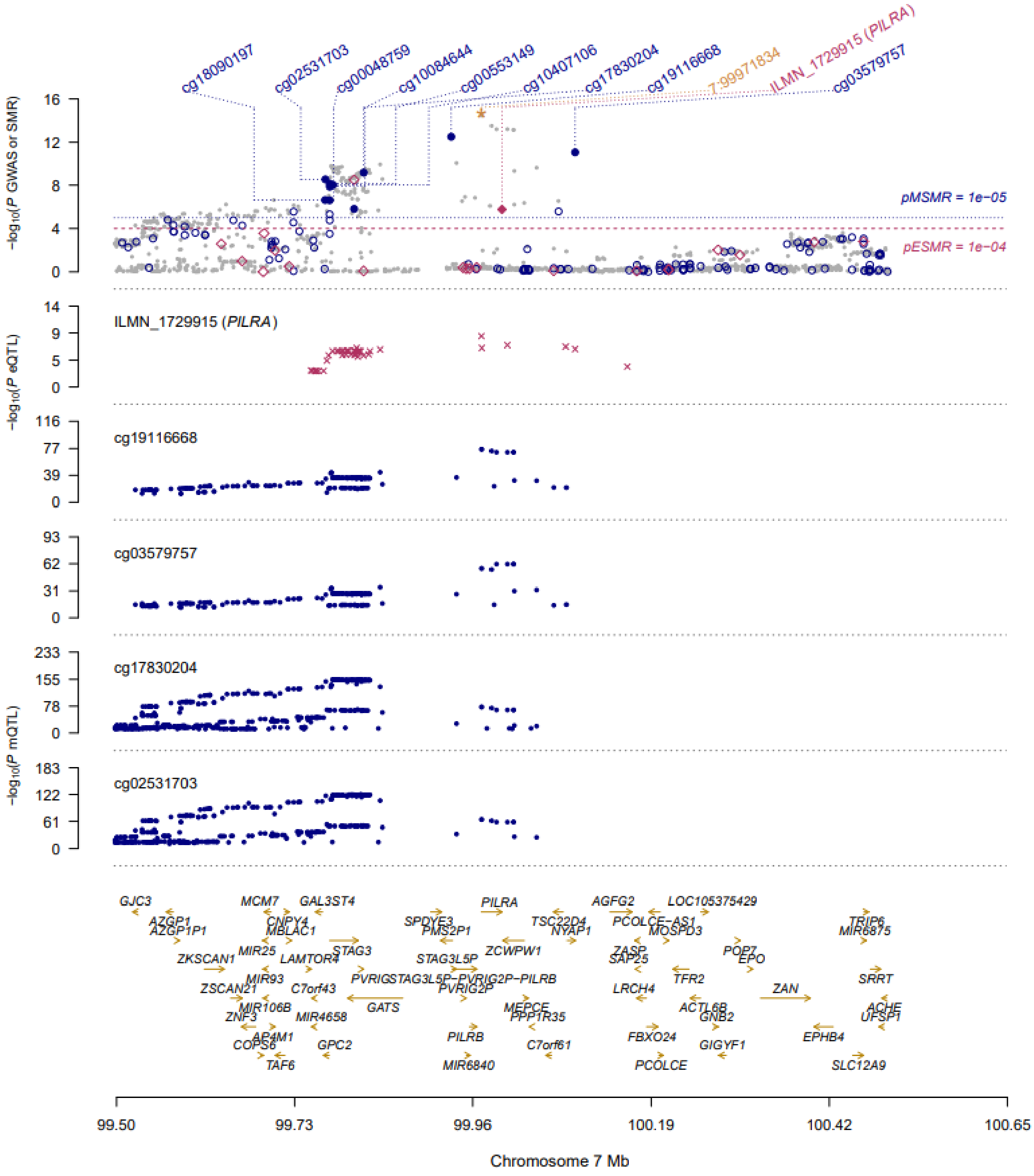


Figure 5.38: Results of SNP and SMR associations across mQTL, eQTL and GWAS in the whole blood datasets using the Jansen et al GWAS. *The top plot shows −log10 (P) of SNPs from the Jansen et al LOAD GWAS. The red diamonds and blue circles represent −log10 (P) from SMR tests for associations of gene expression and DNAm probes with LOAD, respectively. The solid diamonds and circles are the probes not rejected by the HEIDI test. The second plot shows −log10 (P) of the SNP association for gene expression probes ILMN_1729915 (tagging PILRA) from the Westra eQTL study. The third plot shows −log10 (P) of the SNP associations for DNAm probes cg1916668, cg03579757, cg17830204 and cg02531703 from the mQTL study. The yellow star indicates the previously reported causal variant rs1859788 (7: 99971834, annotated to gene ZCWPW1).*

5.4.15 Identifying putative pleiotropic relationships between DNA methylation and gene expression in cortex using SMR

Using SMR, I tested 23,333 pairs between 13,210 DNAm sites and 4,265 gene expression probes (annotated to 4,251 unique genes). I identified 18,412 pairs which had a significant SMR results (Bonferroni p = 2.14e-6). I identified 8,321 pairs, comprised of 6,326 DNAm sites and 2,786 expression sites which had a significant SMR result and which passed the HEIDI test (HEIDI p > 0.01). DNAm sites which have a pleiotropic association with expression are enriched in gene bodies (OR=1.89; p= 1.82e-187) and the transcription start sites (OR=2.63; p= 1.04e-113) but are less represented in intergenic regions (OR=0.81; p= 1.26e-15). 3,782 DNAm sites (45%) have positive effects on gene expression. In situations where DNAm is negatively associated with gene expression (55%), there was evidence that these sites are enriched in promoter and enhancer regions including transcription start sites (200bp from transcription start site OR=2.32, p=1.42E-65; 1500bp from transcription start site, OR=2.07, p= 8.71E-95), 5'UTR (OR=1.81; p=1.71E-45), 3'UTR (OR=1.82; p=3.05E-14) and 1stExon (OR=2.27; p=5.28E-32). This is concordant with the whole blood results and again goes in line with previous research suggesting that promoter DNAm represses gene expression (Tate & Bird, 1993).

5.4.15.1 SMR results using the Kunkle et al GWAS

In order to identify pleiotropic relationships between DNAm, gene expression and AD in cortex I applied the same method as described in **5.4.15**. First, I tested these relationships using the cortex SMR results which were generated using the Kunkle *at al.* GWAS. By incorporating the cortex mQTL-eQTL analysis with the cortex Kunkle *at al.* SMR mQTL and eQTL analyses, I identified four pleiotropic relationships where DNAm, a transcript and LOAD were associated because of a shared causal variant (see **Table 5.23**). The four DNAm sites were associated with the expression of two gene expression probes (ILMN_1671710 and ILMN_1808122). There was one DNAm probe which was not annotated from the SMR mQTL analysis: cg24977308.

Using the mQTL-eQTL SMR results, this probe has an effect on the expression of *C1QTNF4* and this gene is potentially the most functionally relevant in this region. The DNAm site cg09507712 was associated with expression of ILMN_1671710 which is annotated to *MADD* and not *C1QTNF4* (the illumina gene annotation). However, *MADD* was not tested in the mQTL dataset and therefore we cannot robustly state this is the only relevant gene annotation for this DNAm probe in relation to LOAD. These results suggest there is likely a role for two genes within the same region: *MADD* and *C1QTNF4* based on pleiotropic relationships between DNAm, gene expression and AD.

5.4.15.2 SMR results using the Jansen et al GWAS

I ran the same analysis as above using the Jansen et al SMR results. No results passed both the p SMR thresholds set; there currently is not enough evidence to make conclusions about the effect of DNAm on gene expression and LOAD with these data.

Table 5.22 SMR test results for pleiotropic effects between DNAm and gene expression in cortex based on the Kunkle et al SMR results. *Shown are the SMR results for pleiotropic effects between DNAm and gene expression, where the DNAm sites were previously identified as having a pleiotropic association with LOAD when incorporating the Kunkle at al. GWAS, mQTLs and DNAm in SMR analysis. Beta SMR: estimate for effect of DNA methylation on gene expression; P SMR: p-value for SMR association; HEIDI: p-value for the HEIDI test; N SNP HEIDI; number of SNPs included in the HEIDI test.*

DNAm Site				**Expression Site**				**GWAS**		**SMR**			**HEIDI**	
Probe ID	**CHR**	**BP**	**Annotated Gene**	**Probe ID**	**CHR**	**BP**	**Annotated Gene**	**Instrument (CHR:BP)**	**Ref MAF**	**Beta**	**SE**	**P**	**P**	**N SNP**
cg09507712	11	47616693	*C1QTNF4*	ILMN_1671710	11	47321147	*MADD*	11:47606483	0.35	4.56	0.97	2.34E-06	2.20E-02	20
cg07409245	11	47616751	*C1QTNF4*	ILMN_1808122	11	47613713	*C1QTNF4*	11:47687147	0.35	-5.57	0.97	9.36E-09	4.50E-02	20
cg15575356	11	47616757	*C1QTNF4*	ILMN_1808122	11	47613713	*C1QTNF4*	11:47687147	0.35	-6.87	1.22	1.99E-08	4.40E-02	20
cg24977308	11	47636548	-	ILMN_1808122	11	47613713	*C1QTNF4*	11:47486885	0.35	7.92	1.44	3.46E-08	2.69E-01	20

5.5 Discussion

5.5.1 Overview of results

In this Chapter I conducted an assessment of the genetic architecture of DNAm in both whole blood and the cortex and identified genome-wide associations between common genetic variants and DNAm sites (mQTLs) utilising the Illumina EPIC array. I took forward the two mQTL databases generated in my analyses and characterised the relationship between neighbouring DNAm sites using a Bayesian colocalisation approach, identifying many occurrences where proximally located DNAm sites are genetically co-regulated with the same causal variant. This goes in line with previous research which identified that a high proportion of DNAm sites are influenced by shared genetic variants (Hannon *at al.*, 2018; Liu *at al.*, 2014). These results suggest that differentially methylated regions associated with traits identified by EWAS are potentially genetically mediated. The mQTLs were then used for SMR analyses, which identified multiple situations where SNPs are pleiotropically associated with LOAD; I found evidence that this relationship is mediated by either DNAm, gene expression or both. A number of pathways which have previously been implicated in AD were enriched for SMR prioritised genes including lipid and cholesterol metabolism (Di Paolo & Kim, 2011; Jones *at al.*, 2010; Penke *at al.*, 2018), Aβ (Kunkle *at al.*, 2019; Sadigh-Eteghad *at al.*, 2015), tau (Kosik, Joachim, & Selkoe, 1986; Kunkle *at al.*, 2019) and *APP* processing (Eggert, Thomas, Kins, & Hermey, 2018), neuronal, glial and oligodendrocyte cell processes (Nordengen *at al.*, 2019) and innate immune response (Cao & Zheng, 2018; Jones *at al.*, 2010; Kunkle *at al.*, 2019).

The whole blood and cortex mQTL databases showed high overlap, suggesting that QTL effects are generally conserved across tissues. This goes in line with previous studies demonstrating that a relatively high proportion of mQTLs co-vary across tissues (Hannon *at al.*, 2016; Smith *at al.*, 2014). These data further support the notion that whole blood may be a valid correlate of physiological processes in other tissues for genetically mediated sites. However, I found evidence for cortex-specific genetic effects at certain loci suggesting there is some heterogeneity between whole blood and cortex. Since LOAD is a disease of the brain this highlights the importance of utilising the relevant tissue where possible. Additionally, I found a significant enrichment of LOAD-associated GWAS variants in both whole blood and cortex mQTLs, indicating that common genetic variants conferring risk for LOAD are associated with variable DNAm in both tissues. These results further support the utility of applying mQTLs to refine GWAS signals and goes in line with existing literature suggesting LOAD GWAS variants act via mechanisms of gene regulation.

To refine the LOAD GWAS signals I considered multiple mechanisms of gene regulation and applied SMR, incorporating DNAm and gene expression data. The SMR analyses identified several situations where SNPs were pleiotropically associated with DNAm, gene expression and LOAD. The associations nominated sites which were in the vicinity (within 250kb) of GWAS SNPs. Several associations were identified in both the whole blood and cortex analyses along chromosome 11, which contains four established AD GWAS loci. In the whole blood analyses *MS4A4A* and *MS4A6A* were identified as having a pleiotropic association with LOAD mediated by DNAm and gene expression. Interestingly, a genome-wide-significant LOAD SNP (rs1582763; 11:60021948) has previously been associated with both *MS4A4A* and *MS4A6A* (Deming *at al.*, 2019). Evidence suggests that differential expression of *MS4A4A* and *MS4A6A* alter *TREM2* concentrations and this has been associated with increased levels of tau (Deming *at al.*, 2019). These results further support these two genes being the most appropriate annotation for the MSA4 region and good candidates for future functional studies. In the cortex analysis *C1QTNF4* was identified as having a pleiotropic association with LOAD mediated by DNAm and gene expression. *C1QTNF4* is located in the same LD block as *SPI1*, which is currently the gene annotated to a top LOAD GWAS loci. However, AD genetic studies have found eight independent variants in LD within the *SPI1* region which are also eQTLs for *C1QTNF4* (Rosenthal & Kamboh, 2014). *SPI1* may be acting in combination with (or serving as a proxy for) other genes around this region that mediate AD risk. Our results suggest *C1QTNF4* may be a functionally relevant gene in this region, however since not all genes in the region were tested in the gene expression analysis it may not be the only relevant gene. Overall, these SMR results show the utility of applying multi-omics methods to refine and fine-map GWAS signals to identify potential targets for future functional studies.

Several of the DNAm sites identified as having pleiotropic relationships with LOAD currently are not annotated to a specific gene as they are classified as intergenic. However, mounting evidence suggest proximity may not necessarily be the most relevant in terms of functional effects on transcription. For example, a number of DNAm sites were associated with the expression of genes which are not proximally located to these sites. In addition, not all DNAm sites were associated with gene expression, suggesting that differential methylation may not always be involved in transcription. These results go in line with research by Hannon *at al.* (2018) and Bonder *at al.* (2017), who also conducted analysis of DNAm on gene expression. They found similar patterns of results to the analysis in this Chapter, where there were both positive and negative associations between DNAm and gene expression and the sites with a DNAm-expression relationship were enriched in transcription start sites and enhancers. The SMR results could be applied to EWAS data to refine the annotations of DNAm sites and improve the interpretation of LOAD GWAS data.

5.5.2 Limitations

There are several limitations to consider within the work presented in this Chapter. First, the SMR & HEIDI analyses used to identify DNAm–LOAD, gene expression–LOAD and DNAm–gene expression relationships were conducted separately as there is currently no tool which enables all three to be tested simultaneously. The analyses were based on identifying consistent association signals at a particular locus. This is not the most optimal method and results in a loss of power due to the strict Bonferroni thresholding at each step. Hence, I relaxed the threshold of the gene expression–LOAD analyses, although this could potentially introduce false positive results into the analysis. In future, a tool which combines all three genomic data-types would be a more optimal strategy for identifying situations were a SNP is pleiotropically associated with both DNAm, gene expression and a trait. We also need to consider that the Wald ratio used to detect an effect (the influence of the SNP-outcome effect divided by the mQTL or eQTL effect) can help identify pleiotropic associations between two traits, but it does not provide information about the intermediate mechanisms involved in any potential ‘causal’ process. Additionally, there is a limitation of using proxy instrumental variables. If multiple mQTLs have the same significant threshold and LD relationships there is an element of randomness for which one is picked. Leading on from this we also need to take into consideration that SMR does not identify “causal” relationships as it cannot distinguish between two variants in high LD and this is why the relationships identified are referred to as “pleiotropic” associations. It must be noted that SMR and other MR tools are hypothesis generating; the results from SMR are not definitive and are not necessarily biologically relevant. However, they do provide a list of prioritised genes for follow-up functional studies.

Within the analyses in this Chapter I only considered *cis* mQTL effects. Although *cis* effects are easier to detect in the current sample sizes available for mQTLs (as they are generally larger than *trans* effects), studies suggest *trans* effects ultimately account for a higher proportion of variation (Gaunt *at al.*, 2016). In future, *trans* effects could also be considered as QTL datasets become larger.

The samples used in this Chapter are all of European ancestry, which reduces the population validity as these results do not necessarily translate across different ethnicities. Additionally, although I utilised the latest Illumina technology (the EPIC array) only a small proportion of the total number of DNAm sites across the genome were assayed and there is light coverage of certain regulatory features.

Since the analyses in the Chapter have been conducted, additional LOAD GWAS have been published (Bellenguez *at al.*, 2020; de Rojas *at al.*, 2020; Schwartzentruber *at al.*, 2021; Wightman *at al.*, 2020). My analyses should be conducted again using a more comprehensive dataset of variants associated with LOAD. This would increase the power and potentially lead to the identification of more pleiotropic SMR associations, advancing our understanding of the regulatory mechanisms involved in disease pathogenesis.

5.5.3 Conclusion

To my knowledge, I have generated the largest mQTL databases in both whole blood and the cortex. The results in this Chapter have further demonstrated that genetic variants have genome-wide effects on DNAm. Additionally, by integrating SNPs and DNAm with LOAD GWAS and gene expression I have been able to explore the mechanisms underlying disease, advancing our understanding of the interaction between gene regulation and expression, and enabling the prioritisation of candidate genes involved in disease aetiology.

6 Methylomic variation associated with polygenic risk for Alzheimer's disease

This Chapter includes a methods (see **6.3.4**) and results (see **6.4.4**) section taken directly from a peer-reviewed publication I co-authored which has been published in 'Brain Communications' (Hannon *at al.*, 2020). See **Appendix A** for the full published manuscript.

6.1 Introduction

Although genetic studies have identified risk factors for early onset familial Alzheimer's disease (fEOAD) including mutations in *APP*, *PSEN1* and *PSEN2* (which are all involved in amyloid-beta [Aβ] processing), the genetic aetiology of late onset sporadic Alzheimer's disease (LOAD) is less well defined. LOAD is a complex disease; risk is mediated by a combination of genetic, lifestyle and environmental factors (Sims, Hill, & Williams, 2020). The first gene implicated in LOAD was Apolipoprotein E (*APOE*), which is located on chromosome 19 and remains the strongest risk locus for LOAD. Following the discovery of *APOE*, genome wide association studies (GWAS) - which systematically identify common genetic variants associated with disease (see Chapter 1 section **1.2** for more details on GWAS) - have identified >75 SNPs associated with LOAD (Bellenguez *at al.*, 2020; de Rojas *at al.*, 2020; Jansen *at al.*, 2019; Kunkle *at al.*, 2019; Schwartzentruber *at al.*, 2021).

To further investigate the genetic architecture of polygenic traits including LOAD, polygenic risk scores (PRS) have been developed which combine the information provided by independent SNPs. PRS quantify genetic burden as an accumulative genetic score for each individual in a sample. They are calculated as a sum of trait-associated variants, weighted by effect sizes estimated from a discovery GWAS (see Chapter 1 section **1.2.2** for more details). There has been considerable success in applying PRS in LOAD to significantly predict disease status (Escott-Price *at al.*, 2019; Escott-Price, Myers, *at al.*, 2017; Escott-Price *at al.*, 2015). Although the variance explained by a PRS is generally low (~1-5%) (Escott-Price *at al.*, 2015), PRS are currently one of the best predictors for LOAD and have aided understanding about the contributors to variation in the disease. See Chapter 1 section **1.2.3** for more details on PRS in LOAD.

In addition to the robust associations between PRS and disease status, LOAD PRS have been shown to associated with several other AD related traits including MCI and cognitive decline (Felsky *at al.*, 2018; Ge *at al.*, 2018; Kauppi *at al.*, 2020; Marioni *at al.*, 2017; Mormino *at al.*, 2016); memory impairment (Marioni *at al.*, 2017; Mormino *at al.*, 2016); brain measurements such as hippocampal volume (Axelrud *at al.*, 2018) and cortical thickness (Sabuncu *at al.*, 2012); cerebrospinal biomarkers (Martiskainen *at al.*, 2015); inflammatory biomarkers (Morgan *at al.*, 2017); and neuropathological measures (Desikan *at al.*, 2017; Felsky *at al.*, 2018; Hannon *at al.*, 2020). The extensiveness of the associations with LOAD PRS highlights the complexity of understanding the genetic pathways involved in LOAD pathogenesis.

The integration of genetic and DNA methylation (DNAm) data could help us better understand the molecular mechanisms involved in AD, particularly due to the high heritability of LOAD (h^2=56-79% (Gatz *at al.*, 2006)) and research showing how the epigenome is directly influenced by genetic variation (Hannon *at al.*, 2016; Hannon, Gorrie-Stone, *at al.*, 2018b; Hannon, Schendel, *at al.*, 2018). It is possible to look at the effect of genome-wide genetic risk on regulatory processes in order to identify the molecular consequences of elevated risk burden. Recent studies have explored the effects of variable PRS on genome-wide DNAm to explore the molecular genomic mechanisms involved in disease pathogenesis (Hannon *at al.*, 2016; Hannon, Schendel, *at al.*, 2018; Viana *at al.*, 2017). PRS-associated epigenetic variation is potentially less affected by factors associated with the disease itself (such as medication and other confounders) making it an ideal experimental design to explore disease phenotypes. Recent research by Hannon *at al.* used this approach to identify methylomic variation associated with a PRS for schizophrenia (Hannon *at al.*, 2016); they conducted an EWAS of SCZ PRS against genome-wide DNAm identifying multiple differentially methylation position (DMPs) associated with elevated genetic burden (Hannon *at al.*, 2016). PRS EWAS relationships in brain disorders have also been investigated in peripheral tissues. Hannon, Schendel and colleagues (2018) conducted an autism spectrum disorder (ASD) – PRS study whereby they quantified neonatal methylomic variation from archived blood spots in 1263 infants (50% later developed ASD) and identified multiple sites where the ASD PRS was associated with variable DNAm. PRS have proven a useful tool for associating the genetic contribution of a disease - essentially as a biomarker - with epigenetic variation, enabling the exploration of molecular genomic mechanisms driving disease pathogenesis.

In this chapter I aimed to identify how known genetic risk factors for Alzheimer's disease (AD) influence the development of different aspects of neuropathology and how methylomic variation in the cortex and blood associates with polygenic risk burden for LOAD.

6.2 Chapter aims

1. To identify if AD PRS significantly predicts disease status in BDR, using neuropathologically defined disease status.
2. To explore how known genetic risk factors for AD influence the development of different aspects of neuropathology in the BDR cohort.
3. To identify if methylomic variation in whole blood is associated with polygenic burden for LOAD.
4. To identify if methylomic variation in the cortex is associated with polygenic burden for LOAD.
5. To identify differences and similarities between the methylomic signatures of elevated polygenic burden for LOAD in whole blood and cortex.

6.3 Methods

6.3.1 Whole blood data

For the whole blood datasets, I utilised matched DNAm and SNP data from eight studies, totalling 6,106 unrelated European samples (56% male; age range=18-98, mean age=52.2). A breakdown of the cohorts included in my analyses is shown in **Table 6.1**. Several of these cohorts have been described in detail elsewhere, either within this thesis or in published papers (see 'Reference' column of **Table 6.1**). Briefly, the 'Aberdeen' case-control schizophrenia cohort contains individuals who were born in the British Isles (95% in Scotland) (International Schizophrenia Consortium, 2008; Stefansson *at al.*, 2008). The 'Dublin' case-control schizophrenia cohort was selected from the Irish Schizophrenia Genomics consortium (Morris *at al.*, 2014). The 'European Network of National Schizophrenia Networks Studying Gene-Environment Interactions (EU-GEI) cohort was set up to test hypotheses about variations in incidence within and between countries and the role of multiple environmental and genetic risk factors, and their interactions, in the development of psychotic disorders (Jongsma *at al.*, 2018). The 'Exeter 10,000 (EXTEND)' study is a research biobank funded by the National Institute for Health Research (NIHR) and is a population study of >10,000 individuals >18 years of age who live within 25 miles of Exeter (Devon, UK; https://exetercrfnihr.org/about/exeter-10000/); for more details on the genotyping and DNAm data see Chapter 5 section **5.3.1.3.2**. The 'Kings College London (KCL) 1 and 2' cohorts are part of ProjectMinE (Project MinE ALS Sequencing Consortium, 2018) which is a collaboration of international groups with the aim of collecting 22,500 DNA profiles to investigate rare and common genetic and epigenetic variation contributing to the development of Amyotrophic Lateral Sclerosis (ALS) and the data in this chapter consisted of a subset of individuals of UK nationality from the UK National DNA Bank for MND Research who were put forward for DNAm profiling. The University College London (UCL)' case-control schizophrenia cohort is comprised of unrelated ancestrally matched cases and controls from the UK (Datta *at al.*, 2010; International Schizophrenia Consortium, 2008). The 'UK Household Longitudinal Study (UKHLS)' was established in 2009 and is a longitudinal panel survey of 40,000 UK households from England, Scotland, Wales and Northern Ireland (Buck & McFall, 2011). Although several of these cohorts are either schizophrenia (SCZ) or ALS case-control cohorts, none has a phenotypic focus on LOAD. Of note, AD is not genetically correlated with

SCZ after correcting for multiple testing (r_g=0.103; p=0.042) (Jansen *at al.*, 2019) but is modestly correlated with ALS when using the most recent ALS GWAS (r_g = 0.31; p=9.6e-03) (van Rheenen *at al.*, 2021).

6.3.2 Human cortex dataset

For the cortex dataset I utilised the matched DNAm and SNP data from the Brains for Dementia Research (BDR) cohort (see Chapter 4 section **3.2.2** and Chapter 5 section **5.3.2.2**), totalling 1,047 European unrelated samples (53% male; age range=41-104, mean age=83.37; see **Table 6.2**). Briefly, the BDR cohort was established with the aim of generating a large comprehensive neuropathological dataset from multiple brain banks using standardised procedures to enable the investigation of dementia through detailed phenotypic and multi-omics datasets (Francis, Costello, & Hayes, 2018). For more details on this cohort see Chapter 4 section **4.3.1.**

Table 6.1: Sample characteristics of the whole blood cohorts included in my EWAS meta-analysis of variable DNAm associated with AD PRS.

Cohort	N	Male	Female	Age (years)			Illumina DNAm array	Genotyping Array(s)	References
				Range	Median	Mean (SD)			
Aberdeen	797	568	229	18.3-80.7	45.6	44.63 (12.93)	450K	HumanHap300 HumanHap550 Affymetrix GeneChip	(International Schizophrenia Consortium, 2008; Stefansson *at al.*, 2008)
Dublin	634	454	180	17-70.9	58	56.39 (11.93)	450K	Affymetrix 6.0 Illumina 1.2M-Duo	(Morris *at al.*, 2014)
EuGEI	634	346	288	18-64	45.6	36.39 (12.92)	EPIC	Illumina HumanCoreExome-24 BeadChip	(Jongsma *at al.*, 2018)
EXTEND	983	467	516	19-80	58	56.42 (11.72)	EPIC	Illumina Infinium Global Screening Array	Chapter 5 **5.3.1**
KCL1	698	398	300	24-91	63	62.34 (11.33)	450K	Illumina OmniExpress	(Project MinE ALS Sequencing Consortium, 2018; van Rheenen *at al.*, 2016)
KCL2	612	341	271	27-87	63	62.47 (11.30)	450K	Illumina OmniExpress	Project MinE ALS Sequencing Consortium, 2018; van Rheenen *at al.*, 2016)
UCL	649	383	266	18-90	37	40.43 (15.10)	450K	Affymetrix Genome Wide Human SNP Array 5.0 Affymetrix 500K	(Datta *at al.*, 2010; International Schizophrenia Consortium, 2008)
UKHLS	1099	455	644	28-98	59	58.5 (14.74)	EPIC	Illumina Infinium HumanCoreExome BeadChip	(Hannon *at al.*, 2018)
TOTAL	**6106**	**3412**	**2695**	**18-98**	**58**	**52.2 (12.75)**	**-**	**-**	**-**

Table 6.2: Sample characteristics of the subset of the BDR cohort included in the polygenic risk score EWAS analysis.

Cohort	**Donors**	**Samples**			**M**	**F**	**Age (years)**			**Illumina DNAm array**	**Genotyping Array**	**Reference**
		Total	**OCC**	**PFC**			**Range**	**Median**	**Mean (SD)**			
BDR	541	1047	525	522	493	554	41-104	84	83.37 (9.03)	EPIC	Illumina NeuroChip	(Shireby *at al.*, 2020)

Table 6.3: APOE genotypes for the BDR cohort. Number of alleles = the number of ε2 or ε4 alleles carried by an individual.

Phenotype	**Number of alleles**	**Occipital**	**Prefrontal**	**Total Samples**	**Donors**
APOE-ε2	0	536	534	1070	553
	1&2	61	62	123	65
Total	**-**	**597**	**596**	**1193**	**618**
APOE-ε4	0	279	277	556	285
	1	275	278	553	290
	2	43	41	84	43
Total	**-**	**597**	**596**	**1193**	**618**

6.3.2.1 DNA methylation data pre-processing

Unless otherwise reported, all statistical analysis was conducted in the R statistical environment (version 3.5.2; https://www.r-project.org/). Raw data for all datasets were used, prior to any QC or normalisation, and processed using the *wateRmelon* (Pidsley *at al.*, 2013) and *bigmelon* (Gorrie-Stone *at al.*, 2019) packages. The stringent DNAm QC pipeline described in Chapter 3 section **3.2.4** was used on each dataset. Smoking score was estimated in the blood cohorts using the algorithm which is based on the DNAm profile at 183 sites known to be associated with smoking (Elliott *at al.*, 2014). For more details on the QC pipeline see Chapter 2 section **2.1.4.**

6.3.3 Genotyping and imputation

Genotype QC and imputation was run individually on each of the datasets using the same pipeline as described in Chapter 5 section **5.3.1.3**. For more details on the genotyping QC see Chapter 2 section **2.2.1**.

6.3.3.1 *APOE* genotyping

In order to determine *APOE* status in the BDR cohort, samples were genotyped for *APOE* ε2, ε3 and ε4 alleles using the TaqMan assay for SNPs rs7412 and rs429358 (Applied Biosystems). The genotype call rate was 99.7% (Brookes *at al.*, 2018). *APOE* status was modelled as two numeric variables counting the number ε2 and ε4 alleles each individual had. ε2 is rare and only 3 donors had an ε2/ε2 genotype, therefore the ε2/ε2 individuals were combined with the individuals with one ε2 allele (for total samples in each group see **Table 6.3**).

6.3.3.2 Polygenic risk scores

Polygenic risk scores were calculated using an additive method (i.e. by summing up trait associated variants, weighted by effects sizes estimated from GWAS) for each individual in the whole blood and cortex datasets. GWAS results from two of the latest LOAD GWAS with publicly available data were used to calculate the LOAD PRS for each individual: the Kunkle *at al.* LOAD GWAS (Kunkle *at al.*, 2019) which is based on clinically defined cases compared to controls; and the Jansen *at al.* LOAD GWAS (Jansen *at al.*, 2019) which is based on clinically defined cases and controls and proxy AD based on family history (see Chapter 1 section **1.2** for more details). To distinguish the effects of *APOE* from other genetic variants associated with AD I ran all PRS analyses first including, and second excluding, variants in the *APOE* region (chr19:45,116,911–46,318,605) (Kunkle *at al.*, 2019). I generated PRS using *PRSice* (v2.0) (Choi & O'Reilly, 2019) which 'clumps' the AD GWAS summary statistics such that the most significant variant in each LD block is retained. The PRS was calculated in each dataset for each individual, as the number of reference alleles multiplied by the log odds ratio for that SNP (taken from the Kunkle *at al.* and Jansen *at al.* AD GWAS), and then summed across all retained clumped variants with a GWAS p-value < p-value threshold (P_T). A range of P_Ts (p< 5e-08, 1e-05, 1e-04, 1.5e-04, 2e-04, 5e-04, 0.001, 0.05, 0.09, 0.1, 0.2, 0.4, 0.5) were used initially in the BDR cohort to generate multiple possible PRS, and the optimal PRS was selected as the score that explained the highest proportion of variance (Nagelkerke's pseudo R^2) in AD case control status. In this analysis, AD cases and controls were defined as Braak high (Braak NFT stages V-VI) and low (Braak NFT stages 0-II) respectively, and PRS was tested using a logistic regression model with the first eight genetic principal components as covariates. In the BDR cohort the optimal threshold for selecting SNPs for the PRS was p< 5e-8 both when excluding *APOE* from the PRS (see **Figure 6.2** and **Figure 6.3** in **Results section 6.4.1**) and when including *APOE* in the PRS (see **Figure 6.4** and **Figure 6.5** in **Results section 6.4.1**) for both the Kunkle *at al.* (2019) and Jansen *at al.* (2019) GWAS. Therefore, this threshold was used in all analyses to generate PRS values for use in subsequent EWAS analyses. Prior to analysis the PRS calculated at this threshold was standardised to have a mean of 0 and SD of 1, and therefore the interpretation is in units of SDs.

6.3.4 Genetic analysis of neuropathology

Genetic associations between either *APOE* status or Alzheimer's disease PRS (generated using the Kunkle *at al.* GWAS excluding the *APOE* region) and the continuous neuropathology variables (Braak NFT stage, Thal Aβ stage, CERAD stage, Braak Lewy body [LB] stage) and the Clinical Dementia Rating (CDR global rating) were tested using a linear regression model. TDP-43 proteinopathy (a binary variable) was analysed with logistic regression, but the model framework was the same. Up to four regression models were fitted for each variable. First, the effects of *APOE* status and Alzheimer's disease PRS were estimated separately using Model 1 and Model 2 below.

Model 1:

$$variable \sim APOE\varepsilon 2 + APOE\varepsilon 4 + covariates + genetic\ PCs1 - 8$$

Model 2:

$$variable\ \sim\ PRS + covariates + genetic\ PCs1 - 8.$$

If *APOE* (either variable) and PRS were significantly associated with an outcome, then a multiple regression analysis was additionally fitted testing *APOE* and PRS simultaneously to confirm these were independent associations (Model 3).

Model 3:

$$variable\ \sim\ APOE\varepsilon 2 + APOE\varepsilon 4 + PRS + covariates + genetic\ PCs1 - 8$$

Finally, an interaction model (Model 4) between *APOE* and PRS was fitted to test if PRS associations differed depending on *APOE* genotype.

Model 4:

$$variable\ \sim\ APOE\varepsilon 2 + APOE\varepsilon 4 + PRS + APOE\varepsilon 2 * PRS + APOE\varepsilon 4 * PRS +\ covariates \\ + genetic\ PCs1 - 8$$

All analyses included age at death, sex and BDR centre as covariates and the first eight genetic principal components.

6.3.5 Regression models against LOAD polygenic risk

I looked at associations between the genetic loading for LOAD (quantified as PRS; see section **6.3.3.2**) and variable DNAm in both whole blood and cortex tissues.

In whole blood I conduced EWAS for each of the four LOAD PRS (Kunkle *at al.* GWAS, Jansen et al GWAS, with and without *APOE*) using linear regressions controlling for age, sex, batch (96 well plate on which the sample was run), smoking score and derived cell proportions. Equations for the linear regression models are shown below.

EWAS of LOAD PRS, where i = DNAm probe and PRS = Kunkle *at al.* LOAD PRS including *APOE*, Kunkle *at al.* LOAD PRS excluding *APOE*, Jansen *at al.* LOAD PRS including *APOE* and Jansen *at al.* LOAD PRS excluding *APOE*.

$$DNA\ methylation[i]\sim PRS + age + sex + batch + smoking\ score + \ cell\ proportions$$

In the cortex I conducted and EWAS for LOAD PRS (for both GWAS and including and excluding the *APOE* region) using mixed effect regression models where age, sex, batch, PC1 and derived cell proportions were included as fixed effects and individual was included as a random effect. To identify the p-value I used an ANOVA comparing the full model including the PRS to a null model in which this measure was excluded. Equations for the mixed effect regression models are shown below.

EWAS of LOAD PRS full model, where i = DNAm probe and PRS = LOAD PRS including *APOE*, LOAD PRS excluding *APOE*.

$$DNA\ methylation[i]\sim PRS + age + sex + batch + \ cell\ proportions + PC1\ + (1|individual)$$

EWAS of LOAD PRS null model, where i = DNAm probe.

$$DNA\ methylation[i]\sim age + sex + batch + \ cell\ proportions + PC1\ + (1|individual)$$

6.3.6 Meta-analysis of whole blood LOAD PRS EWAS

The whole blood PRS EWAS results from individual cohorts were subsequently meta-analysed using an inverse variance weighted (IVW) method which summarises effect sizes from multiple independent studies by calculating the weighted mean of the effect sizes using the inverse of the variance of each study as weights. Probes were limited to those present in at least two of the cohorts (n= 484,581 DNAm probes) and the P value was Bonferroni corrected to control for this number of sites (p<0.05/484,581=1.03e-07). The inverse variance is roughly proportional to sample size and therefore more weight is generally given to larger studies. I used *plink1.9* (Chang *at al.*, 2015) to run the meta-analysis applying the --meta-analysis flag. I report the results from the random effects and fixed effects model in the tables. However, I focus on the random effect in the interpretation of results due to the number of cohorts included in the meta-analysis being more than three, meaning accurate heterogeneity statistics can be estimated. In addition, this model permits the true effect size to differ between cohorts.

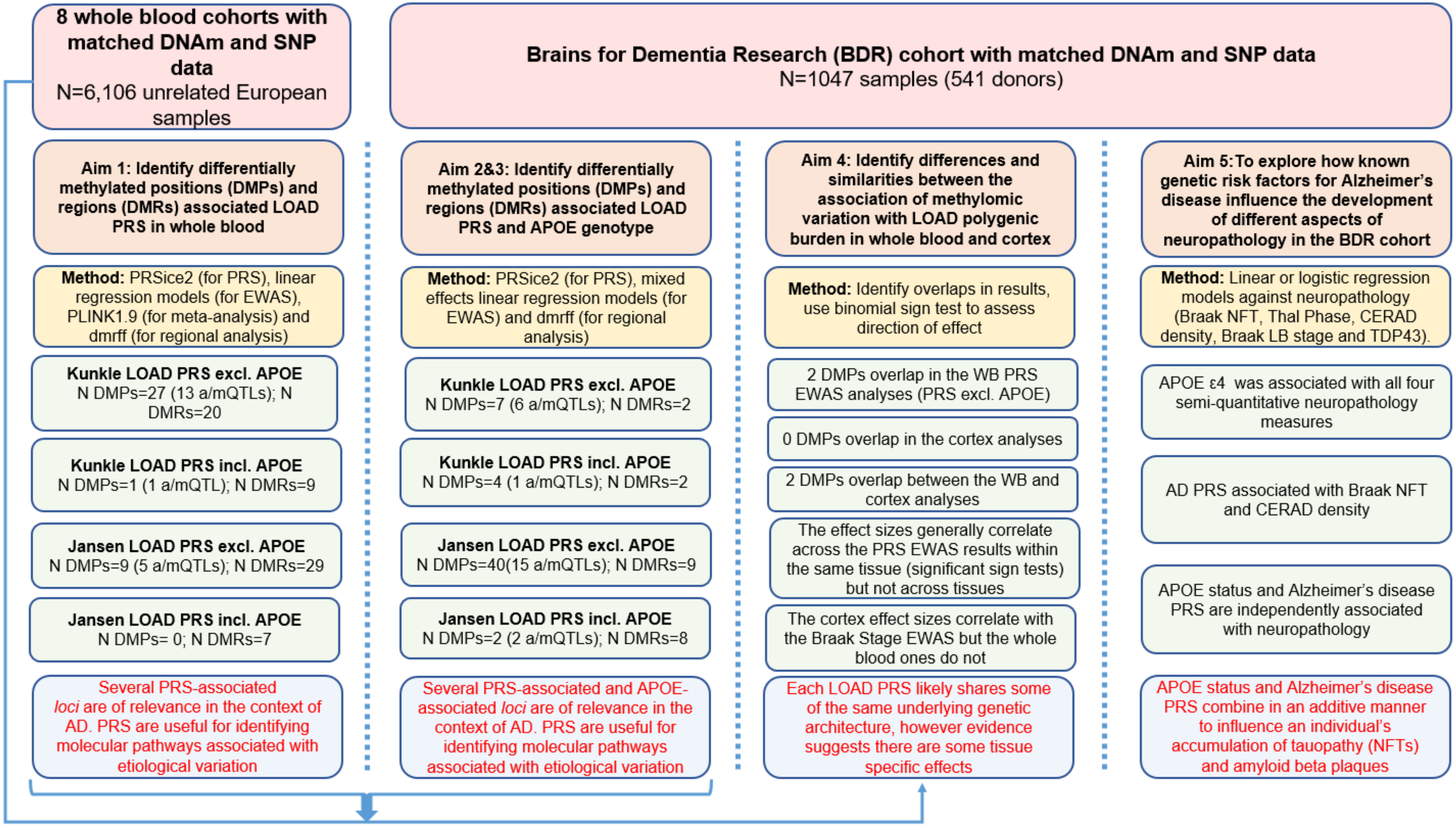
6.3.7 Identifying differentially methylated regions

In order to identify differentially methylated regions (DMRs) I used *dmrff* (Suderman *at al.*, 2018). *Dmrff* identifies regions by combining summary statistics from proximally located DNAm sites. *Dmrff* identifies candidate regions (regions containing ≥3 probes) as sequences of DNAm sites with EWAS values that reach a certain P_t. I used the input threshold p<0.05. For more details on dmrff see Chapter 2 section **2.3.3.** I applied *dmrff* to each of the EWAS run above (to each of the meta-analysed whole blood EWAS and each cortex EWAS) whereby I input the beta estimates, p-values estimated in each of the EWAS as well as the normalised DNAm data matrix.

6.3.8 Comparing results between and across tissues and with recent Alzheimer's disease EWAS

To identify differences and similarities across the results for each PRS and tissue, I first explored if there was any overlap in the significant DMPs identified in each EWAS. Then within each tissue and across tissues I compared the effect sizes of the DMPs reaching suggestive significance (p<5e-05) to those for the same DNAm sites in other EWAS analyses. To calculate the strength of the correlation I used Pearson's correlation and to determine if there was enrichment for the same direction of effect, I used a binomial sign test. Next, I compared results from the PRS EWAS analyses to those from two recent EWAS of AD pathology. First, I compared against the results from **Chapter 4** where I conducted a Braak NFT Stage EWAS in the BDR cohort (see section **4.4.5**). I then compared results to a recent whole blood AD EWAS conducted by Nabais and colleagues (Nabais *at al.*, 2021) using three AD cohorts: The Australian Imaging, Biomarkers and Lifestyle (AIBL) cohort (Ellis *at al.*, 2009), The Alzheimer's Disease Neuroimaging Initiative (ADNI) (Vasanthakumar *at al.*, 2020) and AddNeuroMed - The European Collaboration for the Discovery of Novel Biomarkers for Alzheimer's Disease (Lovestone *at al.*, 2009) .

6.4 Results



8 whole blood cohorts with matched DNAm and SNP data
N=6,106 unrelated European samples

Brains for Dementia Research (BDR) cohort with matched DNAm and SNP data
N=1047 samples (541 donors)

Aim 1: Identify differentially methylated positions (DMPs) and regions (DMRs) associated LOAD PRS in whole blood

Method: PRSice2 (for PRS), linear regression models (for EWAS), PLINK1.9 (for meta-analysis) and dmrff (for regional analysis)

Kunkle LOAD PRS excl. APOE
N DMPs=27 (13 a/mQTLs); N DMRs=20

Kunkle LOAD PRS incl. APOE
N DMPs=1 (1 a/mQTL); N DMRs=9

Jansen LOAD PRS excl. APOE
N DMPs=9 (5 a/mQTLs); N DMRs=29

Jansen LOAD PRS incl. APOE
N DMPs= 0; N DMRs=7

Several PRS-associated *loci* are of relevance in the context of AD. PRS are useful for identifying molecular pathways associated with etiological variation

Aim 2&3: Identify differentially methylated positions (DMPs) and regions (DMRs) associated LOAD PRS and APOE genotype

Method: PRSice2 (for PRS), mixed effects linear regression models (for EWAS) and dmrff (for regional analysis)

Kunkle LOAD PRS excl. APOE
N DMPs=7 (6 a/mQTLs); N DMRs=2

Kunkle LOAD PRS incl. APOE
N DMPs=4 (1 a/mQTLs); N DMRs=2

Jansen LOAD PRS excl. APOE
N DMPs=40(15 a/mQTLs); N DMRs=9

Jansen LOAD PRS incl. APOE
N DMPs=2 (2 a/mQTLs); N DMRs=8

Several PRS-associated and APOE-associated *loci* are of relevance in the context of AD. PRS are useful for identifying molecular pathways associated with etiological variation

Aim 4: Identify differences and similarities between the association of methylomic variation with LOAD polygenic burden in whole blood and cortex

Method: Identify overlaps in results, use binomial sign test to assess direction of effect

2 DMPs overlap in the WB PRS EWAS analyses (PRS excl. APOE)

0 DMPs overlap in the cortex analyses

2 DMPs overlap between the WB and cortex analyses

The effect sizes generally correlate across the PRS EWAS results within the same tissue (significant sign tests) but not across tissues

The cortex effect sizes correlate with the Braak Stage EWAS but the whole blood ones do not

Each LOAD PRS likely shares some of the same underlying genetic architecture, however evidence suggests there are some tissue specific effects

Aim 5: To explore how known genetic risk factors for Alzheimer's disease influence the development of different aspects of neuropathology in the BDR cohort

Method: Linear or logistic regression models against neuropathology (Braak NFT, Thal Phase, CERAD density, Braak LB stage and TDP43).

APOE ε4 was associated with all four semi-quantitative neuropathology measures

AD PRS associated with Braak NFT and CERAD density

APOE status and Alzheimer's disease PRS are independently associated with neuropathology

APOE status and Alzheimer's disease PRS combine in an additive manner to influence an individual's accumulation of tauopathy (NFTs) and amyloid beta plaques

Figure 6.1: Overview of experimental design and summary of results for Chapter 6: Methylomic variation associated with polygenic risk for Alzheimer's disease. *PRS = polygenic risk scores. LOAD = late onset Alzheimer's disease. a/mQTLs = associated with a methylation quantitative trait loci (mQTL) which is included in the PRS. WB = whole blood.*

6.4.1 PRS derived at genome-wide significance is the optimal threshold

In order to derive the optimal p-value threshold (i.e. the threshold which explains the most variation in the phenotype of interest) I generated PRS for each individual in the BDR cohort at several significance thresholds (p< 5e-08, 1e-05, 1e-04, 1.5e-04, 2e-04, 5e-04, 0.001, 0.05, 0.09, 0.1, 0.2, 0.4, 0.5) using *PRSice*(v2.0)(Choi & O'Reilly, 2019) incorporating both the Kunkle et al and Jansen at all summary statistics. Each PRS was tested against AD status using a logistic regression model. The most significant threshold for each of the four PRS generated (LOAD PRS Kunkle *at al.* excluding *APOE* [PRS-Kunkle]; LOAD PRS Kunkle *at al.* including *APOE* [PRS-Kunkle$_{APOE}$]; LOAD PRS Jansen *at al.* excluding *APOE* [PRS-Jansen]; and LOAD PRS Jansen *at al.* including *APOE* [PRS-Jansen$_{APOE}$]) was p=5e-08 (see **Figure 6.2; Figure 6.3; Figure 6.4; and Figure 6.5**). Of note, the variance explained for both the PRS-Kunkle and PRS-Jansen was around 6% (see **Figure 6.2** and **Figure 6.4**), which is fourfold smaller than the PRS including *APOE* (~22% for PRS-Kunkle$_{APOE}$ and PRS-Jansen$_{APOE}$; see **Figure 6.3 and Figure 6.5**). Both the PRS-Jansen and PRS-Jansen$_{APOE}$ did not significantly predict AD status or explain a substantial amount of variation in AD at p-value thresholds of p>0.001, suggesting that including genetic variants with p>0.001 introduces a lot of noise into the Jansen derived PRS. By including many variants which are unlikely to have true associations with AD, there is reduced power to detect an association with disease status.

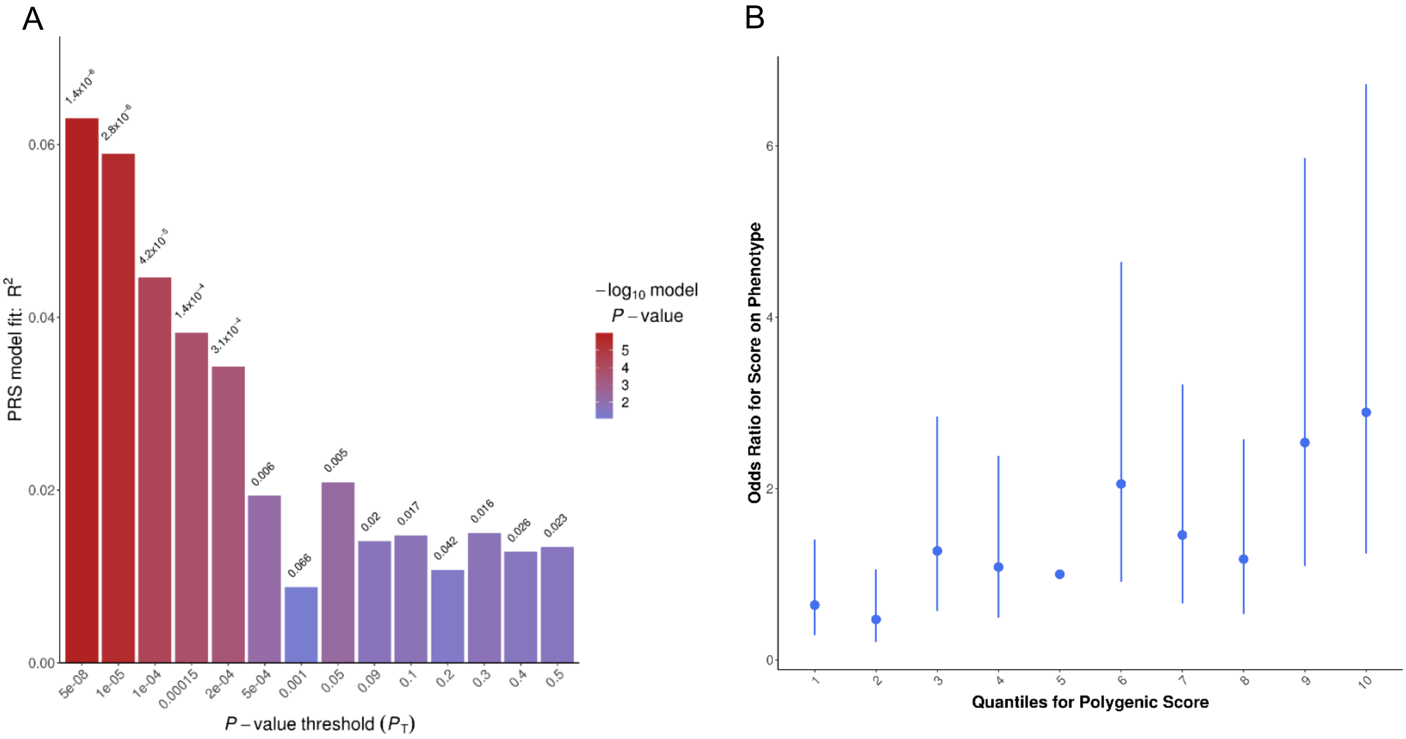


Figure 6.2: The PRS-Kunkle significantly and optimally predicts Alzheimer's disease status when including genome wide significant (p<5e-08) GWAS variants. *PRS-Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle et al. GWAS excluding the APOE region.* ***(A)*** *Bar plot from PRSice2 showing results at broad P-value thresholds for the PRS-Kunkle predicting LOAD status. The y-axis indicates the proportion of variance explained and the colour of the bar indicates the P-value of the association, which is also provided above each bar. The most significant P-value threshold is 5e-08 at which the PRS explains 6.3% of the variation in LOAD and is significantly associated with AD status (p=1.4e-06)* ***(B)*** *PRS-Kunkle deciles against odds ratio for developing LOAD. The X-axis shows the range of different deciles for PRS for individuals in the BDR cohort where PRS were generated using the most predictive threshold (p < 5e-08). The Y-axis shows the odds ratio for developing LOAD when comparing PRS from different deciles with the reference deciles (quantile 5 = average PRS in sample) with the bars corresponding to 95% confidence intervals of the odds ratio.*

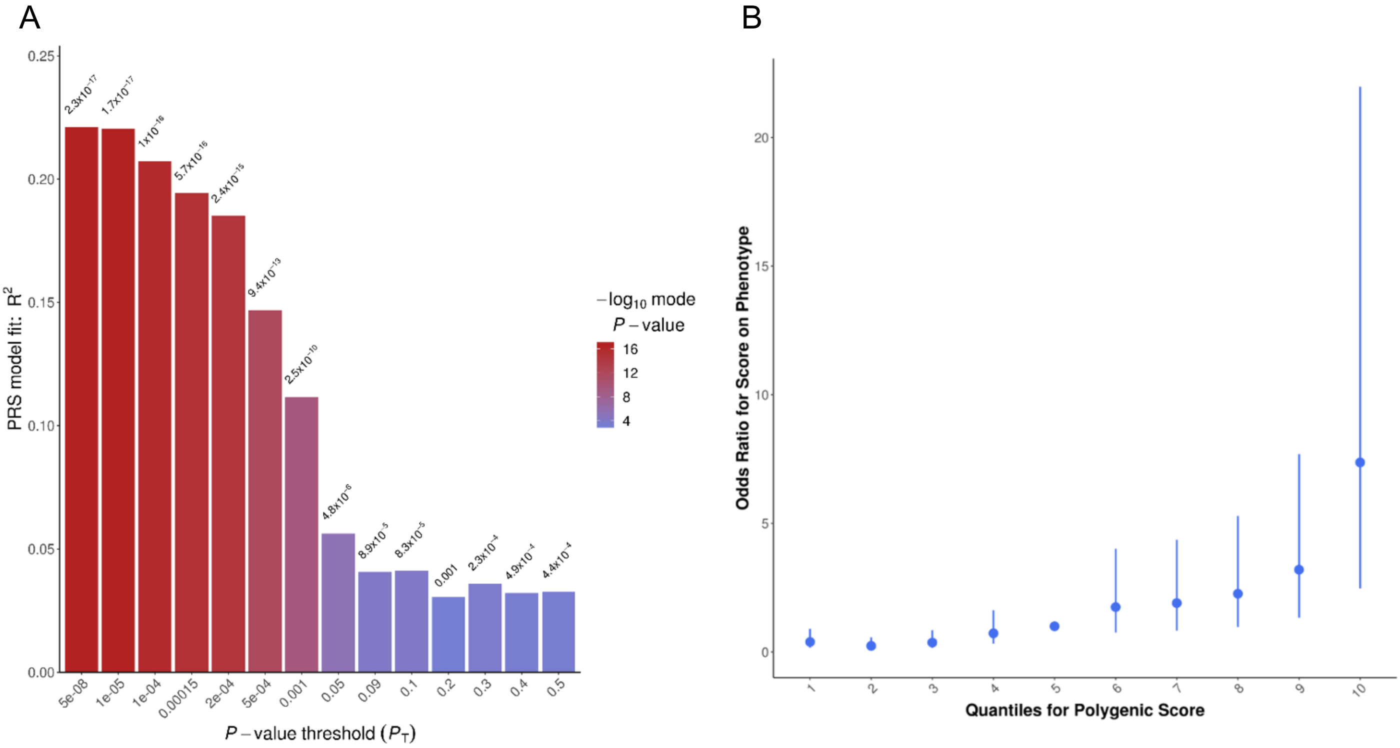


Figure 6.3: The PRS-Kunkle$_{APOE}$ significantly and optimally predicts Alzheimer's disease status when including genome wide significant (p<5e-08) GWAS variants. *PRS-Kunkle$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle et al. GWAS including the APOE region.* ***(A)*** *Bar plot from PRSice2 showing results at broad P-value thresholds for the PRS-Kunkle predicting LOAD status. The y-axis indicates the proportion of variance explained and the colour of the bar indicates the P-value of the association, which is also provided above each bar. The most significant P-value threshold is 5e-08 at which the PRS explains 22.1% of the variation in LOAD and is significantly associated with AD status (p=2.3e-17)* ***(B)*** *PRS-Kunkle deciles against odds ratio for developing LOAD. The X-axis shows the range of different deciles for PRS for individuals in the BDR cohort where PRS were generated using the most predictive threshold (p < 5e-08). The Y-axis shows the odds ratio for developing LOAD when comparing PRS from different deciles with the reference deciles (quantile 5 = average PRS in sample) with the bars corresponding to 95% confidence intervals of the odds ratio.*

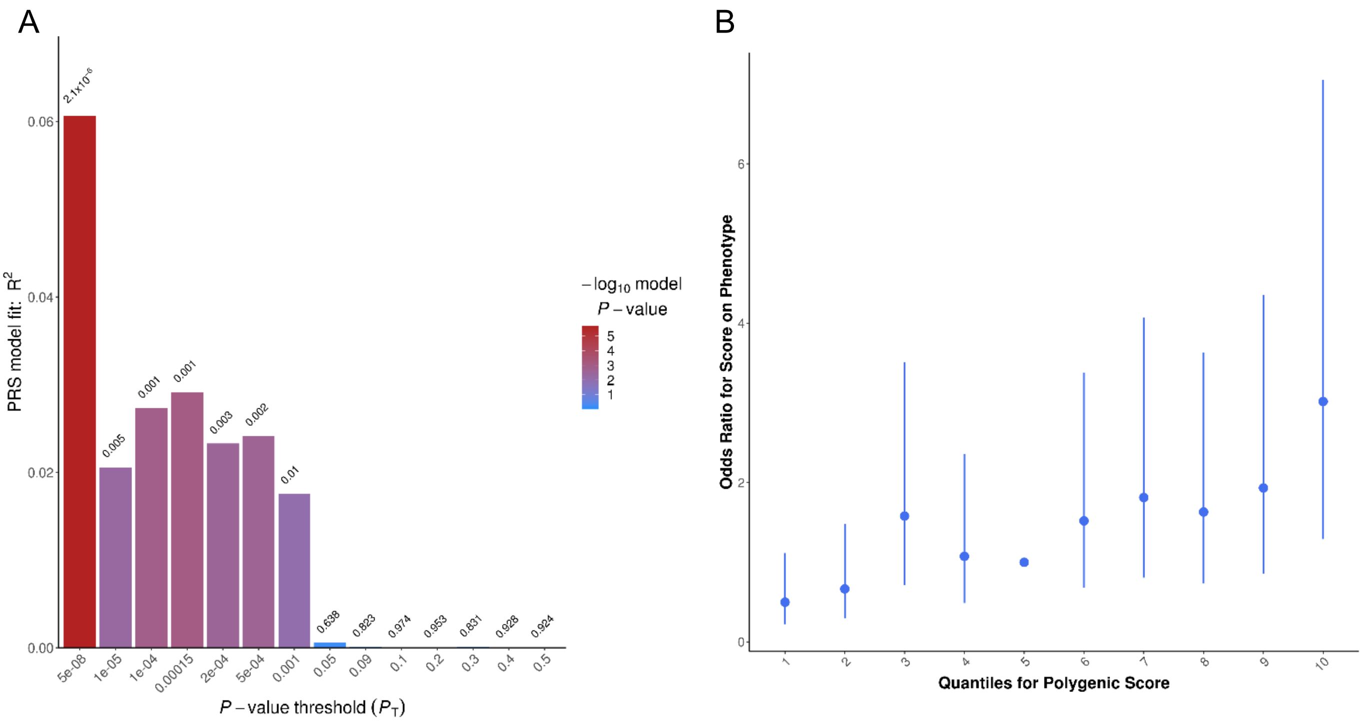


Figure 6.4: The PRS-Jansen significantly and optimally predicts Alzheimer's disease status when including genome wide significant (p<5e-08) GWAS variants. *PRS-Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen et al. GWAS excluding the APOE region.* ***(A)*** *Bar plot from PRSice2 showing results at broad P-value thresholds for the PRS-Jansen predicting LOAD status. The y-axis indicates the proportion of variance explained and the colour of the bar indicates the P-value of the association, which is also provided above each bar. The most significant P-value threshold is 5e-08 which explains 6.1% of the variation in developing load and is significantly associated with AD status (p=2.1e-06)* ***(B)*** *PRS-Jansen deciles against odds ratio for developing LOAD. The X-axis shows the range of different deciles for PRS for individuals in the BDR cohort where PRS were generated using the most predictive threshold (5e-08). The Y-axis shows the odds ratio for developing LOAD when comparing PRS from different deciles with the reference deciles (quantile 5 = average PRS in sample) with the bars corresponding to 95% confidence intervals of the odds ratio.*

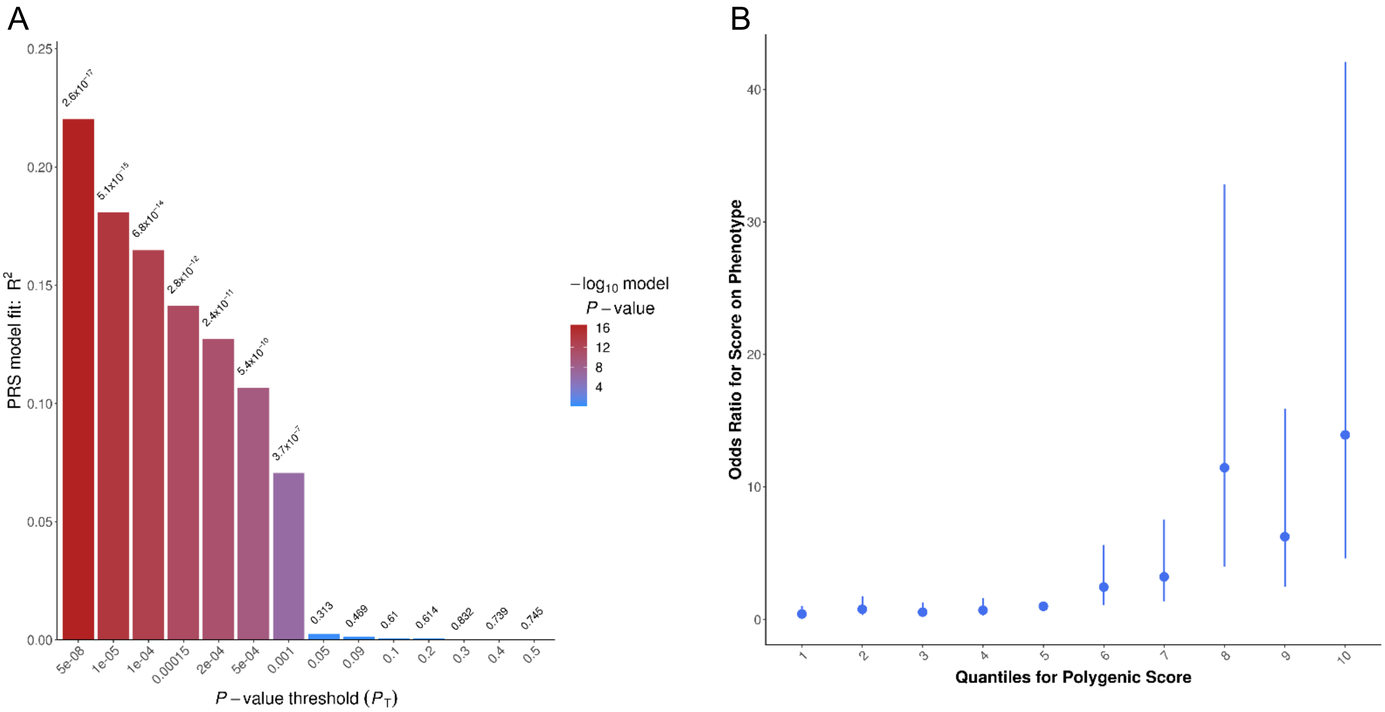


Figure 6.5: The PRS-Jansen*APOE *significantly and optimally predicts Alzheimer's disease status when including genome wide significant (p<5e-08) GWAS variants. *PRS-Jansen*APOE *= Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen et al. GWAS including the APOE region* ***(A)*** *Bar plot from PRSice2 showing results at broad P-value thresholds for the PRS-Jansen*APOE *predicting LOAD status. The y-axis indicates the proportion of variance explained and the colour of the bar indicates the P-value of the association, which is also provided above each bar. The most significant P-value threshold is 5e-08 which explains 22% of the variation in developing load and is significantly associated with AD status (p=2.6e-17)* ***(B)*** *PRS-Jansen*APOE *deciles against odds ratio for developing LOAD. The X-axis shows the range of different deciles for PRS for individuals in the BDR cohort where PRS were generated using the most predictive threshold (5e-08). The Y-axis shows the odds ratio for developing LOAD when comparing PRS from different deciles with the reference deciles (quantile 5 = average PRS in sample) with the bars corresponding to 95% confidence intervals of the odds ratio.*

6.4.2 PRS is elevated in LOAD cases compared to low pathology controls

There were significant differences between cases and controls for each of the four PRS, with cases having a higher PRS compared to controls (see **Figure 6.6**). Both the PRS-KunkleAPOE (p=2.3e-17) and the PRS-JansenAPOE (p=2.6e-17) had a large difference between disease status, with cases having on average a PRS 1 SD higher than controls; this is nearly double the magnitude of effect compared to the PRS where *APOE* was excluded. The PRS-Kunkle was characterised by marginally stronger differences between cases and controls (beta=0.32; p=1.4e-06) than the PRS-Jansen (beta=0.51; p=2.1e-06) (see **Figure 6.6**). These results are as we would expect; although previous studies have consistently shown that PRS can significantly differentiate between cases and controls of disease phenotypes including AD (Escott-

Price *at al.*, 2019; Escott-Price, Myers, *at al.*, 2017; Escott-Price *at al.*, 2015), generally these differences are small and are close to the population mean (see **Figure 6.7**). Of note, the larger differences between cases and controls when *APOE* was included in the PRS probably reflects the very strong effect of this gene compared to any of the other AD risk variants (Strittmatter *at al.*, 1993; van der Lee *at al.*, 2018).

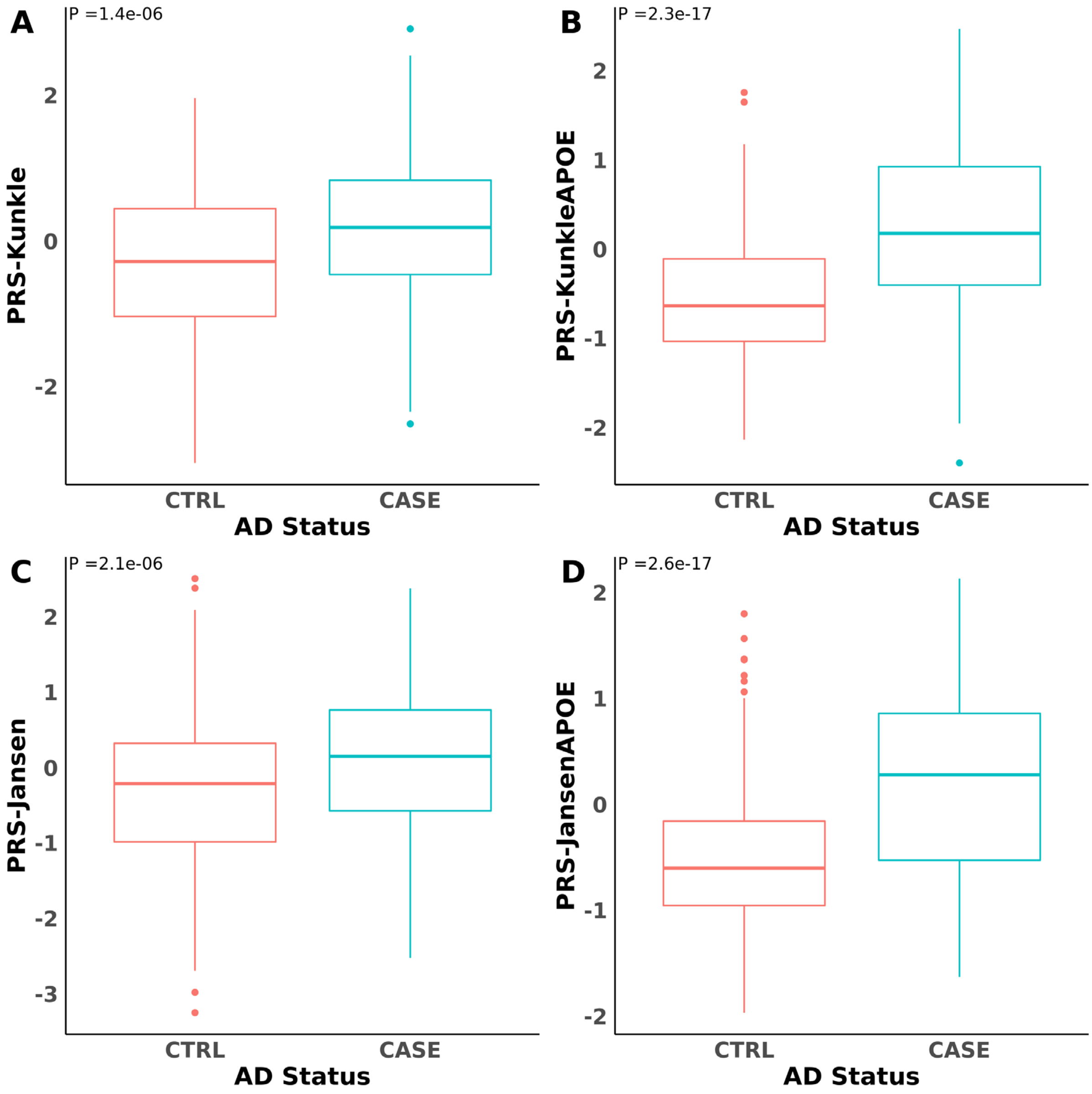


Figure 6.6: All four polygenic risk scores (PRS) for LOAD were significantly higher in AD cases compared to controls. *Shown are the results for analyses stratifying samples based on* ***(A)*** *a PRS generated using the Kunkle at al. GWAS excluding variants in the APOE region;* ***(B)*** *a PRS generated using the Kunkle at al. GWAS including variants in the APOE region;* ***(C)*** *a PRS generated using the Jansen at al. GWAS excluding variants in the APOE region; and* ***(D)*** *a PRS generated using the Jansen at al. GWAS including variants in the APOE region. AD = Alzheimer's disease. AD status CTRL = Braak neurofibrillary tangle stage < III and AD status CASE = Braak neurofibrillary tangle stage > IV.*

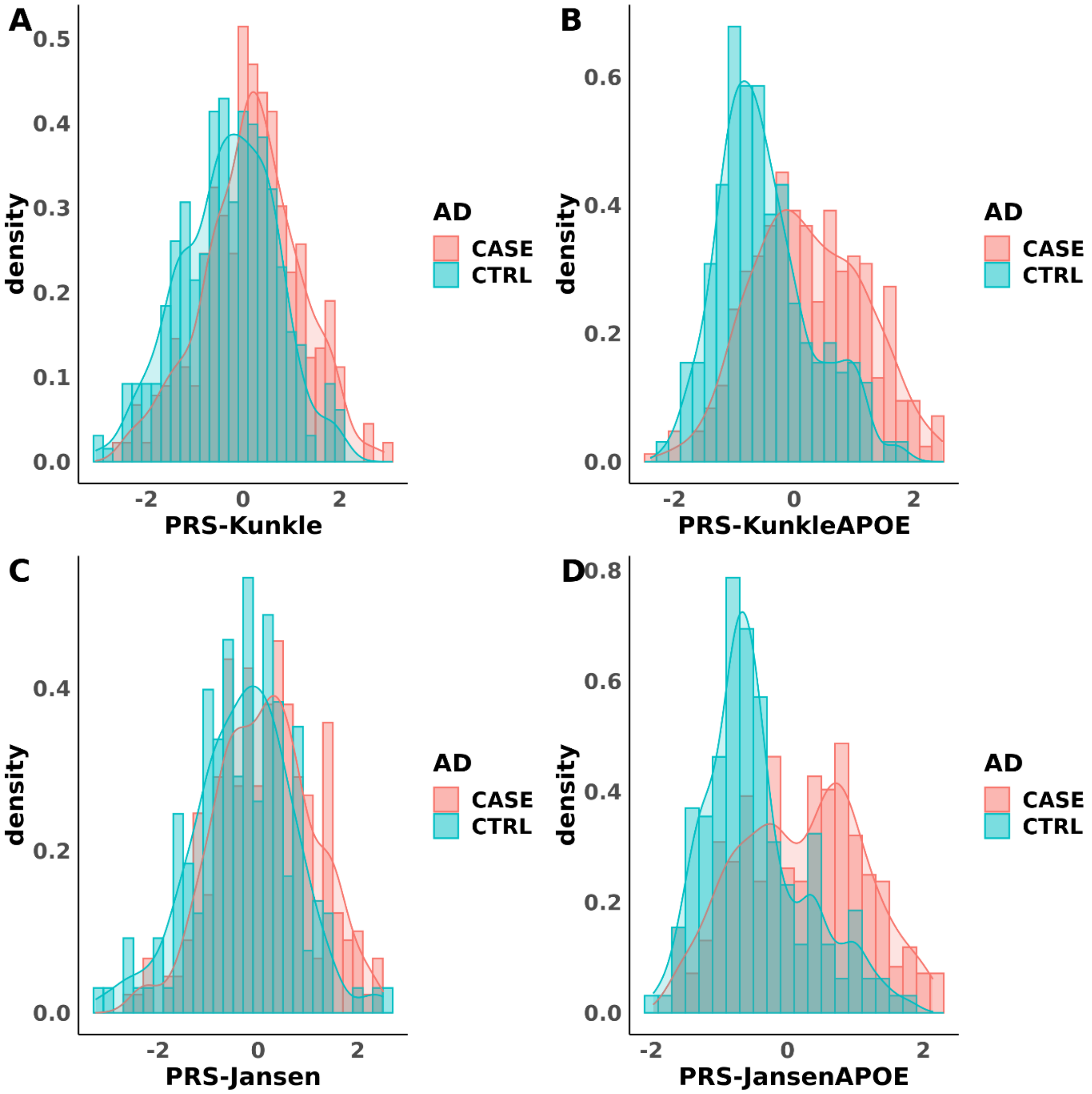


Figure 6.7: The distribution of polygenic risk scores (PRS) for LOAD were significantly higher in AD cases compared to controls. *Shown are the results for analyses stratifying samples based on* ***(A)*** *a PRS generated using the Kunkle at al. GWAS excluding variants in the APOE region;* ***(B)*** *a PRS generated using the Kunkle at al. GWAS including variants in the APOE region****;*** ***(C)*** *a PRS generated using the Jansen at al. GWAS excluding variants in the APOE region; and* ***(D)*** *a PRS generated using the Jansen at al. GWAS including variants in the APOE region. AD = Alzheimer's disease. AD status CTRL = Braak neurofibrillary tangle stage < III and AD status CASE = Braak neurofibrillary tangle stage > IV.*

6.4.3 The four different LOAD PRS correlate with each other

As expected, the four different PRS were significantly (Bonferroni p<0.05/6=0.0083) correlated with each other (see **Figure 6.8**). The strongest correlation was seen between the PRS-Kunkle$_{\text{APOE}}$ and the PRS-Jansen$_{\text{APOE}}$ (r=0.91; p=2.2e-308; see **Figure 6.8**). The next strongest correlation was between the PRS-Kunkle and the PRS-Jansen (r=0.65; p=7.4e-126), suggesting the inclusion or exclusion of *APOE* likely leads to a similar set of genetic variants in the PRS (see **Figure 6.8**) and that it is likely to be a major driver of the PRS. This is further supported by the weaker correlations between the PRS-Kunkle and PRS-Kunkle$_{\text{APOE}}$ (r=0.33; p=1.97e-27) and the PRS-Jansen and PRS-Jansen$_{\text{APOE}}$ (r=0.34; p=5.6e-29; see **Figure 6.8**). The weakest correlations were between the PRS-Jansen and PRS-Kunkle$_{\text{APOE}}$ (r=0.26; p=1.42e-17) and the PRS-Kunkle and PRS-Jansen$_{\text{APOE}}$ (r=0.24; p=5.6e-15), which likely reflects the differences in the set of genetic variants included in each PRS (see **Figure 6.8**).

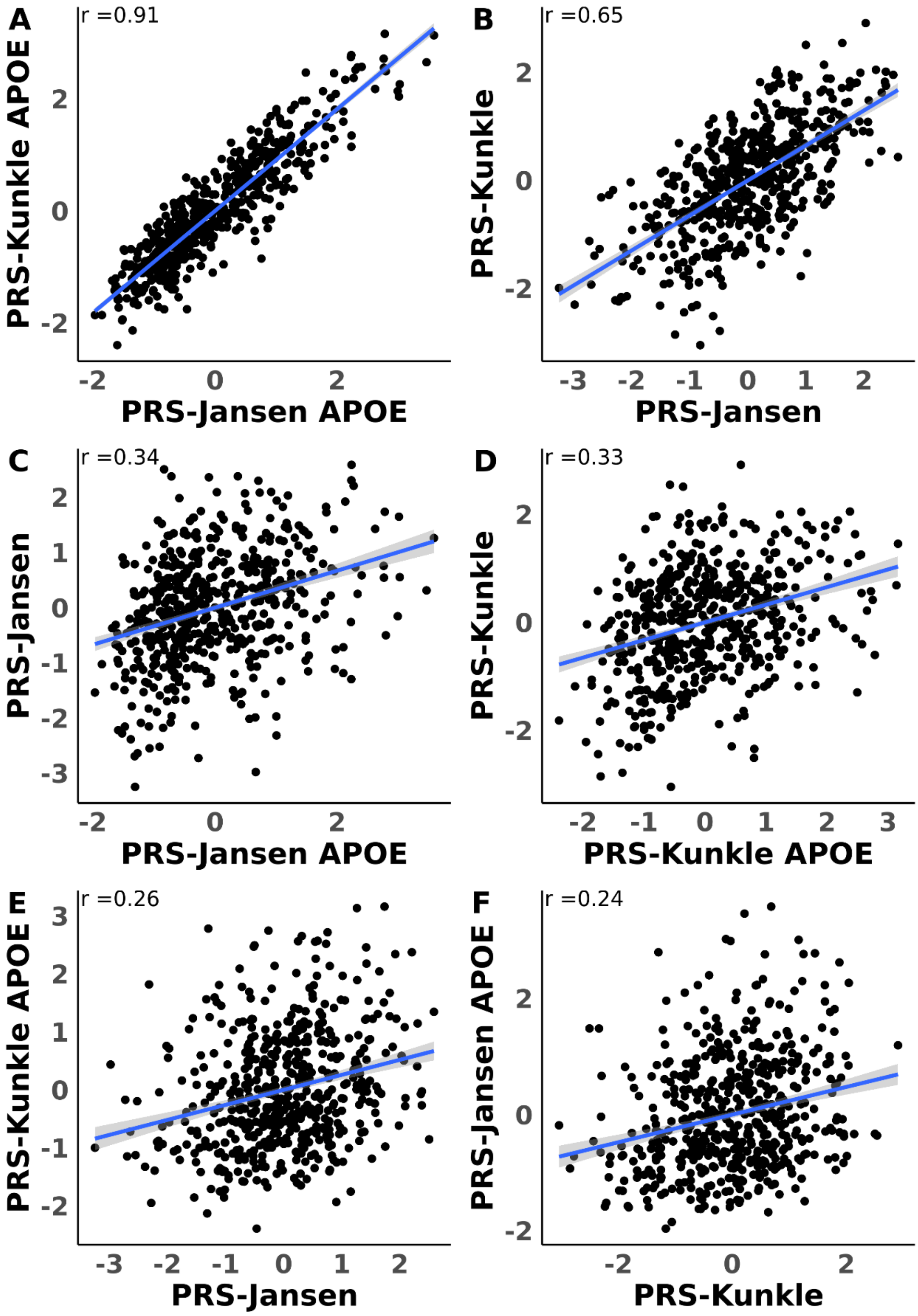


Figure 6.8: Correlations between PRS values generated two different GWAS datasets, including and excluding variants in the APOE region, in BDR donors. *Shown are correlations between PRS derived using data from **(A)** the Kunkle et al GWAS and the Jansen at al. GWAS including variants in the APOE region, **(B)** the Kunkle at al. GWAS and the Jansen et al GWAS excluding variants in the APOE region, **(C)** the Jansen at al. GWAS including variants in the APOE region and the Jansen at al. GWAS excluding variants in the APOE region, **(D)** the Kunkle at al. GWAS including variants in the APOE region and the Kunkle at al. GWAS excluding variants in the APOE region, **(E)** the Kunkle at al. GWAS including variants in the APOE region and the Jansen at al. GWAS excluding variants in the APOE regions, and **(F)** the Kunkle at al. GWAS excluding variants in the APOE region and the Jansen at al. GWAS including variants in the APOE region.*

6.4.4 Common genetic risk factors for Alzheimer’s disease are associated with multiple aspects of neuropathology

The number of *APOE* ε4 alleles was positively associated (Bonferroni p< 0.05/36=0.00014) with all four semi-quantitative neuropathology measures (see **Table 6.4**). The most significant association was with Braak NFT stage: each ε4 allele was associated with an increase in 1.16 Braak NFT stages (p=4.16e−24). Associations were also found between ε4 status and Thal Aβ phase (mean difference per ε4 allele = 0.981 phases; p=3.96 e−20), neuritic plaque density (mean difference per ε4 allele = 0.713 stages; p=1.03e−19) and Braak Lewy body stage (mean difference per ε4 allele = 0.555 stages; p=2.64e−04). PRS-Kunkle was associated with two measures of neuropathology (see **Table 6.4**): a higher polygenic burden was associated with Braak NFT stage (mean difference per SD of PRS = 0.354 stages; p=1.36e−6) and neuritic plaque density (mean difference per SD of PRS = 0.202 stages; p=5.27 e−5). TDP-43 was not associated with either *APOE* genotype or Alzheimer’s disease PRS. Although variants in the *APOE* region were excluded from the PRS, we tested both *APOE* and PRS against Braak NFT stage and neuritic plaque density simultaneously to confirm that the identified associations were independent. The estimated effects of ε4 on both Braak NFT stage and neuritic plaque density were unaffected, while the Alzheimer’s disease PRS associations were slightly attenuated (see **Table 6.4**) but remained significant. In addition to an additive model, we tested whether there was evidence for a multiplicative effect between Alzheimer’s disease PRS and *APOE* genotype on neuropathological burden to explore the hypothesis that in individuals with protective *APOE* genotypes, Alzheimer’s disease PRS is more important (i.e. has a larger effect on neuropathology). In this analysis, none of the five neuropathological variables had statistically significant differences across *APOE* genotype groups (p > 0.05) (see **Table 6.5**). Taken together, these results suggest that *APOE* status and Alzheimer’s disease PRS are independently associated with neuropathology, combining in an additive manner to influence an individual’s accumulation of tauopathy (NFTs) and Aβ plaques.

Table 6.4: Common genetic risk factors for Alzheimer’s disease are associated with multiple aspects of neuropathology

Analytical model	**Neuropathological variable**	***APOE***						**PRS**	**Coefficient**	**%VarExp**
		Number of ε 2 alleles			**Number of ε 4 alleles**					
		***P*-value**	**Coefficient**	**%VarExp**	***P*-value**	**Coefficient**	**%VarExp**	***P*-value**		
Model 1	Braak NFT Stage	0.0877	−0.357	0.958	4.16E−24	1.16	15.1			
	Thal amyloid stage	0.00333	−0.562	1.54	3.96E−20	0.981	13.5			
	CERAD stage	0.0224	−0.329	1.99	1.03E−19	0.713	13.4			
	Braak LB stage	0.988	−0.00439	0.0809	0.000264	0.555	2.59			
	TDP-43	0.859	−0.0574	0.00821	0.00158	0.537	2.58			
Model 2	Braak NFT Stage							1.36E−06	3.4	0.354
	Thal amyloid stage							0.00288	1.1	0.201
	CERAD stage							5.27E−05	2.95	0.202
	Braak LB stage							0.267	0.167	0.105
	TDP-43							0.315	0.26	0.104
Model 3	Braak NFT stage	0.0885	−0.3505	0.9580	9.40E−24	1.132	15.119	4.97E−06	0.309	2.465
	CERAD stage	0.0224	−0.3254	1.9865	2.02E−19	0.700	13.402	1.30E−04	0.179	2.192

Table 6.5: No evidence of differential effect on neuropathology by Alzheimer's Disease PRS when stratified by APOE status.

Neuropathology variable	**Main effects**									**Interaction effects**			
	APOE						**Polygenic risk score**			**Number of E2 alleles x PRS**		**Number of E4 alleles x PRS**	
	Number of E2 alleles			**Number of E4 alleles**									
	P-value	**Coeff.**	**%VarExp**	**P-value**	**Coeff.**	**%VarExp**	**P-value**	**Coeff.**	**%VarExp**	**P-value**	**%VarExp**	**P-value**	**%VarExp**
Braak NFT stage	0.0665	-0.381	0.958	2.27e-23	1.13	15.1	3.78e-05	0.387	2.46	0.609	0.461	0.272	0.194
Thal amyloid stage	0.00146	-0.61	1.54	6.49e-19	0.95	13.5	0.00269	0.257	0.589	0.0554	0.167	0.266	0.0559
CERAD stage	0.0172	-0.342	1.99	4.01e-19	0.699	13.4	0.000459	0.223	2.19	0.626	0.425	0.359	0.0985
Braak LB stage	0.941	-0.021	0.0809	0.000147	0.58	2.59	0.0717	0.232	0.0964	0.338	0.198	0.0376	0.104
TDP-43	0.924	-0.0312	0	0.00183	0.534	0	0.798	0.0376	0	0.712	0.0356	0.689	0.0418

Given that the two molecular pathologies—tauopathy and β-amyloidosis—that define Alzheimer's disease are highly correlated (see **Figure 6.9**), we wanted to establish whether *APOE* or Alzheimer's disease PRS had a specific (or primary) effect on a particular aspect of neuropathology. To this end, we repeated the analysis of how Alzheimer's disease PRS and *APOE* influence pathology, sequentially controlling for other neuropathology variables. This analysis revealed some interesting patterns. First, after controlling for any of the other three quantitative neuropathological variables, Braak Lewy body stage was not significantly associated with *APOE* ε4 (see **Table 6.6**) suggesting that the association we detected was largely driven by the fact that individuals with Lewy bodies have also NFTs and Aβ plaques. Second, after we controlled for Braak NFT stage, neither of the plaque measures remained significantly associated with *APOE* ε4. In contrast, Braak NFT stage remained significantly associated with *APOE* ε4 status after controlling for plaque variable (adjusted for Thal phase, mean difference per *APOE* ε4 allele = 0.468; *p*=6.44 -07; adjusted for neuritic plaque density, mean difference per ε4 allele = 0.238; *p*=1.82 e-04), albeit with an attenuated magnitude of effect. Considering the two measures of plaque burden, only Thal Aβ phase remained significantly associated with ε4 after controlling for neuritic plaque density (mean difference per ε4 allele = 0.265; *p*=3.4e-04). Neither Braak NFT stage nor neuritic plaque density remained significantly associated with Alzheimer's disease PRS after controlling for the other measure of pathology. These results indicate that *APOE* ε4 has a specific influence on tauopathy (NFTs) as well as a shared effect on both plaque and NFT development, whereas the PRS is more generally associated with an increased burden of Alzheimer's disease neuropathology.

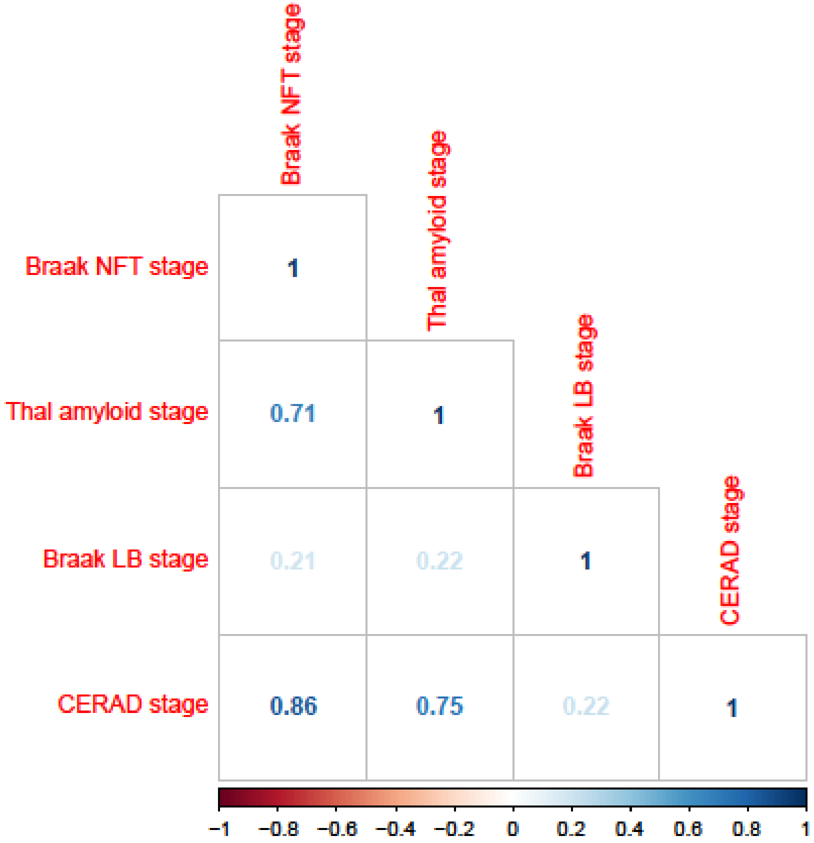


Figure 6.9: Heatmap of correlations between semi-quantitative measures of Neuropathology. NFT- neurofibrillary tangles, LB – Lewy body.

Table 6.6: APOE is associated with a shared effect on neurofibrillary tangles and Aβ. Results from regression models testing for associations between APOE or Alzheimer's disease PRS and neuropathology while controlling for other measures of neuropathology.

		APOE						**Polygenic risk score**		
		Number of E2 alleles			**Number of E4 alleles**					
		P-value	**Coeff.**	**%VarExp**	**P-value**	**Coeff.**	**%VarExp**	**P-value**	**Coeff.**	**%VarExp**
Covary for Braak NFT stage	**Thal amyloid stage**	0.00755	-0.377	1.66	0.00202	0.257	13.7	0.394	-0.0402	0.52
	CERAD stage	0.0431	-0.15	2.07	0.223	0.0515	13.4	0.855	0.00447	2.34
	Braak Lewy body stage	0.863	0.0489	0.0874	0.0625	0.305	2.53	0.804	0.0236	0.093
	TDP-43	0.952	-0.0205	0.00101	0.238	0.221	0.389	0.84	-0.0228	0.0114
Covary for Thal amyloid stage	**Braak NFT stage**	0.486	0.111	0.868	8.08e-07	0.459	16.5	0.000124	0.202	2.33
	CERAD stage	0.951	-0.00601	1.91	0.00265	0.169	13.7	0.00113	0.105	2.25
	Braak Lewy body stage	0.609	0.146	0.0946	0.0999	0.276	2.57	0.396	0.081	0.18
	TDP-43	0.73	0.12	0.0355	0.401	0.163	0.212	0.81	0.0275	0.0173
Covary for CERAD stage	**Braak NFT stage**	0.335	0.108	0.87	0.000159	0.239	15	0.0304	0.0799	2.54
	Thal amyloid stage	0.0533	-0.247	1.61	0.000336	0.265	13.6	0.379	-0.0374	0.5
	Braak Lewy body stage	0.695	0.11	0.0735	0.0896	0.273	2.51	0.745	0.0305	0.111
	TDP-43	0.808	0.0832	0.0168	0.134	0.277	0.646	0.82	-0.0257	0.0149
Covary for Braak LB stage	**Braak NFT stage**	0.269	-0.238	0.656	9.48e-19	1.07	15.5	3.10e-05	0.3	2.41
	Thal amyloid stage	0.00506	-0.541	1.44	5.37e-17	0.923	14.3	0.0327	0.138	0.576
	CERAD stage	0.0442	-0.29	1.82	6.32e-17	0.669	14	0.000664	0.163	2.01
	TDP-43	0.806	0.0817	0.0176	0.00281	0.539	2.61	0.651	0.0509	0.0605
Covary for TDP-43	**Braak NFT stage**	0.16	-0.294	0.787	7.90e-21	1.07	15.1	3.09e-05	0.285	2.16
	Thal amyloid stage	0.0113	-0.49	1.23	1.39e-17	0.919	13.2	0.0285	0.139	0.478
	CERAD stage	0.0373	-0.3	1.8	3.62e-17	0.659	13.7	0.000192	0.175	2.24
	Braak Lewy body stage	0.988	0.00446	0.11	0.000263	0.573	3.02	0.443	0.0741	0.0544

6.4.5 Methylomic variation is associated with LOAD PRS in the cortex

To identify if polygenic burden for LOAD is associated with variable DNAm in the cortex I first calculated PRS using summary statistics from two recent LOAD GWAS (Kunkle et. al and Jansen *at al.*) for individuals in the BDR cohort which consists of 522 PFC and 525 OCC samples (see Methods section **6.3.2** for more details). To assess the effects of including *APOE* in the PRS on DNAm, PRS were calculated both including and excluding the *APOE* locus. Mixed effects regression models were then run against genome-wide DNAm controlling for age, sex, batch, derived neural cell proportions and genetic PC1 as fixed effects and individual as a random effect (see Chapter 4 section **4.4.2** for details on the covariates chosen). To identify if any DMPs reflect a direct *cis*-genetic effect on DNAm, I utilised the cortex mQTL results generated in Chapter 5 (see section **5.4.3**). I only considered methylation quantitative trait loci (mQTLs; SNPs associated with DNAm) which were included in the PRS or variants which are in high LD with these SNPs ($r^2 \geq 0.8$), which I refer to as PRS-mQTLs.

6.4.5.1 Multiple differentially methylated positions are associated with PRS-Kunkle in the cortex

I identified seven experiment-wide significant DMPs associated with PRS-Kunkle (see **Figure 6.10**). Three (43%) of the DMPs were significantly hypermethylated with elevated PRS and the other four (57%) were hypomethylated (see **Figure 6.11**; **Table 6.7**). There was no evidence of statistical inflation in this analysis, as shown by the quantile-quantile plots (λ =1.03; see **Figure 6.12**). The average absolute magnitude of effect for the significant DMPs per SD increase in LOAD PRS was 1.11% (inter-quartile range [IQR] = 0.85-1.29%). To identify if any of these DMPs reflect a direct *cis*-genetic effect on DNAm, I utilised the cortex mQTL results generated in Chapter 5 (section **5.4.3**). There was evidence of a relationship between a PRS-mQTL and DNAm for six (86%) of the DMPs, indicating that these sites are under direct genetic control at the SNP level. Three DMPs had no gene annotation but the other four (cg09507712, cg07409245, cg02848401 and cg27051260) are annotated to *C1QTNF4* (chromosome 11) and were hypomethylated with elevated PRS, with a 0.64-1.39% decrease in DNAm per SD increase in PRS (see **Figure 6.13**). *C1QTNF4* has been associated with AD in eQTL studies and resides around the *SPI1* GWAS locus. The expression of several genes within this locus are highly correlated with one another and have been associated with AD status (Karch *at al.*, 2016). These results suggest *C1QTNF4* is potentially a LOAD risk locus in addition to the currently implicated *SPI1* locus. No DMRs were significantly associated with PRS-Kunkle in the cortex.

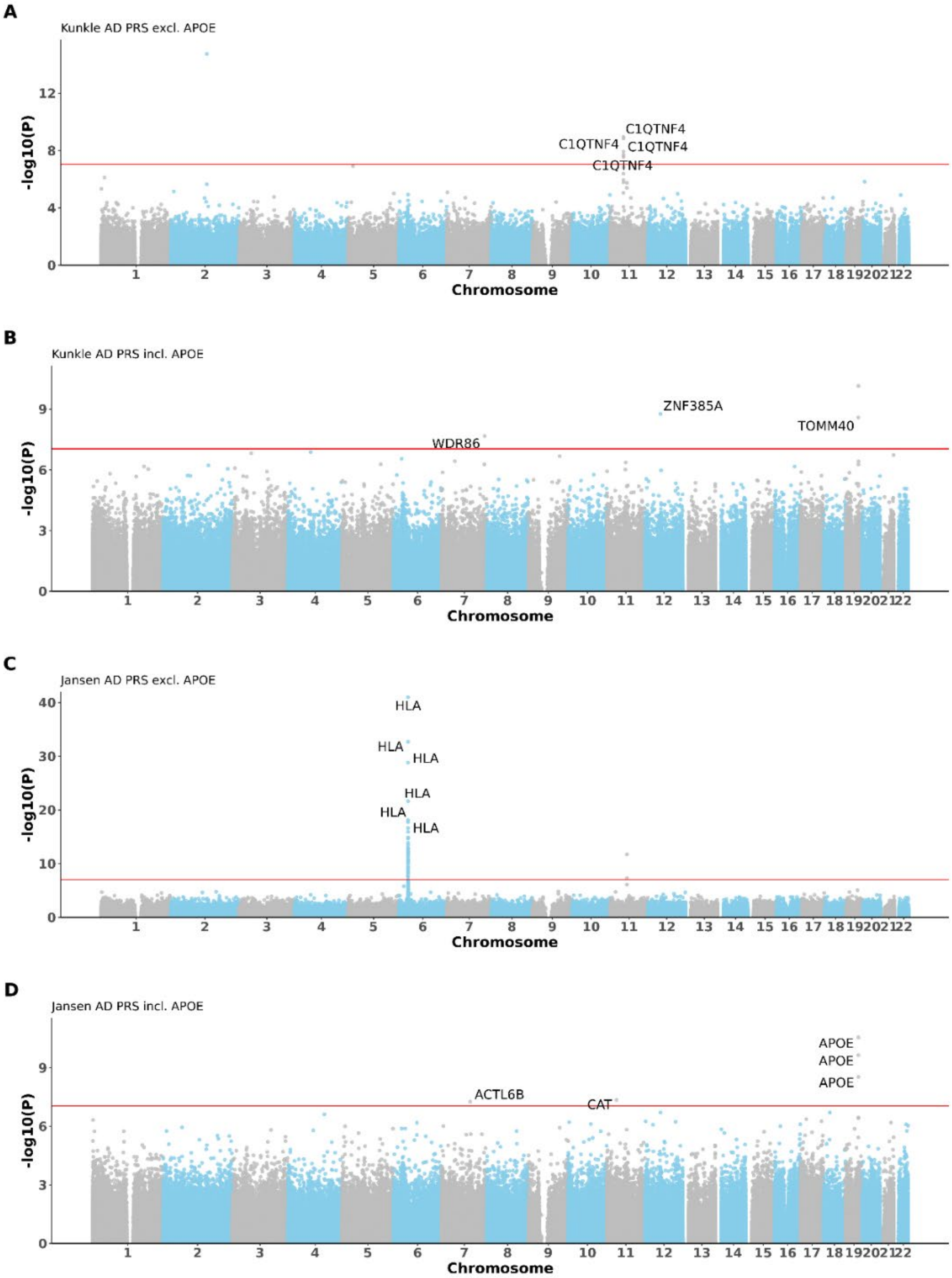


Figure 6.10: Cortex EWAS of LOAD PRS highlights experiment-wide significant differentially methylated positions. *Manhattan plots showing results of four EWAS across two cortical regions (prefrontal cortex and occipital cortex). DNA methylation was regressed against PRS where in **(A)** LOAD PRS was generated using the Kunkle at al. GWAS excluding APOE **(B)** LOAD PRS was generated using the Kunkle at al. GWAS including APOE; **(C)** LOAD PRS was generated using the Jansen at al. GWAS excluding APOE; and **(D)** LOAD PRS was generated using the Jansen at al. GWAS including APOE. The significant differentially methylated positions are annotated with their Illumina UCSC gene name, unless they are unannotated to a gene. The X-axis shows chromosomes 1-22 and the Y-axis shows -log10(P), with the horizontal red line representing experiment wide significance (p< 9e-8). Late onset Alzheimer's disease. PRS = Polygenic risk scores.*

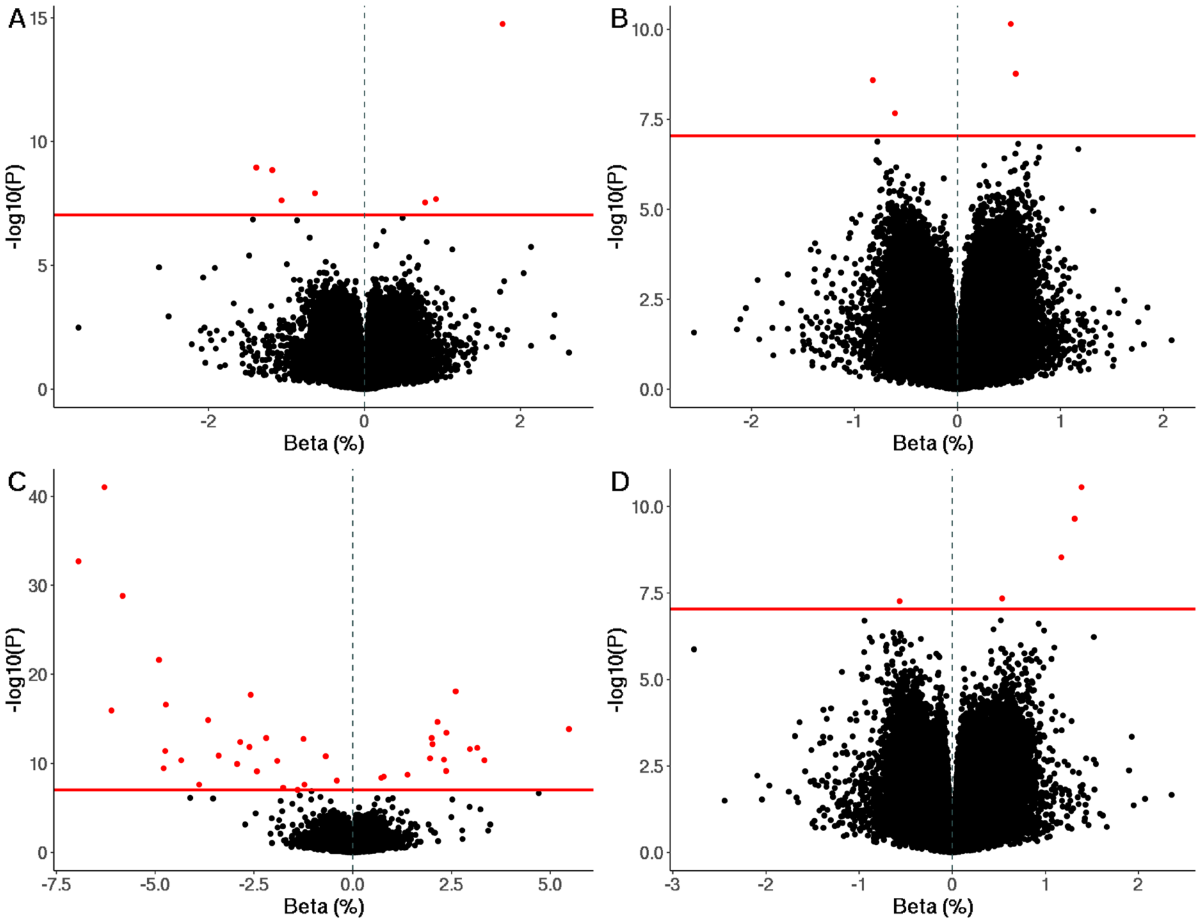


Figure 6.11: Volcano plot of differentially methylated positions (DMPs) identified in the cortex LOAD PRS EWAS. *DNA methylation was regressed against PRS where in* ***(A)*** *LOAD PRS was generated using the Kunkle at al. GWAS excluding APOE* ***(B)*** *LOAD PRS was generated using the Kunkle at al. GWAS including APOE;* ***(C)*** *LOAD PRS was generated using the Jansen at al. GWAS excluding APOE; and* ***(D)*** *LOAD PRS was generated using the Jansen at al. GWAS including APOE. The X-axis shows beta effect size (ES) and the Y-axis shows -log10(p). Red probes indicate a p-value that reaches experiment-wide significance (EWS) (p < 9e-08). LOAD = Late onset Alzheimer's disease. PRS = Polygenic risk scores.*

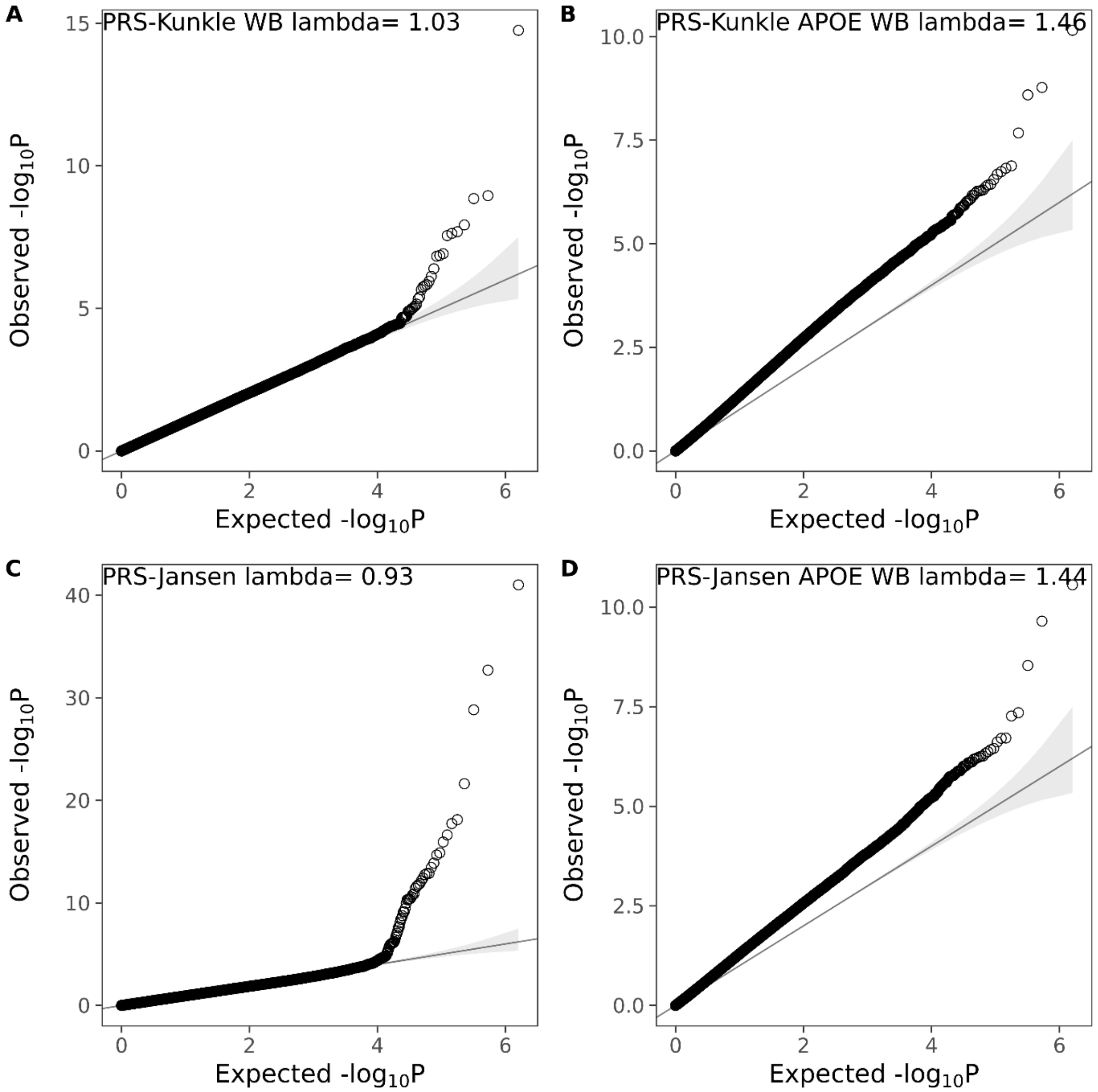


Figure 6.12: Quantile-quantile plots of the LOAD-PRS EWAS conducted in the cortex. *Shown are the expected (x-axis) against the observed (y-axis) quantiles in each EWAS against LOAD PRS where* ***(A)*** *LOAD PRS was generated using the Kunkle at al. GWAS excluding APOE* ***(B)*** *LOAD PRS was generated using the Kunkle at al. GWAS including APOE;* ***(C)*** *LOAD PRS was generated using the Jansen at al. GWAS excluding APOE; and* ***(D)*** *LOAD PRS was generated using the Jansen at al. GWAS including APOE. Lambda= genomic inflation factor.*

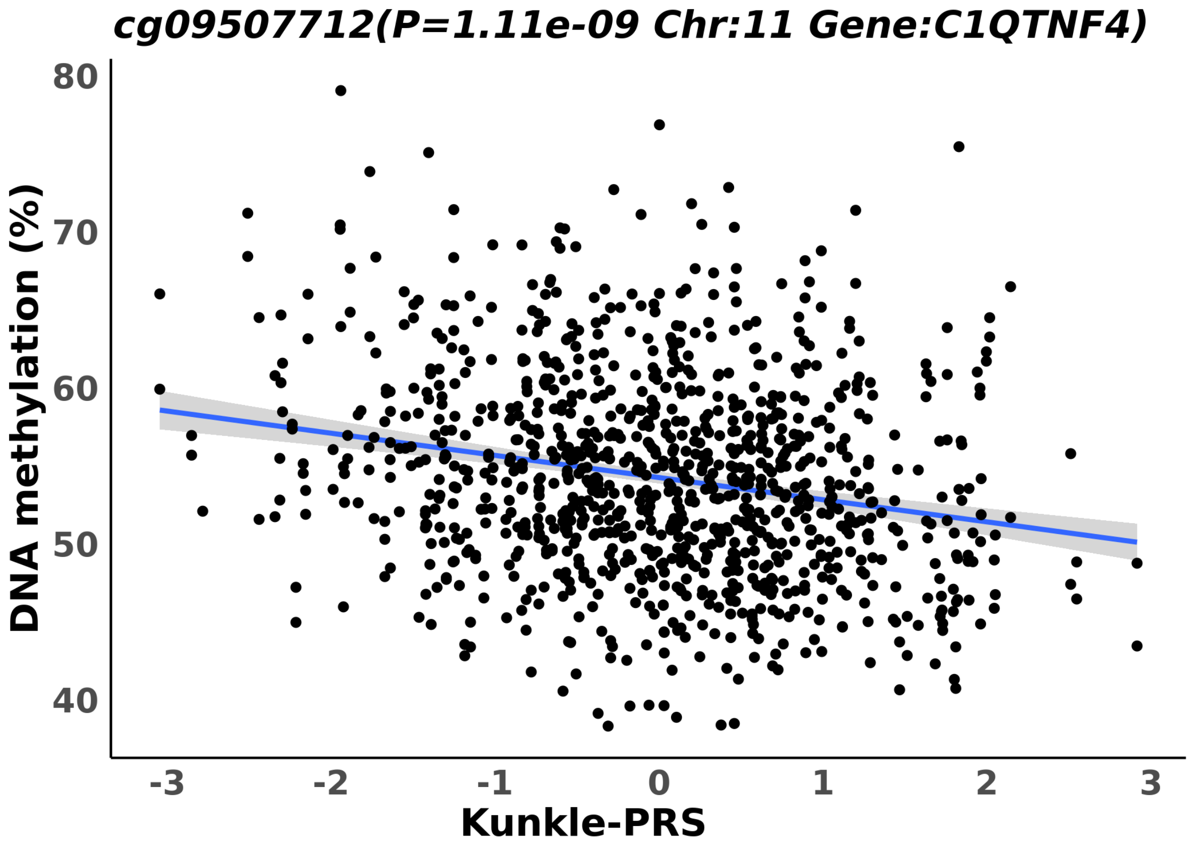


Figure 6.13: cg09507712 is hypomethylated with elevated PRS-Kunkle in the cortex. *The x-axis represents the standardised (mean of 0, standard deviation=1) PRS-Kunkle and the y-axis represents DNA-methylation adjusted for confounders. PRS-Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS excluding the APOE region*

Table 6.7: Differentially methylated positions were associated with the PRS-Kunkle at experiment wide significance (p< 9E− 8) in the cortex. *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. Assoc. mQTL = associated with a methylation quantitative trait loci (mQTL) which was included in the PRS. PRS-Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS excluding the APOE region.*

CpG	Chr	BP	Gene	Gene Region	Beta (%)	SE (%)	P	Assoc. mQTL
cg14317533	2	127886316	-	-	1.77	0.22	1.77e-15	mQTL
cg09507712	11	47616693	*C1QTNF4*	TSS1500	-1.39	0.23	1.11e-09	mQTL
cg07409245	11	47616751	*C1QTNF4*	TSS1500	-1.18	0.20	1.41e-09	mQTL
cg02848401	11	47617346	*C1QTNF4*	TSS1500	-0.64	0.11	1.18e-08	-
cg24977308	11	47636548	-	-	0.92	0.16	2.04e-08	mQTL
cg27051260	11	47616825	*C1QTNF4*	TSS1500	-1.07	0.19	2.35e-08	mQTL
cg17688768	11	47628441	-	-	0.77	0.14	2.80e-08	mQTL

6.4.5.2 Multiple differentially methylated positions and regions were associated with PRS-Kunkle$_{APOE}$ in the cortex

I identified four experiment-wide significant DMPs associated with PRS-Kunkle$_{APOE}$ (see **Figure 6.10**). Two (50%) of the DMPs were significantly hypermethylated with elevated PRS and the other two (50%) were hypomethylated (see **Figure 6.11; Table 6.8**). There was slight p-value inflation in this analysis, as shown by the quantile-quantile plots (λ =1.46; see **Figure 6.12**), however, the test statistics were normally distributed suggesting the results have not been influenced by bias. The average absolute magnitude of effect for the significant DMPs per SD increase in LOAD PRS was 0.62% (inter-quartile range [IQR] = 0.55-0.66%). To identify if any of these DMPs reflect a direct cis-genetic effect on DNAm, I utilised the cortex mQTL results generated in Chapter 5 (section **5.4.3**). There was evidence of a relationship between a PRS-mQTL and DNAm for one of these DMPs - cg17928676 - which is located 2.8kb away from *APOE*. Of the four DMPs, three were annotated to genes. Two of the PRS-Kunkle$_{APOE}$ associated DMPs are annotated to genes which are of relevance in the context of AD and associated neurobiological functions. These include:

- cg07332724 located on chromosome 12 and annotated to *ZNF385A* was hypermethylated with increasing PRS, with a 0.56% increase (p= 1.69e-09) for every SD increase in PRS. Elevated DNAm in *ZNF385A* has previously been associated with Braak stage in the PFC (Smith *at al.*, 2018).
- cg02613937 located on chromosome 19 and annotated to *TOMM40* was hypomethylated for PRS, with a 0.82% decrease (p= 2.55e-09) for every SD increase in PRS (see **Figure 6.14**). *TOMM40* is in the same LD region as *APOE* (located 2.1kb away from *APOE*) and has been associated with age of onset of LOAD (Roses *at al.*, 2010). Multiple regulatory elements, spanning the *TOMM40-APOE*-APOC2 cluster have been shown to regulate gene expression across this region (Shao *at al.*, 2018). In addition, variable DNAm in this region correlates with AD-related biomarkers and *TOMM40*/ *APOE* expression in AD (Shao *at al.*, 2018).

cg13258599 located on chromosome 7 and annotated to *WDR86*, was hypomethylated with elevated PRS, with a 0.61% decrease (p= 2.11e-08) for every SD increase in PRS. *WDR86* is expressed throughout the brain although it has not previously been implicated in AD or other neurodegenerative diseases and its functions are unknown.

Variation in PRS-Kunkle$_{APOE}$ was associated with two DMRs. One region was significantly hypermethylated with elevated PRS (p=5.29e-10) and annotated to *KLHL33* (chromosome 14), which is a protein coding gene predominantly expressed in the brain in primary hippocampal neurons, astrocytes and oligodendrocytes and studies suggests it plays a role in the functioning and development of the nervous system and in the differentiation of oligodendrocytes (Jiang *at al.*, 2005; Soltysik-Espanola *at al.*, 1999). The other was significantly hypomethylated with elevated PRS (p=2.78e-09) and was annotated to *PITPNM2* (chromosome 12). *PITPNM2* is a part of the phosphatidylinositol pathway, which have been implicated in neuropsychiatric disorders such as bipolar, depression and schizophrenia (Lescai *at al.*, 2017; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

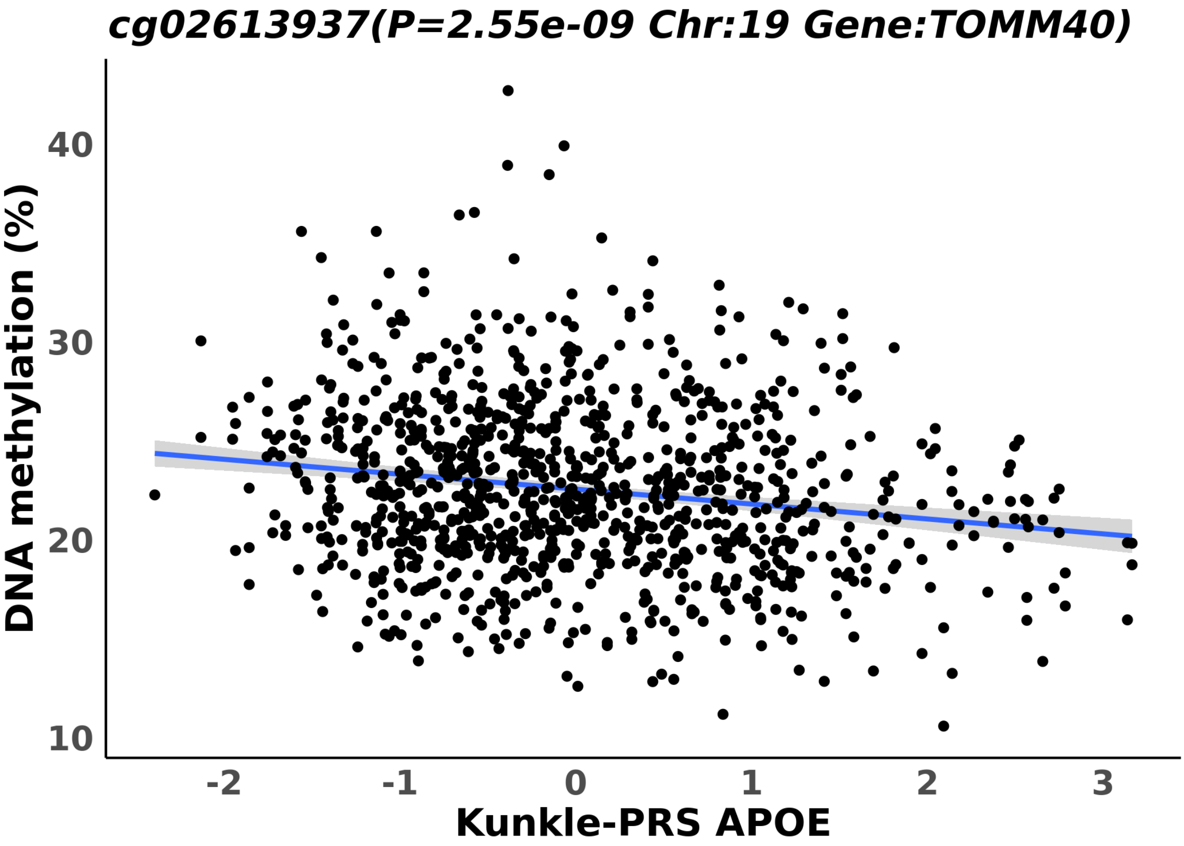


Figure 6.14: cg02613937 is hypomethylated with elevated PRS-Kunkle$_{APOE}$ in the cortex. *The x-axis represent the standardised (mean of 0, standard deviation=1) PRS-Kunkle$_{APOE}$ and the y-axis represents DNA-methylation adjusted for confounders. PRS-Kunkle$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS including the APOE region.*

Table 6.8: Differentially methylated positions associated with the PRS-Kunkle$_{APOE}$ in the cortex at experiment wide significance (p< 9E− 8). *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. Assoc. mQTL = associated with a methylation quantitative trait loci (mQTL) which was included in the PRS. PRS-Kunkle$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS including the APOE region.*

CpG	Chr	BP	Gene	Gene Region	Beta (%)	SE (%)	P	Assoc. mQTL
cg17928676	19	45414742	-	-	0.52	0.08	7.07e-11	mQTL
cg07332724	12	54773114	*ZNF385A;LOC102724050*	Body	0.56	0.09	1.69e-09	-
cg02613937	19	45395297	*TOMM40*	Body	-0.82	0.14	2.55e-09	-
cg13258599	7	151089761	*WDR86*	Body	-0.61	0.11	2.11e-08	-

Table 6.9: Differentially methylated regions associated with PRS-Kunkle$_{APOE}$ in the cortex. *Two regions were identified. Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected P value adjusted for the number of independent tests. PRS- KunkleAPOE = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS including the APOE region.*

Chr	BP start	BP end	n	Gene	Beta (%)	SE (%)	P	P adjusted
14	20903410	20904169	11	*KLHL33*	8.23	1.33	5.29e-10	4.80e-04
12	123469284	123469669	3	*PITPNM2*	-4.6	0.77	2.78e-09	2.52e-03

6.4.5.3 Multiple differentially methylated positions were associated with PRS-Jansen

I identified 40 experiment-wide significant DMPs associated with PRS-Jansen (see **Figure 6.10**). 15 (37.5%) of the DMPs were characterised by significant hypermethylation with elevated PRS and the other 25 (62.5%) were hypomethylated (see **Figure 6.11; Table 6.10**). There was no evidence of statistical inflation in this analysis, as shown by the quantile-quantile plots (λ =0.93; see **Figure 6.12**).The average absolute magnitude of effect for the significant DMPs per SD increase in LOAD PRS was 2.99% (inter-quartile range [IQR] = 1.95-4%). To identify if any of these DMPs reflect a direct *cis*-genetic effect on DNAm, I utilised the cortex mQTL results generated in Chapter 5 (section **5.4.3**). There was evidence of a relationship between a PRS-mQTL and DNAm for 15 of the DMPs, indicating that these sites are under direct genetic control at the SNP level. Of the 40 DMPs, 32 were annotated to genes, and several are of relevance in the context of AD and associated neurobiological functions including:

- 27 DMPs (68%) were annotated to the HLA region (chromosome 6) with the most significant being cg13076785 which is annotated to *HLA-DRB6* and was hypomethylated with elevated PRS with a 6.29% decrease PRS in DNAm for every SD increase in PRS (p=9.74e-42; see **Figure 6.15**). cg08265274, which is annotated to *HLA-DRB5* was hypomethylated with elevated PRS with a 5.83% decrease in DNAm for every SD increase in PRS (p=1.48e-29). Several of these PRS-associated DMPs were also associated with a PRS-mQTL, suggesting there are *cis*-genetic effects at this locus. In concordance with the Kunkle-PRS, these results support a role for the function of the immune system in the aetiology of AD. Epidemiological studies have long suggested a role of immune system dysregulation in AD and the HLA region encodes several proteins that play major roles in the immune system and is highly expressed in microglia and other myeloid cells. The HLA region has been associated with AD and the top candidate gene in HLA locus is *HLA-DRB1* (Jansen *at al.*, 2019; Kunkle *at al.*, 2019) and four DMPs were annotated to this gene. This region is known for its complex organisation and fine mapping analyses used by Kunkle and colleagues (2019) of HLA and genetic correlations suggest LOAD has a shared genetic architecture with several immune-mediated and cognitive phenotypes.

- cg20636526, cg09452510 and cg25457674 located on chromosome 6 and annotated to *C6orf10*. All three DMPs are associated with a PRS-mQTL, suggesting there is a *cis*-genetic effect at this locus. cg20636526 was hypomethylated with elevated PRS (2.19% decrease per SD in PRS; p=1.33e-13) whereas cg09452510 and cg25457674 were hypermethylated with elevated PRS (1.96% and 0.78% increase in DNAm per SD in PRS, respectively; p=2.59e-11, p=3.00e-09). *C6orf10* is an AD risk gene, which has been strongly associated with Tau pathology (Ma *at al.*, 2020).
- cg09618893 located on chromosome 6 and annotated to *EHMT2* was significantly hypomethylated with elevated PRS with a decrease of 1.23% in DNAm (p=2.26e-08) for every SD increase in PRS. *EHMT2* is an AD risk gene, and has been associated with differential histone methylation in AD cases (Coneys & Wood, 2020), and mice deficient in *EHMT2* have learning deficits (Maze *at al.*, 2010). In addition, evidence suggests in AD patients high *EHMT2* levels results in reduced synapse strength leading to a detrimental effect on cognition (Coneys & Wood, 2020).
- cg02521229 located on chromosome 11 had no gene annotation, however it resides in the vicinity (within 50kb) of the MS4A locus which contains several genes which have been implicated in AD via GWAS (Jansen *at al.*, 2019; Kunkle *at al.*, 2019).

Several of the identified DMPs are annotated to genes which have not previously been implicated in neurodegeneration. For example, cg10640833 located on chromosome 6 and annotated to *SKIV2L* was significantly hypermethylated with elevated PRS with an increase of 1.39% in DNAm for every SD increase in PRS (p=1.79e-09). *SKIV2L* is a DEAD box protein is involved in nucleic acid binding and helicase activity and there is no literature suggesting it have been implicated in neurodegenerative disease.

I identified one significant DMR, which was annotated to the HLA region (see **Table 6.11**).

Table 6.10: Differentially methylated positions associated with PRS-Jansen in the cortex at experiment wide significance (p< 9E−8). *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. Assoc. mQTL = associated with a methylation quantitative trait loci (mQTL) which was included in the PRS. PRS-Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region.*

CpG	Chr	BP	Gene	Gene Region	Beta (%)	SE (%)	P	Assoc. mQTL
cg13076785	6	32520916	*HLA-DRB6*	Body	-6.29	0.43	9.74e-42	mQTL
cg23365293	6	32489984	*HLA-DRB5*	Body	-6.94	0.54	2.01e-33	-
cg08265274	6	32490444	*HLA-DRB5*	Body	-5.83	0.49	1.48e-29	mQTL
cg20022036	6	32549496	*HLA-DRB1*	Body	-4.91	0.48	2.37e-22	-
cg15710545	6	32578114	-	-	2.60	0.28	7.60e-19	mQTL
cg05938207	6	32489750	*HLA-DRB5*	Body	-2.59	0.29	1.82e-18	-
cg09139047	6	32552042	*HLA-DRB1*	Body	-4.73	0.54	2.35e-17	mQTL
cg26036029	6	32552443	*HLA-DRB1*	Body	-6.11	0.72	1.15e-16	-
cg17360552	6	32725332	*HLA-DQB2*	Body	-3.67	0.45	1.30e-15	-
cg07007382	6	32578070	-	-	2.14	0.26	2.12e-15	mQTL
cg07984380	6	32547019	*HLA-DRB1*	Body	5.47	0.69	1.31e-14	mQTL
cg12296550	6	32728862	*HLA-DQB2*	Body	2.37	0.31	3.47e-14	-
cg20636526	6	32305145	*C6orf10*	Body	-2.19	0.29	1.33e-13	mQTL
cg15074838	6	32406521	*HLA-DRA*	TSS1500	1.99	0.26	1.35e-13	-
cg13778567	6	32609783	*HLA-DQA1*	Body	-1.25	0.17	1.78e-13	mQTL
cg15602423	6	32552095	*HLA-DRB1*	Body	-2.85	0.38	3.81e-13	-
cg14255617	6	32729118	*HLA-DQB2*	Body	2.01	0.28	6.90e-13	-
cg00551313	6	32521737	*HLA-DRB6*	Body	-2.62	0.36	1.36e-12	-
cg02521229	11	60019236	-	-	3.15	0.44	1.74e-12	mQTL
cg00440797	6	32493873	*HLA-DRB5*	Body	2.96	0.42	2.42e-12	-
cg22233843	6	32632565	*HLA-DQB1*	Body	-4.75	0.67	3.81e-12	-
cg11986643	6	32634316	*HLA-DQB1*	1stExon	-3.40	0.49	1.32e-11	-
cg24283019	6	32381449	-	-	-0.69	0.10	1.56e-11	-

cg09452510	6	32330188	*C6orf10*	Body	1.96	0.29	2.59e-11	mQTL
cg10466124	6	32498285	*HLA-DRB5*	TSS1500	2.30	0.34	3.76e-11	-
cg00119778	6	32466447	-	-	3.33	0.50	4.29e-11	mQTL
cg14645244	6	32552205	*HLA-DRB1*	Body	-4.34	0.65	4.42e-11	mQTL
cg11404906	6	32551749	*HLA-DRB1*	Body	-1.92	0.29	4.93e-11	mQTL
cg01815645	6	32548627	*HLA-DRB1*	Body	-2.93	0.45	1.11e-10	-
cg16345566	6	32633102	*HLA-DQB1*	Body	-4.79	0.75	3.49e-10	-
cg24631579	6	32609181	*HLA-DQA1*	Body	2.36	0.38	7.15e-10	-
cg24242384	6	32551954	*HLA-DRB1*	Body	-2.43	0.39	7.48e-10	-
cg10640833	6	31930323	*SKIV2L*	Body	1.39	0.23	1.79e-09	-
cg25457674	6	32303820	*C6orf10*	Body	0.78	0.13	3.00e-09	mQTL
cg08188015	6	32489553	*HLA-DRB5*	Body	0.72	0.12	3.98e-09	-
cg09666540	6	32381456	-	-	-0.40	0.07	8.30e-09	-
cg09618893	6	31856773	*EHMT2*	Body	-1.23	0.22	2.26e-08	-
cg15820961	6	32558459	*HLA-DRB1*	TSS1500	-3.89	0.69	2.32e-08	-
cg08238829	11	60011506	-	-	-1.76	0.32	5.10e-08	-
cg20946741	6	32428328	-	-	-1.40	0.26	8.83e-08	mQTL

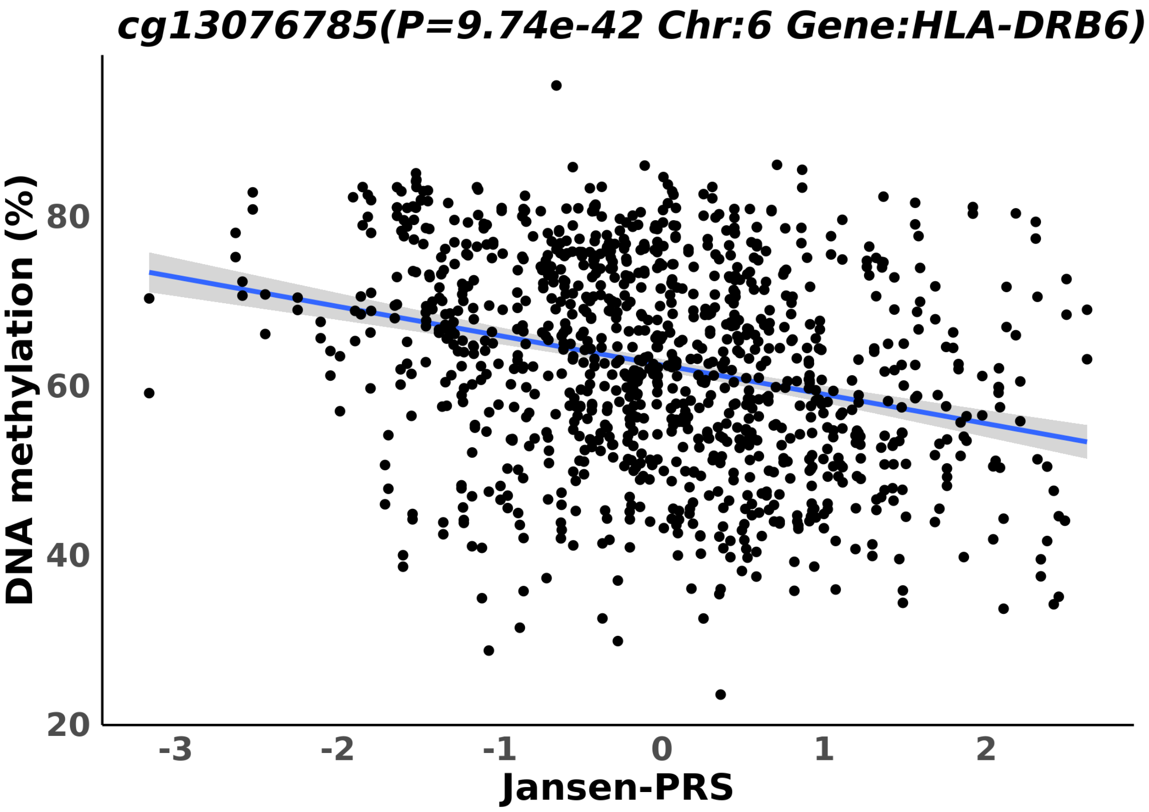


Figure 6.15: cg13076785 is hypomethylated with elevated PRS-Jansen in the cortex. *The x-axis represents the standardised (mean of 0, standard deviation=1) PRS-Jansen and the y-axis represents DNA-methylation adjusted for confounders. PRS- Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region.*

Table 6.11: Differentially methylated region associated with PRS-Jansen in the cortex. *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected p-value adjusted for the number of independent tests. PRS- Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region.*

Chr	BP start	BP end	n	Gene	Beta (%)	SE (%)	P	P adjusted
6	32551749	32551954	3	*HLA-DRB1*	-23.72	3.37	1.94e-12	1.60e-06

6.4.5.4 Multiple differentially methylated positions and regions were associated with PRS-Jansen$_{APOE}$

I identified two experiment-wide significant DMPs associated with PRS-Jansen$_{APOE}$ (see **Figure 6.10**). Both DMPs were significantly hypermethylated with elevated PRS (see **Figure 6.11; Table 6.12**). Although there was slight inflation of p-values, as shown by the quantile-quantile plots (λ =1.44; see **Figure 6.12**), the test statistics were normally distributed suggesting my results are not affected by bias. The average absolute magnitude of effect for the significant DMPs per SD increase in LOAD PRS was 1.35% (inter-quartile range [IQR] = 1.33-1.37%). To identify if any of these DMPs reflect a direct *cis*-genetic effect on DNAm, I utilised the cortex mQTL results generated in Chapter 5 (section **5.4.3**). There was evidence of a relationship between a PRS-mQTL and DNAm for both DMPs (cg14123992 and cg20051876), indicating they are under direct genetic control at the SNP level. Both DMPs were annotated to the same gene – *APOE*– and were hypermethylated with elevated PRS, with a 1.31-1.39% increase in DNAm per SD increase in PRS respectively (p=6.56e-09; p=3.47e-08; see **Figure 6.16** for plot of DNAm against PRS at cg14123992).

I identified eight DMRs (38% hypermethylated) which were associated with PRS-Jansen$_{APOE}$ (see **Table 6.13**). Of these, seven were annotated to genes including *APOE* (chromosome 19), SSBP3 (chromosome 1), *LOC100130872; LOC100130872-SPON2* (chromosome 4), *CCDC102B*; TMX3 (chromosome 18), *STAT5A* (chromosome 18), GPR68 (chromosome 17), and *KLHL33* (chromosome 14). *CCDC102B* is of interest in the context of neurodegenerative disease and has been associated with frontotemporal lobar degeneration (Andrés-Benito *at al.*, 2019). Folate deficiency has been associated with AD risk and *STAT5A* has been shown to be differentially expressed when there is dysregulated folate metabolism (Li *at al.*, 2016). The remaining DMRs are annotated to genes which have not previously been implicated in AD and currently little is known about their function. *KLHL33* was also identified in the cortex PRS-Kunkle$_{APOE}$ EWAS.

Table 6.12: Differentially methylated positions associated with the PRS-Jansen$_{APOE}$ in the cortex at experiment wide significance (p< 9E− 8). *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. Assoc. mQTL = associated with a methylation quantitative trait loci (mQTL) which was included in the PRS. PRS-Jansen$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region.*

CpG	Chr	BP	Gene	Gene Region	Beta (%)	SE (%)	P	Assoc. mQTL
cg14123992	19	45407868	*APOE*	TSS1500	1.39	0.21	6.56e-09	mQTL
cg20051876	19	45407860	*APOE*	TSS1500	1.31	0.20	3.47e-08	mQTL

Table 6.13: Differentially methylated regions associated with PRS-Jansen$_{APOE}$ in the cortex. *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected P value adjusted for the number of independent tests. PRS- JansenAPOE = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS including the APOE region.*

Chr	BP start	BP end	n	Gene	Beta (%)	SE (%)	P	P adjusted
19	45407860	45407945	3	*APOE*	20.66	2.82	2.50e-13	2.25e-07
1	54821853	54822103	5	*SSBP3*	-6.37	1.05	1.25e-09	1.12e-03
4	1202509	1203104	16	*LOC100130872;* *LOC100130872-SPON2*	-4.36	0.72	1.28e-09	1.16e-03
18	66382226	66382465	5	*CCDC102B;TMX3*	-5.68	0.94	1.48e-09	1.33e-03
17	40439132	40439492	3	*STAT5A*	5.03	0.83	1.72e-09	1.55e-03
14	91720173	91720578	9	*GPR68*	-4.68	0.78	1.80e-09	1.62e-03
5	132155233	132155460	4	-	-5.66	0.98	7.70e-09	6.93e-03
14	20903410	20904169	11	*KLHL33*	7.42	1.33	2.32e-08	2.09e-02

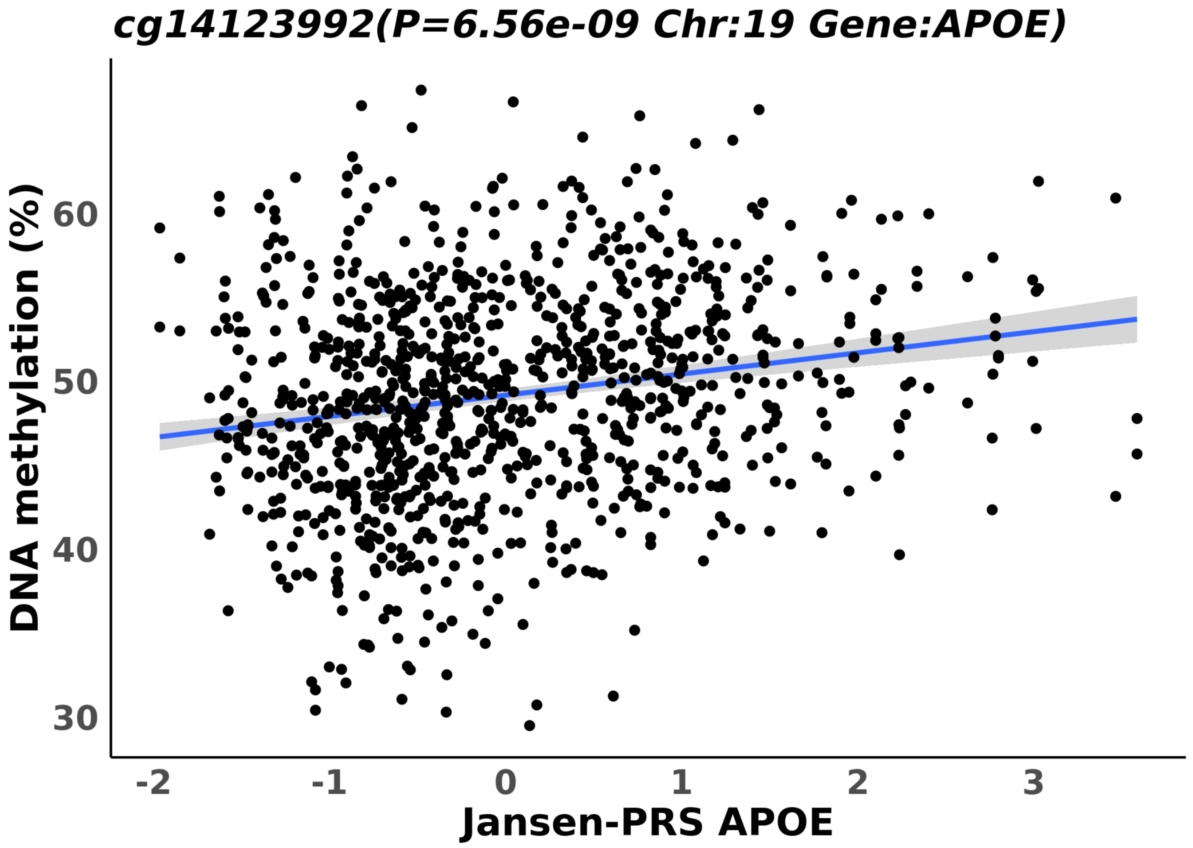


Figure 6.16: cg14123992 is hypermethylated with elevated PRS-Jansen$_{APOE}$ in the cortex. *The x-axis represents the standardised (mean of 0, standard deviation=1) PRS-Jansen$_{APOE}$ and the y-axis represents DNA-methylation adjusted for confounders. PRS-Jansen$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS including the APOE region.*

6.4.5.5 The effect sizes were generally consistent across the four cortex PRS EWAS

Although there was no direct overlap in the significant DMPs in the cortex associated with each of the AD PRS, the direction of the effect was consistent across analyses (as evaluated by a binomial sign-test and Pearson's correlation) of the DMPs reaching suggestive significance (p<5e-05) across the cortex PRS EWAS (see **Figure 6.17**). The most significant effects were between the PRS-Kunkle$_{APOE}$ and PRS-Jansen$_{APOE}$ EWAS (sign test p=5.1e-144; see **Figure 6.17**). These results suggest that regardless of LOAD PRS, similar effects on DNAm occur, suggesting they are not just reflecting mQTLs as different variants were included in the PRS. Interestingly, there was evidence for opposite directions of effect between the PRS-Kunkle$_{APOE}$ and the PRS-Jansen (see **Figure 6.17**), which is also supported by the non-significant sign test. This may be a result of the differing combinations of SNPs included from each GWAS in addition to the inclusion of the *APOE* region.

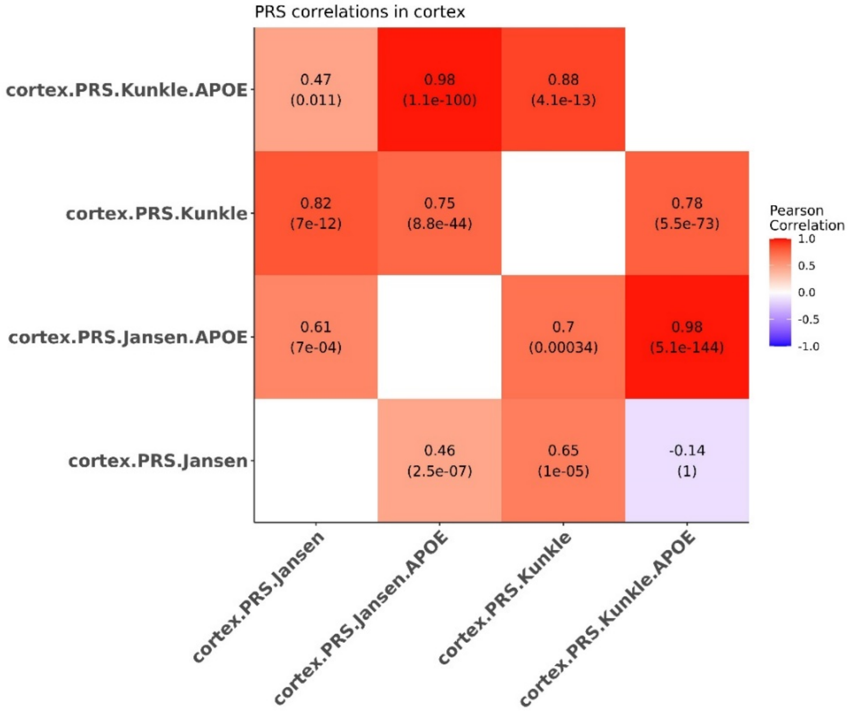


Figure 6.17: Comparing the effect sizes of DMPs reaching suggestive significance (p<5e-05) in each LOAD-PRS EWAS in cortex. *PRS.Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS excluding the APOE region. PRS.Kunkle.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS including the APOE region. PRS.Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region. PRS.Jansen.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS including the APOE region. Cortex = cortex EWAS. Colour represents the strength of the Pearson correlation. The numbers inside the squares represent the Pearson's correlation coefficient and the binomial sign test p-values which are within parentheses.*

6.4.6 Methylomic variation is associated with LOAD PRS in whole blood

I next explored if methylomic variation in whole blood is also associated with LOAD polygenic burden by undertaking an EWAS of variable PRS in eight whole blood cohorts totalling 6,106 unrelated European samples (see section **6.3.1** for more details on the cohorts). PRS were calculated using two recent LOAD GWAS (Kunkle et. al. (2019) and Jansen *at al.* (2019)). To assess the effects of including *APOE* in the PRS on DNAm, PRS were calculated both including and excluding the *APOE* locus. Linear regressions were then run against genome-wide DNAm controlling for age, sex, batch, smoking score and derived cell proportions. The results were meta-analysed using the IVW method (see section **6.3.6**).

To identify if any DMPs identified reflect a direct *cis*-genetic effect on DNAm, I utilised the whole blood mQTL results generated in Chapter 5 (see sections **5.4.2**), using the same method as described above in **6.4.6.**

6.4.6.1 Multiple differentially methylated positions and regions were associated with PRS-Kunkle in whole blood

I identified 27 experiment-wide significant DMPs (Bonferroni p<1.03e-07) in whole blood associated with PRS-Kunkle (see **Figure 6.18**). 16 (59%) of the DMPs were significantly hypermethylated with elevated PRS (see **Table 6.14** and **Figure 6.19**) and the other 11 (41%) were hypomethylated with elevated PRS; see **Table 6.14** and **Figure 6.19**). The direction of effects were consistent across cohorts as shown by the heterogeneity statistics (see columns Q and I^2 in **Table 6.14**) with the majority of the Q statistics (calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies) being >0.05 and I^2 (which describes the percentage of variation across studies that is due to heterogeneity) being below 50%. There was no evidence of p-value inflation in this analysis, as shown by the quantile-quantile plots and the genomic inflation factor - lambda (λ) - being ≤ 1 (λ =0.84; see **Figure 6.20**). The average absolute magnitude of effect for the significant DMPs per SD in PRS-Kunkle was 1.18% (inter-quartile range [IQR] = 0.43-1.94%). To identify if any of these DMPs reflect a direct *cis*-genetic effect on DNAm, I utilised the whole blood mQTL results generated in Chapter 5 (see **5.4.2**). There was evidence of a relationship between a PRS-mQTL and DNAm for 13 of the DMPs, providing evidence that these sites are under direct genetic control at the SNP level. Of the 27

DMPs, 20 were annotated to genes and 11 of these were unique. The majority of the PRS-Kunkle associated DMPs are annotated to genes that are of relevance in the context of AD and associated neurobiological functions. These include:

- cg09580214 and cg20307385 which are located on chromosome 11, are annotated to *PSMC3*. Interestingly they have an opposite direction of effect where cg09580214 was significantly hypomethylated with elevated PRS-Kunkle, with a decrease in DNAm of 1.65% (p=1.28e-21) for every SD increase in PRS (see **Figure 6.22**), whereas cg20307385 was significantly hypermethylated with elevated PRS-Kunkle, with an increase of 0.21% (p=1.71e-15) for every SD increase in PRS. *PSMC3* is an AD GWAS gene, residing around the *SPI1* locus (nearest gene(s) identified to top GWAS loci in the Kunkle *at al.* (2019) GWAS) and is thought to mediate tau toxicity (Karch *at al.*, 2016). See section **6.4.5.1** for more details on *SPI1*.
- cg20796544 was which is located on chromosome 7 and annotated to *HIP1*, was significant hypermethylated with elevated PRS, with a 0.56% increase (p= 2.52e-09) in DNAm per SD increase in PRS. *HIP1* has been shown to be hypermethylated for multiple system atrophy – a fatal late onset neurodegenerative disease and similarly to Parkinson's and DLB it is a α-synucleinopathy (Bettencourt *at al.*, 2020). In addition cell studies suggest there is a role for *HIP1* for initiating apoptosis in the pathogenesis of Huntington's disease (Wanker, 2002).
- cg08168897 and cg02887598 located on chromosome 2 and annotated to *BIN1*. cg08168897 was hypermethylated with elevated PRS, with an 0.47% increase (p= 4.81e-18; **see Figure 6.22**) per SD in PRS and this concurs with previous evidence suggesting there is hypermethylation of *BIN1* in AD patients (De Jager *at al.*, 2014). cg02887598 has the opposite direction of effect, where it was hypomethylated with elevated PRS, with a 2.34% decrease (p=1.15e-11) per SD increase in PRS. Interestingly, increased expression of *BIN1* has been found to mediate AD risk by modulating Tau pathology.
- cg18774435 and cg02771260 which are located on chromosome 11 and annotated to *MS4A3*. Both sites were hypermethylated with elevated PRS, with cg02771260 having a threefold larger effect (0.97% increase per SD in PRS; p= 7.89e-11) in comparison the cg18774435 (0.3% increase per SD in PRS; p=

3.66e-12). *MS4A3* resides in the LD block of the nearest gene to a top AD loci (Jansen *at al.*, 2019; Kunkle *at al.*, 2019) and is part of the MS4A locus which contains several genes implicated in immune modulation. Increased *MS4A3* expression is associated with more advanced brain pathology in AD patients, and in addition elevated *MS4A3* levels in both blood and brain tissue have been associated with AD risk (Villegas-Llerena, Phillips, Garcia-Reitboeck, Hardy, & Pocock, 2016).

- cg05377527 which is located on chromosome 11 and is annotated to *PTPMT1*. cg05377527 was hypomethylated with elevated PRS with a -0.19% decrease (p= 6.11e-10) in DNAm per SD increase in PRS. *PTPMT1* expression is altered in AD brains (Karch *at al.*, 2016).
- cg20172563 which is located on chromosome 6, is annotated to *CD2AP*, was significantly hypermethylated with elevated PRS, with a 0.33% increase (p= 7.01e-09) in DNAm per SD increase in PRS. *CD2AP* is the nearest gene to a top AD GWAS loci (Jansen *at al.*, 2019; Kunkle *at al.*, 2019) and has been associated with AD phenotypes including neuritic plaque burden (Shulman *at al.*, 2013) and tau biomarkers (Ramos de Matos *at al.*, 2018). This DMP is associated with a PRS-mQTL.
- cg12568536 which is located on chromosome 11 and has no gene annotation based on proximity, however it resides within the vicinity (within 100kb) of *PICALM*, an AD GWAS gene involved in *APP* metabolism (Jansen *at al.*, 2019; Kunkle *at al.*, 2019).cg09139047, cg14645244, cg17369694, cg17369694, cg22933800, cg17416722, cg15708909, cg13423887, cg19575208 cg02919082 and cg18816397 are located on chromosome 6 and are annotated to genes within the HLA region (*HLA-DRB1*; *HLA-DRB5; HLA-DQA1l HLA-DQB1*). cg09139047 and cg14645244 have an mQTL association. See section **6.4.5.3** for more details on HLA.cg24672777 which is located on chromosome 11, is annotated to a gene that has not previously been implicated in dementia: *OR4C45*. cg24672777 was hypomethylated with elevated PRS, with a 0.52% decrease (p= 6.60e-09) in DNAm per SD increase in PRS. *OR4C45* is an olfactory receptor and has been associated with Aicardi syndrome, a severe neurodevelopmental disorder with a complex and largely unknown aetiology (Piras *at al.*, 2017).

In addition to identifying DMPs, I used the software tool *dmrff* to identify DMRs. I identified 10 regions (60% hypermethylated) using the PRS-Kunkle in whole blood (see **Table 6.15**). The majority of regions were annotated to genes identified in the DMP analysis including the HLA region, *PSMC3* and *BIN1*.

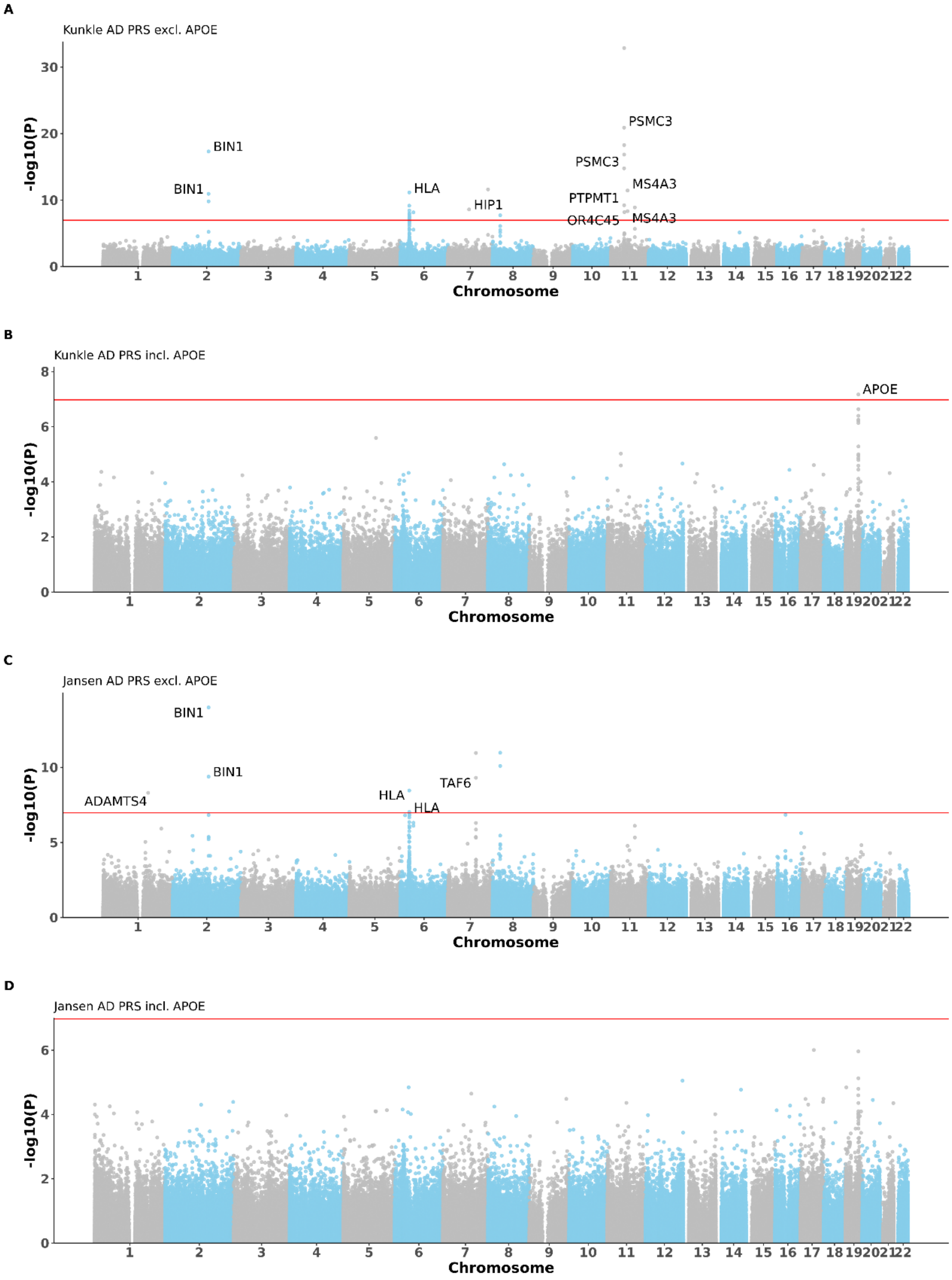


Figure 6.18: Whole blood EWAS of polygenic variation associated with LOAD highlights experiment-wide significant differentially methylated positions. *Manhattan plots showing results of EWAS meta-analyses of*

*polygenic burden for AD conducted in whole blood using **(A)** a PRS generated using the Kunkle at al. GWAS excluding variants in the APOE region, **(B)** a PRS generated using the Kunkle at al. GWAS including variants in the APOE region, **(C)** a PRS generated using the Jansen at al. GWAS excluding variants in the APOE region, and **(D)** a PRS generated using the Jansen at al. GWAS including variants in the APOE region. Significant differentially methylated positions are annotated to their proximal gene, unless they are unannotated. The X-axis shows the probe location across chromosomes 1-22 and the Y-axis shows EWAS significance (-log10(P)), with the horizontal red line representing an experiment wide significance threshold (p< 1.03e-07).*

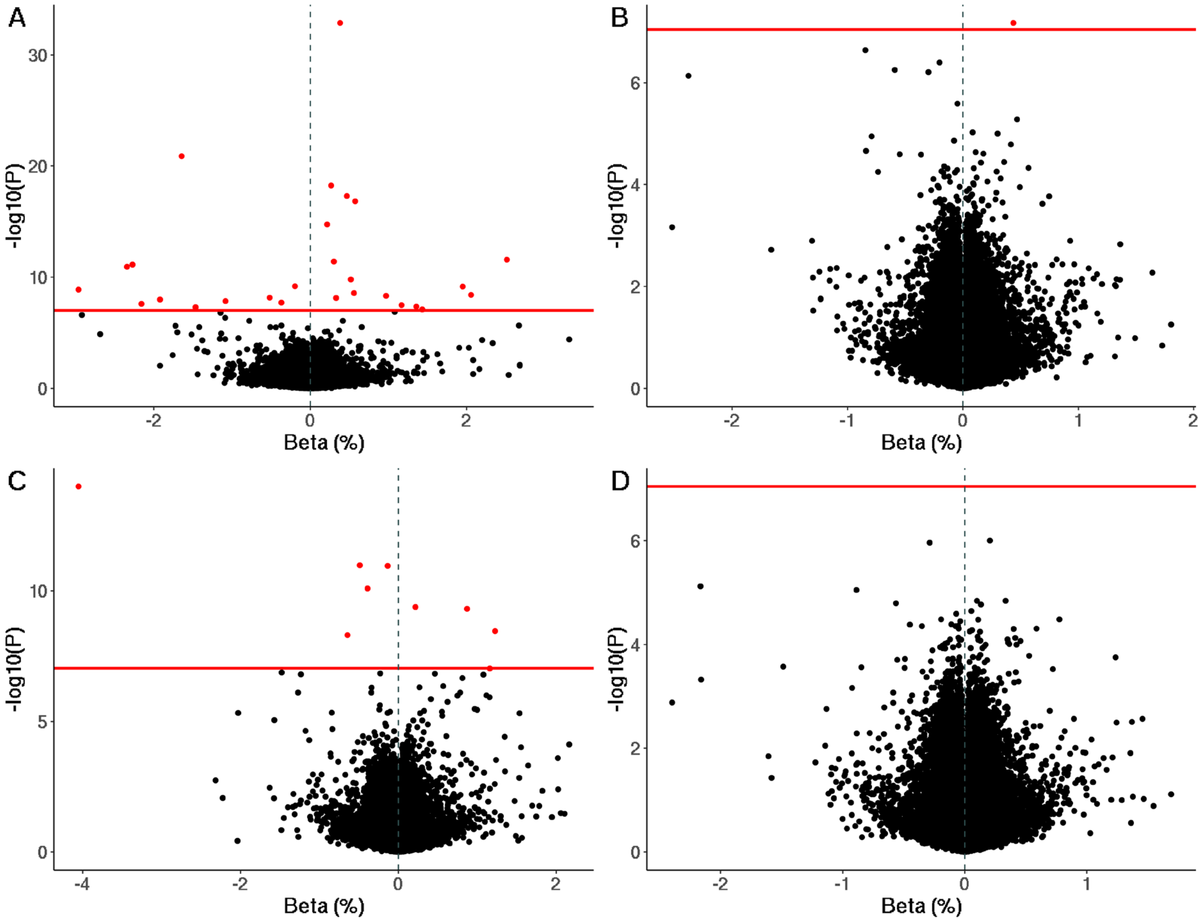


***Figure 6.19: Volcano plot of differentially methylated positions (DMPs) identified in the whole blood LOAD PRS EWAS meta-analyses.** DNA methylation was regressed against PRS where in **(A)** LOAD PRS was generated using the Kunkle at al. GWAS excluding APOE **(B)** LOAD PRS was generated using the Kunkle at al. GWAS including APOE; **(C)** LOAD PRS was generated using the Jansen at al. GWAS excluding APOE; and **(D)** LOAD PRS was generated using the Jansen at al. GWAS including APOE. The X-axis shows beta effect size (ES) and the Y-axis shows -log10(p). Red probes indicate a p-value that reaches experiment-wide significance (EWS) (p < 1.03e-07). LOAD = Late onset Alzheimer's disease. PRS = Polygenic risk scores.*

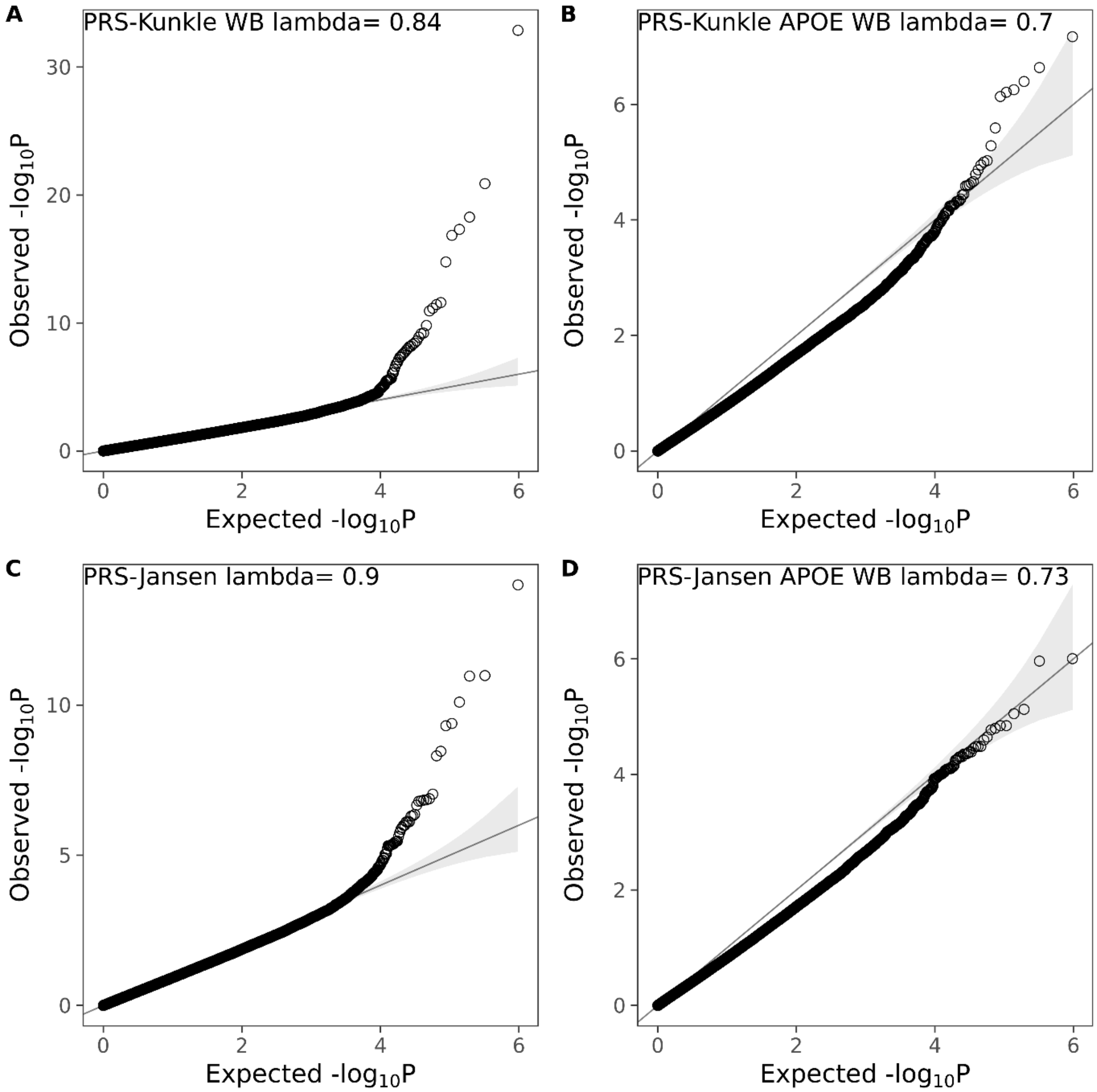


Figure 6.20: Quantile-quantile plots of p-values from AD PRS EWAS meta-analyses conducted in whole blood. *Shown are the expected (x-axis) against the observed (y-axis) quantiles in each EWAS of polygenic burden where **(A)** PRS were generated using the Kunkle at al. GWAS excluding APOE variants, **(B)** PRS were generated using the Kunkle at al. GWAS including APOE variants, **(C)** PRS were generated using the Jansen at al. GWAS excluding APOE variants, and **(D)** PRS were generated using the Jansen at al. GWAS including APOE variants. Lambda= genomic inflation factor.*

Table 6.14: Differentially methylated positions (DMPs) associated with the PRS-Kunkle in whole blood at experiment wide significance (p< 1.03e-07). *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. FE refers to statistics for the fixed effects meta-analysis model. RE refers to statistics from the random-effects meta-analysis model. Q = Cochran's Q – a measure of heterogeneity. I = the percentage of variation across studies that is due to heterogeneity. Assoc. mQTL = associated with a methylation quantitative trait loci (mQTL) which was included in the PRS. PRS-Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS excluding the APOE region.*

CpG	Chr	BP	Gene	Gene Region	Beta FE (%)	P FE (%)	Beta RE (%)	P RE (%)	Q	I	Assoc. mQTL
cg09580214	11	47448534	*PSMC3*	TSS1500	-1.64	8.34e-38	-1.65	1.28e-21	0.08	44.64	-
cg20307385	11	47447363	*PSMC3*	Body	0.20	2.44e-37	0.21	1.71e-15	0.01	64.57	mQTL
cg20135002	11	47629003	-	-	0.38	1.35e-33	0.38	1.35e-33	0.58	0	mQTL
cg05585544	11	47624801	-	-	0.26	7.17e-24	0.27	5.33e-19	0.26	21.34	mQTL
cg20796544	7	75253334	*HIP1*	Body	0.56	1.48e-18	0.56	2.52e-09	0.09	50.32	-
cg08168897	2	127865431	*BIN1*	TSS1500	0.47	4.81e-18	0.47	4.81e-18	0.89	0	mQTL
cg18512352	11	47633146	-	-	0.57	1.42e-17	0.57	1.42e-17	0.64	0	mQTL
cg06223080	2	127868745	-	-	0.48	1.67e-15	0.52	1.57e-10	0.11	39.98	mQTL
cg09139047	6	32552042	*HLA-DRB1*	Body	-2.28	2.60e-14	-2.27	7.12e-12	0.30	17.93	mQTL
cg12568536	11	85873778	-	-	-2.98	3.03e-14	-2.96	1.27e-09	0.19	34.62	-
cg16618979	7	143108841	-	-	2.51	2.52e-12	2.51	2.52e-12	0.41	0	-
cg18774435	11	59838540	*MS4A3*	3'UTR	0.30	3.66e-12	0.30	3.66e-12	0.44	0	mQTL
cg14645244	6	32552205	*HLA-DRB1*	Body	-1.93	4.22e-12	-1.92	1.00e-08	0.21	30.12	mQTL
cg02887598	2	127841945	*BIN1*	Body	-2.34	1.15e-11	-2.34	1.15e-11	0.71	0	-
cg05377527	11	47586666	*PTPMT1*	TSS1500	-0.20	4.62e-11	-0.19	6.11e-10	0.35	10.56	mQTL
cg02771260	11	59836817	*MS4A3*	Body	0.97	7.89e-11	0.97	4.62e-09	0.28	18.53	mQTL
cg17369694	6	32485396	*HLA-DRB5*	3'UTR	1.95	2.84e-10	1.95	6.90e-10	0.38	4.18	-
cg22933800	6	32605704	*HLA-DQA1*	Body	2.06	4.95e-10	2.05	3.67e-09	0.36	8.65	-
cg17416722	6	32554385	*HLA-DRB1*	Body	1.36	4.96e-10	1.36	4.31e-08	0.28	21.24	-
cg15708909	6	32487314	*HLA-DRB5*	Body	-1.48	4.99e-10	-1.47	4.82e-08	0.28	20.99	-
cg13423887	6	32632694	*HLA-DQB1*	Body	-2.17	6.26e-10	-2.16	2.43e-08	0.30	17.52	-
cg20172563	6	47487173	*CD2AP*	Body	0.33	6.66e-10	0.33	7.01e-09	0.33	12.5	mQTL
cg13879655	8	27450777	-	-	-0.37	1.11e-09	-0.37	1.92e-08	0.31	14.87	mQTL
cg24672777	11	48374446	*OR4C45*	TSS1500	-0.52	6.60e-09	-0.52	6.60e-09	0.72	0	-

cg19575208	6	32551888	*HLA-DRB1*	Body	-1.08	1.17e-08	-1.08	1.33e-08	0.40	0.73	-
cg02919082	6	32605694	*HLA-DQA1*	Body	1.16	3.08e-08	1.16	3.08e-08	0.43	0	-
cg18816397	6	32489555	*HLA-DRB5*	Body	1.43	7.45e-08	1.43	7.45e-08	0.63	0	-

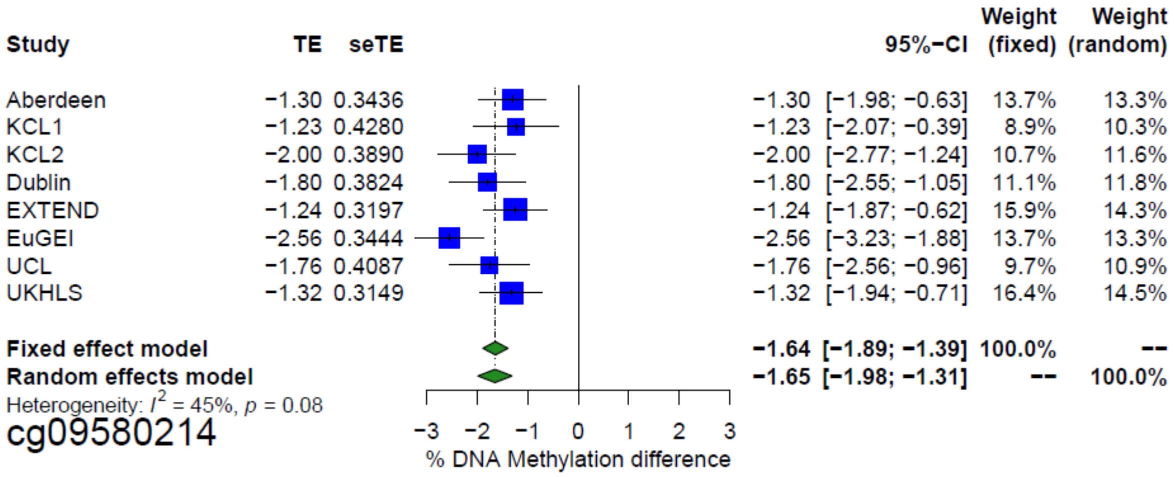


Figure 6.21: A forest plot of the most significant DMP identified in the PRS-Kunkle EWAS meta-analysis in whole blood (cg09580214 located at chr11:47448534 and annotated to PSMC3, p = 1.28e-21). *Across all studies PRS-Kunkle was associated with hypomethylation at this DNA methylation site. The X-axis shows the beta effect size (% DNA methylation difference per SD increase in LOAD PRS), with squares representing effect size and arms indicating the 95% confidence intervals. PRS-Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle et al. GWAS excluding the APOE region.*

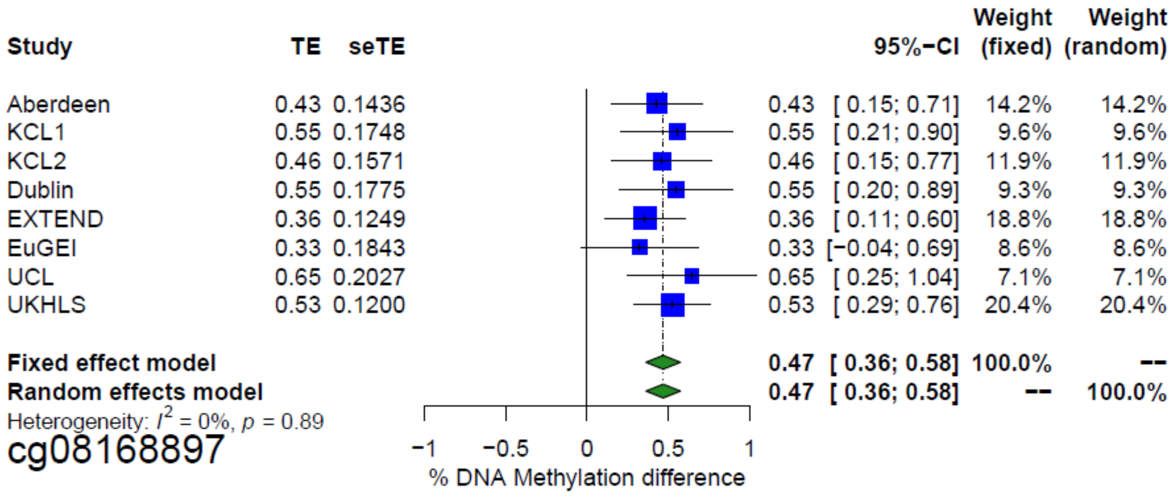


Figure 6.22: A forest plot of a significant DMP identified in the PRS-Kunkle$_{APOE}$ EWAS in whole blood (cg08168897:2:127865431: BIN1, p = 4.81e-18). *Across all studies PRS-Kunkle was associated with hypermethylation at this DNA methylation site. The X-axis shows the beta effect size (% DNA methylation difference per SD increase in LOAD PRS), with squares representing effect size and arms indicating the 95% confidence intervals. PRS-Kunkle$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS including the APOE region.*

Table 6.15: Differentially methylated regions associated with the PRS-Kunkle in a whole blood EWAS meta-analysis. *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected P value adjusted for the number of independent tests. PRS- Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS excluding the APOE region.*

Chr	**BP start**	**BP end**	**n**	**Gene**	**Beta (%)**	**SE (%)**	**P**	**P adjusted**
6	32551749	32552453	15	*HLA-DRB1*	-9.89	0.07	0	0
6	32632565	32633157	18	*HLA-DQB1*	-8.12	0.09	0	0
6	32728862	32729442	11	*HLA-DQB2*	5.44	0.13	0	0
6	32606385	32607509	4	*HLA-DQA1*	12.31	0.2	0	0
11	47448223	47448534	4	*PSMC3*	-17.27	0.55	1.97e-220	9.27e-215
6	32632331	32632338	3	*HLA-DQB1*	8.38	0.35	3.88e-126	1.82e-120
6	32526342	32526414	3	*HLA-DRB6*	-9.64	0.45	2.48e-100	1.17e-94
6	32609094	32609212	4	*HLA-DQA1*	7.26	0.52	7.51e-44	3.53e-38
6	32629786	32629955	4	*HLA-DQB1*	4.88	0.66	1.94e-13	9.13e-08
2	127865248	127865431	3	*BIN1*	11.92	1.8	3.71e-11	1.75e-05

6.4.6.2 One differentially methylated position and multiple differentially regions were associated with PRS-Kunkle$_{APOE}$ in whole blood

I identified a single experiment-wide significant (p<1.03e-07) DMP associated PRS-Kunkle$_{APOE}$ (see **Table 6.16**). This DMP - cg06750524 - was significantly hypermethylated with elevated PRS (p=6.74e-08; see **Figure 6.18**) and is annotated to *APOE*; for every SD increase in PRS there was a 0.44% increase in DNAm (see **Table 6.16**). The direction of effect was consistent across studies (see **Figure 6.18**), although there was evidence of some heterogeneity (Q=0.02; I^2=59.43; see **Table 6.16**). There was no evidence of statistical inflation in this analysis, as shown by the quantile-quantile plots (λ =0.7; see **Figure 6.20**). No other DMPs were identified in this analysis suggesting that *APOE* diminishes the majority of effects from the other genetic variants included in the PRS due to its large effect. There was evidence of a relationship between a PRS-mQTL and DNAm at this site.

In the DMR analysis PRS-Kunkle$_{APOE}$ was associated with 5 regions (40% hypermethylated; see **Table 6.17**). Some of the genes identified by the DMR analysis are relevant in the context of AD and other neurobiological functions:

- The HLA region (chromosome 6) has been associated with AD and is an AD risk locus (Jansen *at al.*, 2019; Kunkle *at al.*, 2019) (see section **6.4.6.1** above for more details). This association suggests that although *APOE* dominates the PRS when it is included, some of the other genetic effects still persist. The association with HLA suggests that the none-*APOE* variants included in the PRS still influence DNAm, although these effects are attenuated when *APOE* is included.
- *APOC1* (chromosome 19) is in LD with the *APOE* region and is known to is known to facilitate dementia onset under oxidative stress (Prendecki *at al.*, 2018).

The remaining DMRs with a gene annotation have not been previously implicated in AD or any other any neurological phenotypes:

- *OR2V2* (chromosome 5) is an olfactory receptor involved in g-protein coupled receptor signalling. Although this specific olfactory gene has not been implicated in AD, high copy number variations in a cluster of olfactory receptors on chromosome 14 have been associated with a younger AD age of onset

(Shaw *at al.*, 2011) supporting the theory that olfactory receptors play a role in the aetiology of AD.

- *D2HGDH* (chromosome 2) regulates alpha-ketoglutarate levels and dioxygenase function and has been associated with a rare neuro-metabolic disorder - D2HGA1 - caused by pathogenic variants in this gene (Lin *at al.*, 2015).

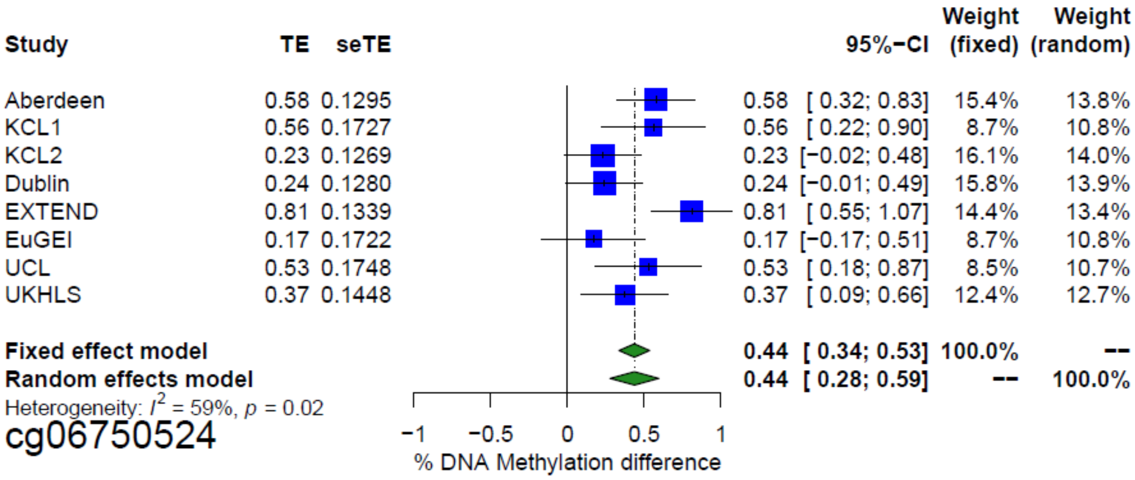


Figure 6.23: A forest plot of the significant DMP identified in the PRS- Kunkle$_{APOE}$ EWAS in whole blood (cg06750524:19:45409955: APOE, p = 6.74e-08). *Across all studies PRS-Kunkle was associated with hypomethylation at this DNA methylation site. The X-axis shows the beta effect size (% DNA methylation difference per SD increase in LOAD PRS), with squares representing effect size 22 and arms indicating the 95% confidence intervals. DMP= differentially methylated position. PRS- Kunkle$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle et al. GWAS including the APOE region.*

Table 6.16: Differentially methylated position associated with the PRS-Kunkle$_{APOE}$ in whole blood at experiment wide significance (p< 1.03e-07). *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. . FE refers to statistics for the fixed effects meta-analysis model. RE refers to statistics from the random-effects meta-analysis model. Q = Cochran's Q – a measure of heterogeneity. I = the percentage of variation across studies that is due to heterogeneityAssoc. mQTL = associated with a methylation quantitative trait loci (mQTL) which was included in the PRS.. PRS- Kunkle$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS including the APOE region.*

CpG	Chr	BP	Gene	Gene Region	Beta FE (%)	P FE (%)	Beta RE (%)	P RE (%)	Q	I	Assoc. mQTL
cg06750524	19	45409955	*APOE*	Body	0.44	1.20e-17	0.44	6.74e-08	0.02	59.43	mQTL

Table 6.17: Differentially methylated regions associated with PRS-Kunkle$_{APOE}$ in whole blood. *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected P value adjusted for the number of independent tests. PRS- Kunkle$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS including the APOE region.*

Chr	BP start	BP end	n	Gene	Beta (%)	SE (%)	P	P adjusted
6	32489991	32490421	3	*HLA-DRB5*	5.33	0.04	0	0
6	32552095	32552453	5	*HLA-DRB1*	-5.1	0.16	7.21e-234	3.34e-228
19	45417587	45418020	5	*APOC1*	-11.66	0.38	3.35e-209	1.55e-203
5	180581301	180581761	4	*OR2V2*	4.9	0.31	2.58e-56	1.20e-50
2	242692882	242693323	3	*D2HGDH*	5.38	0.46	1.65e-31	7.67e-26

6.4.6.3 Multiple differentially methylated positions and regions were associated with PRS-Jansen in whole blood

I identified nine experiment-wide significant (p<1.03e-07) DMPs associated with PRS-Jansen (see **Figure 6.18**). Four (44%) of the DMPs were significantly hypermethylated with elevated PRS and the other five (66%) were hypomethylated (see **Table 6.18**). There was no evidence of statistical inflation in this analysis, as shown by the quantile-quantile plots (λ =0.9; see **Figure 6.11**). There was an average absolute magnitude of effect for the significant DMPs per SD in PRS-Jansen was 1.02% (inter-quartile range [IQR] = 0.39-1.16). To identify if any of these DMPs reflect a direct *cis*-genetic effect on DNAm, I utilised the whole blood mQTL results generated in Chapter 5. There was evidence of a relationship between a PRS-mQTL and DNAm for six of the DMPs, providing evidence that these sites are under direct genetic control at the SNP level. Several of the PRS-Jansen associated DMPs are of relevance in the context of AD. These include:

- cg02887598 and cg00436254 located on chromosome 2 and annotated to *BIN1* – an AD risk gene identified by GWAS and EWAS (for more details on *BIN1* see section **6.4.6.1**). cg02887598 was hypomethylated with elevated PRS with a 4.05% decrease (p= 1.00e-14) in DNAm per SD increase in PRS. cg00436254 was hypermethylated with elevated PRS with a 0.21% increase (p= 4.10e-10) in DNAm per SD increase in PRS and has previously been associated with an mQTL. *BIN1* is as AD risk gene (De Jager *at al.*, 2014; Jansen *at al.*, 2019; Kunkle *at al.*, 2019) and was also identified in the PRS-Kunkle whole blood analysis (see section **6.4.6.1** for more details).
- cg23472400 located on chromosome 7, is annotated to *TAF6* and has previously been associated with an mQTL. Cg23472400 was hypermethylated with elevated PRS, with a 0.87% increase (p= 4.88e-10) in DNAm per SD increase in PRS (see **Figure 6.24**). *TAF6* has been identified as a potential AD risk gene which influences transcriptional regulation (Jones *at al.*, 2010) although its role in neurodegeneration is largely unknown.
- cg24329783 located on chromosome 1, is annotated to *ADAMTS4* and has previously been associated with an mQTL. cg24329783 was hypomethylated with elevated PRS, with a 0.65% decrease in DNAm per SD increase in PRS (see **Figure 6.25**). *ADAMTS4* is the nearest gene to an AD GWAS loci (Jansen

at al., 2019; Marioni *at al.*, 2018). It is highly expressed in the brain and predominately in neurons. It is a proteinase involved in the cleaving of reelin, which has been shown to aggregate and lead to the depositing of Aβ, tau phosphorylation, and neurofibrillary tangle formation in the hippocampus (Kocherhans *at al.*, 2010).

- cg06204447 and cg26331067 located on chromosome 6 and annotated to the HLA region (specifically *HLA-DRB1* and *HLA-DRB6*, respectively). Both were significantly hypermethylated with increasing PRS, with a ~1.2% increase in DNAm per SD increase in PRS. More details on HLA have been describe previously.

I identified 13 DMRs (54% hypermethylated) associated with PRS-Jansen. Six of these were annotated to the HLA region. Four other DMRs were annotated to genes located proximal to the HLA cluster including *HCG27* (chromosome 6), *HCG4P6* (chromosome 6), *LOC285830* (chromosome 6) and *TAP2* (chromosome 6). *TAP2* genotype has been associated with AD in *APOE*-ε4 carriers (Bullido *at al.*, 2007). There were two other DMRs annotated to genes: *C1orf10* (chromosome 1), and *WDR60* (chromosome 7). *WDR60* is a gene involved in the neurodevelopmental processes, and been shown to be differentially expressed (upregulated) in the cortex of AD patients (Sun, Yang, Sun, Li, & Duan, 2019). *C1orf10* encodes a member of the fused gene family of proteins and is involved in cell proliferation (Imai *at al.*, 2005) and has not previously been implicated in AD or neurodegenerative diseases.

6.4.7 Two differentially methylated regions are associated with PRS-Jansen$_{APOE}$ in whole blood

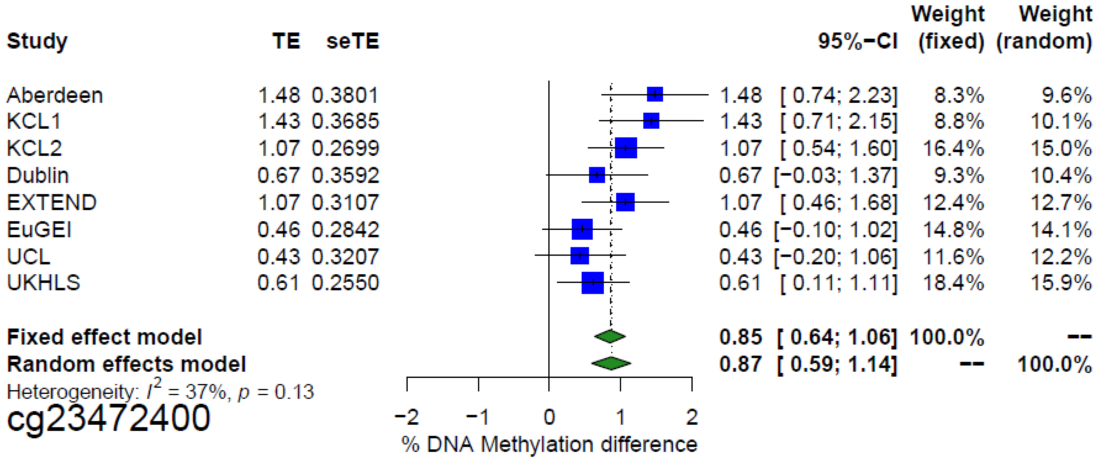
No DMPs were associated with PRS-Jansen$_{APOE}$ at experiment wide significance. I identified two DMRs associated with PRS-Jansen$_{APOE}$ (see **Table 6.20**). One DMR was hypomethylated with elevated PRS and was annotated to *APOC1* (chromosome 19) – a gene which is in the LD block of *APOE* (see section **6.4.6.2** for more details). The other DMR was hypermethylated with elevated PRS and was annotated to *OR2V2* (chromosome 5), an olfactory receptor, which was also identified as a DMR associated with PRS-Kunkle$_{APOE}$ (see section **6.4.6.2** for more details).

Table 6.18: Differentially methylated positions associated with the PRS-Jansen in whole blood at experiment wide significance (p<1.03e-07). *Information is provided corresponding to chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. FE refers to statistics for the fixed effects meta-analysis model. RE refers to statistics from the random-effects meta-analysis model. Q = Cochran's Q – a measure of heterogeneity. I = the percentage of variation across studies that is due to heterogeneityAssoc. mQTL = associated with a methylation quantitative trait loci (mQTL) which was included in the PRS.. PRS-Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region.*

CpG	Chr	BP	Gene	Gene Region	Beta FE (%)	P FE (%)	Beta RE (%)	P RE (%)	Q	I	Assoc. mQTL
cg02887598	2	127841945	*BIN1*	Body	-4.13	1.48e-33	-4.05	1.00e-14	0.06	56.43	-
cg23472400	7	99704848	*TAF6*	3'UTR	0.85	7.98e-15	0.87	4.88e-10	0.13	37.02	mQTL
cg22906224	7	99728672	-	-	-0.14	1.08e-13	-0.14	1.09e-11	0.31	15.13	mQTL
cg04803944	8	27450844	-	-	-0.49	1.05e-11	-0.49	1.05e-11	0.72	0	mQTL
cg24329783	1	161160887	*ADAMTS4*	3'UTR	-0.64	1.65e-11	-0.65	4.92e-09	0.24	23.9	mQTL
cg06204447	6	32546665	*HLA-DRB1*	3'UTR	1.17	1.68e-11	1.16	9.24e-08	0.18	35.71	-
cg00436254	2	127862614	*BIN1*	Body	0.21	3.78e-11	0.21	4.10e-10	0.34	11.15	mQTL
cg13879655	8	27450777	-	-	-0.39	8.00e-11	-0.39	8.00e-11	0.81	0	mQTL
cg26331067	6	32522535	*HLA-DRB6*	Body	1.23	1.65e-10	1.22	3.46e-09	0.33	13.18	-

Table 6.19: Differentially methylated regions associated with PRS-Jansen in the whole blood. *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected p-value adjusted for the number of independent tests. PRS- Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region.*

Chr	BP start	BP end	n	Gene	Beta (%)	SE (%)	P	P adjusted
6	32362407	32362744	9	-	-7.6	0.22	1.05e-256	4.97e-251
6	32725270	32725438	4	*HLA-DQB2*	-6.81	0.28	4.73e-129	2.23e-123
6	29701563	29702509	3	*LOC285830*	3.61	0.15	7.03e-125	3.32e-119
6	32632565	32632694	4	*HLA-DQB1*	-3.86	0.16	1.27e-123	5.99e-118
6	32525813	32526702	7	*HLA-DRB6*	9.66	0.42	4.23e-117	2.00e-111
6	33048444	33048919	16	*HLA-DPB1*	7.04	0.37	1.53e-81	7.24e-76
6	32797476	32797578	3	*TAP2*	-3.53	0.24	1.95e-48	9.18e-43
1	38156462	38156652	4	*C1orf109*	3.82	0.43	5.96e-19	2.81e-13
7	158712251	158712291	3	*WDR60*	-4.25	0.52	2.57e-16	1.21e-10
6	29893944	29894228	9	*HCG4P6*	4.44	0.59	6.35e-14	3.00e-08
6	32632731	32632957	7	*HLA-DQB1*	6.13	0.96	1.74e-10	8.22e-05
6	31166504	31166799	3	*HCG27*	-8.67	1.41	9.03e-10	4.26e-04
6	32728862	32729174	8	*HLA-DQB2*	6	1.03	5.25e-09	2.48e-03

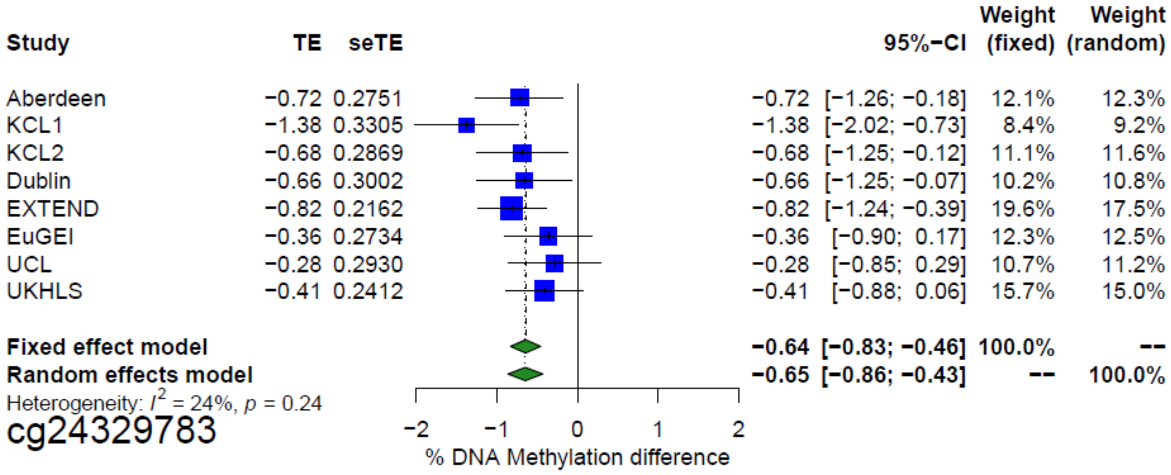


Study	TE	seTE	95%−CI	Weight (fixed)	Weight (random)
Aberdeen	1.48	0.3801	1.48 [0.74; 2.23]	8.3%	9.6%
KCL1	1.43	0.3685	1.43 [0.71; 2.15]	8.8%	10.1%
KCL2	1.07	0.2699	1.07 [0.54; 1.60]	16.4%	15.0%
Dublin	0.67	0.3592	0.67 [−0.03; 1.37]	9.3%	10.4%
EXTEND	1.07	0.3107	1.07 [0.46; 1.68]	12.4%	12.7%
EuGEI	0.46	0.2842	0.46 [−0.10; 1.02]	14.8%	14.1%
UCL	0.43	0.3207	0.43 [−0.20; 1.06]	11.6%	12.2%
UKHLS	0.61	0.2550	0.61 [0.11; 1.11]	18.4%	15.9%
Fixed effect model			**0.85 [0.64; 1.06]**	**100.0%**	**--**
Random effects model			**0.87 [0.59; 1.14]**	**--**	**100.0%**

Heterogeneity: $I^2 = 37\%$, $p = 0.13$

cg23472400

Figure 6.24: A forest plot of a significant DMP identified in the PRS-Jansen EWAS in whole blood (cg23472400:7:99704848:TAF6, p = 1.00e-14). *Across all studies PRS-Jansen was associated with hypermethylation at this DNA methylation site. The X-axis shows the beta effect size (% DNA methylation difference per SD increase in LOAD PRS), with squares representing effect size and arms indicating 95% confidence intervals. PRS-Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region.*



Study	TE	seTE	95%−CI	Weight (fixed)	Weight (random)
Aberdeen	−0.72	0.2751	−0.72 [−1.26; −0.18]	12.1%	12.3%
KCL1	−1.38	0.3305	−1.38 [−2.02; −0.73]	8.4%	9.2%
KCL2	−0.68	0.2869	−0.68 [−1.25; −0.12]	11.1%	11.6%
Dublin	−0.66	0.3002	−0.66 [−1.25; −0.07]	10.2%	10.8%
EXTEND	−0.82	0.2162	−0.82 [−1.24; −0.39]	19.6%	17.5%
EuGEI	−0.36	0.2734	−0.36 [−0.90; 0.17]	12.3%	12.5%
UCL	−0.28	0.2930	−0.28 [−0.85; 0.29]	10.7%	11.2%
UKHLS	−0.41	0.2412	−0.41 [−0.88; 0.06]	15.7%	15.0%
Fixed effect model			**−0.64 [−0.83; −0.46]**	**100.0%**	**--**
Random effects model			**−0.65 [−0.86; −0.43]**	**--**	**100.0%**

Heterogeneity: $I^2 = 24\%$, $p = 0.24$

cg24329783

Figure 6.25: A forest plot of a significant DMP identified in the PRS-Jansen EWAS in whole blood (cg24329783:1:161160887:ADAMTS4). *Across all studies PRS-Jansen was associated with hypomethylation at this DNA methylation site. The X-axis shows the beta effect size (% DNA methylation difference per SD increase in LOAD PRS), with squares representing effect size and arms indicating 95% confidence intervals. PRS-Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region.*

Table 6.20: Differentially methylated regions associated with PRS-Jansen$_{APOE}$ in the whole blood. *Listed for each probe is the chromosomal location (h19/GRCh37 genomic annotation) and the Illumina UCSC gene annotation. BP start = base position the region begins. BP end = base position where the region ends. N = number of probes in the region. P adjusted = Bonferroni corrected p-value adjusted for the number of independent tests. PRS- Jansen$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS including the APOE region.*

Chr	BP start	BP end	n	Gene	Beta (%)	SE (%)	P	P adjusted
19	45417587	45418020	5	*APOC1*	-11.76	0.54	5.85e-107	2.72e-101
5	180581301	180581761	4	*OR2V2*	6.12	0.54	6.29e-30	2.93e-24

6.4.7.1 Effect sizes at individual DMPs are generally consistent across the four whole blood PRS EWAS meta-analyses

Of the experiment wide significant DMPs identified in each of the four whole blood analyses, two directly overlapped: cg02887598 (chr2: 127841945:*BIN1*) and cg13879655 (chr8:27450777) which were significant in both the PRS-Kunkle and PRS-Jansen EWAS. In addition to *BIN1*, there were several DMPs identified around the HLA locus in both analyses, suggesting an overlapping signal for this region of the genome. This hypothesis was further supported when comparing the direction of effect (as evaluated by a binomial sign-test and Pearson's correlation) of the DMPs reaching suggestive significance (p<5e-05) across the PRS EWAS (see **Figure 6.26**). The most consistent overlap of effects were between the PRS-Kunkle and PRS-Kunkle$_{APOE}$ EWAS (sign test p=3e-17; see **Figure 6.26**), however when tested the other way around, using the suggestive significant DMPs from the PRS-Kunkle$_{APOE}$, the sign test was not significant. This is expected since all the SNPs included in the PRS with *APOE* included all of the SNPs in the PRS without *APOE* but not vice versa. In addition, there were Bonferroni significant (sign test p<0.004) positive correlations for the majority of PRS EWAS analyses, and nominally significant (sign test p<0.05) positive correlations for all analyses with the exception of between the PRS-Kunkle$_{APOE}$ and the PRS-Jansen. These results support the notion that a similar genetic signal is being captured in the PRS calculated using either GWAS (see **Figure 6.26**). However, there is evidence of some differences which could be a result of the slightly differing combinations of genetic variants in each PRS. Similarly to the cortex, there was evidence for opposite directions of effect between the PRS-Kunkle$_{APOE}$ and the PRS-Jansen (see **Figure 6.26**).

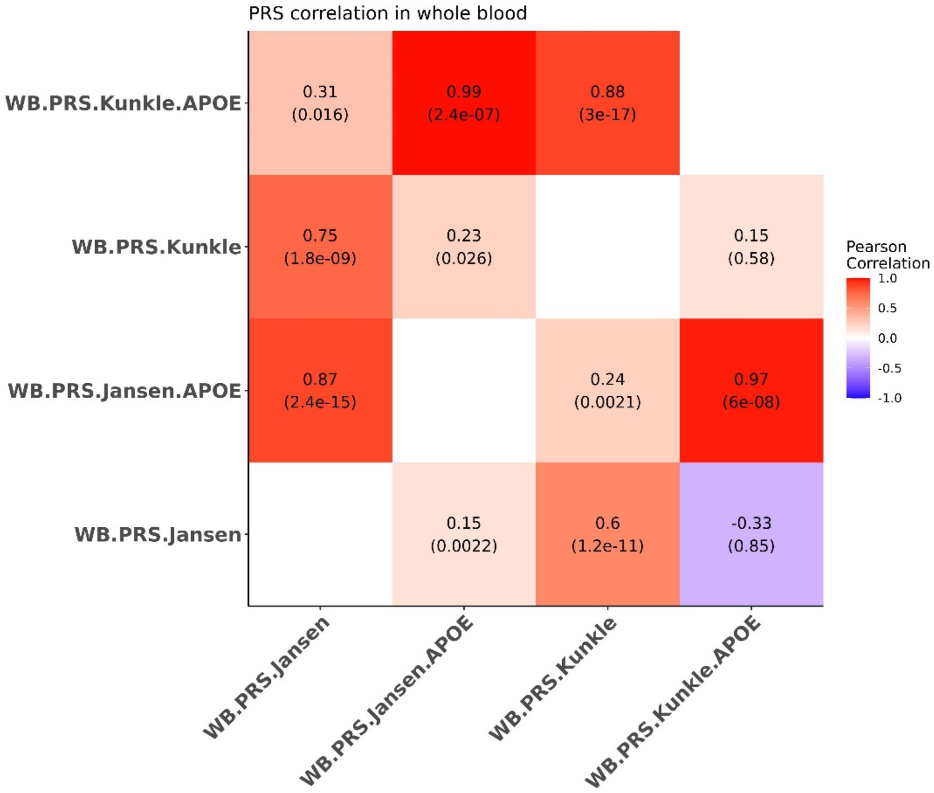


Figure 6.26: Comparing the effect sizes of DMPs reaching suggestive significance (p<5e-05) in each LOAD-PRS EWAS in whole blood (WB). *PRS.Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS excluding the APOE region. PRS.Kunkle.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS including the APOE region. PRS.Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region. PRS.Jansen.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS including the APOE region. WB = whole blood EWAS. Colour represents the strength of the Pearson correlation. The numbers inside the squares represent the Pearson's correlation coefficient and the binomial sign test p-values which are within parentheses.*

6.4.8 There are tissue specific effects when comparing results across whole brain and cortex

Interestingly, the associations with PRS were stronger in the cortex than the blood (see **Figure 6.26**), with more DMPs yielding suggestive significance in the cortex. This might be due to tissue specific effects in the cortex, which is the area primarily affected in AD and the fact the cortex samples are from a dementia cohort so are likely to have higher genetic loading for AD.

Across the significant DMPs identified in the blood and cortex analyses there was a direct overlap of two loci from the PRS-Kunkle whole blood and the PRS-Jansen cortex analyses: cg09139047 (6:32552042) and cg14645244 (6:32552205), which are both annotated to *HLA-DRB1* and were hypomethylated with elevated PRS. Across tissues, multiple DMPs and DMRs annotated to the HLA region, or cluster genes proximally located within this region, further supporting the hypothesis that immune dysfunction is associated with AD. In addition, in the PRS EWAS excluding *APOE* there was evidence for *cis* effects of the genome-wide AD variants on chromosome 11, most notably around the *SPI1* GWAS locus. In contrast, when *APOE* was included in the PRS it seemed to reduce the effects of these variants in the PRS, and signals on chromosome 11 and 6 were diminished in both tissues.

When comparing the overall direction of effect of the suggestive significant probes (sign test p<5e-05) in whole blood against those same probes in the cortex, there was little consistency (**Figure 6.27**). The PRS-Jansen$_{\mathrm{APOE}}$ whole blood and the PRS-Jansen cortex were the only analyses characterised by the same direction of effect (sign test p=0.0037). Similarly, when comparing the direction of effect of the suggestive significant probes (p<5e-05) in cortex against those same probes in whole blood there was little evidence for enrichment of the same direction of effect (**Figure 6.27**). However, the PRS-Jansen cortex and both the PRS-Jansen whole blood (sign test p=0.00018), and PRS-Kunkle whole blood (sign test p=0.0025) were characterised by the same direction of effect. These results suggest that despite some specific overlaps, PRS may have different effects on DNAm in blood and cortex.

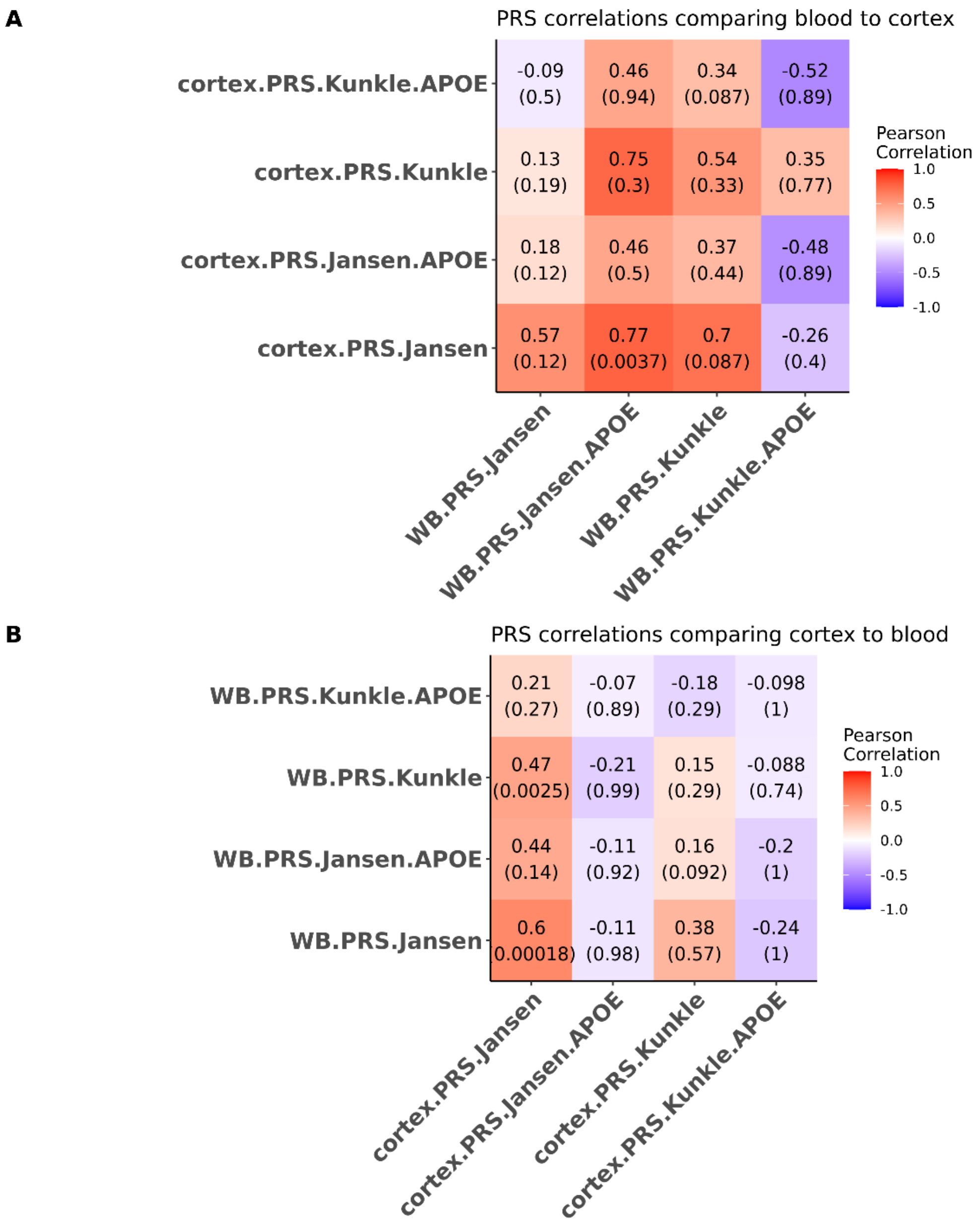


***Figure 6.27: Comparing the effect sizes of DMPs reaching suggestive significance (p<5e-05) in each LOAD-PRS EWAS across tissues (whole blood and cortex). (A)** Effect sizes of top 100 DMPS for each WB EWAS compared to the effect sizes of those same probes in the cortex EWAS. **(B)** Effect sizes of top 100 DMPS for each cortex EWAS compared to the effect sizes of those same probes in the WB EWAS. PRS.Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS excluding the APOE region. PRS.Kunkle.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS including the APOE region. PRS.Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region. PRS.Jansen.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS including the APOE region. WB = whole blood EWAS. Cortex = cortex EWAS. Colour represents the strength of the Pearson correlation. The numbers inside the squares represent the Pearson's correlation coefficient and the binomial sign test p-values which are within parentheses.*

6.4.9 There are tissue specific effects when comparing results to Braak NFT Stage and Alzheimer's disease case-control EWAS

LOAD PRS was significantly associated with AD status (see **Figure 6.6**) and direct measures of AD neuropathology (see section **6.4.4**). Therefore, we can hypothesise that there may be consistent patterns of differential DNAm when comparing the results from the Braak NFT stage EWAS conducted in **Chapter 4** to the PRS EWAS results described in this Chapter. The PRS EWAS and Braak NFT-EWAS were compared using a binomial sign-test and Pearson's correlation including probes reaching suggestive significance (p<5e-05).

Differential DNAm associated with AD PRS in the cortex was characterised by the same direction of effect as Braak NFT stage in the cortex (see **Figure 6.28** and **Figure 6.29**). The strongest association with Braak NFT stage was identified between PRS-Kunkle$_{APOE}$ and (sign test p=8.1e-131), followed by the PRS-Jansen$_{APOE}$ (sign test p=2.1e-88). All cortex PRS EWAS were positively correlated with Braak NFT stage (see **Figure 6.28** and **Figure 6.29**). However, the PRS-Jansen had a non-significant sign test with Braak NFT stage, although when examined vice versa, the DMPs were enriched for the same direction of effect (sign test p=0.003). There was no evidence for Braak NFT stage associated differential DNAm in the cortex at sites found to be associated with AD PRS in whole blood (see **Figure 6.28** and **Figure 6.29**).

I compared all eight sets of EWAS results generated in this Chapter to a case-control EWAS of three AD cohorts conducted in whole blood (Nabais *at al.*, 2021). The correlations with the PRS EWAS were generally weak, around 0 (see **Figure 6.30** and **Figure 6.31**), and the only the evidence of enrichment for the same direction of effect at a nominal p-value (p<0.05) was with the whole blood PRS-Jansen$_{APOE}$ (sign test p=0.016; see **Figure 6.30**). However, the correlations of each whole blood PRS EWAS to the whole blood AD EWAS were generally positive, whereas the relationship with cortex was negative (see **Figure 6.30**), mirroring what I found in the cortex analyses above. Of note, the whole blood PRS EWAS were conducted across none-AD cohorts whereas the cortex EWAS was conducted using the same dementia cohort which may explain some of the differences in the results and the strengths of the associations; the samples are not independent. Again, these results indicate there may be tissue specific effects driven by PRS.

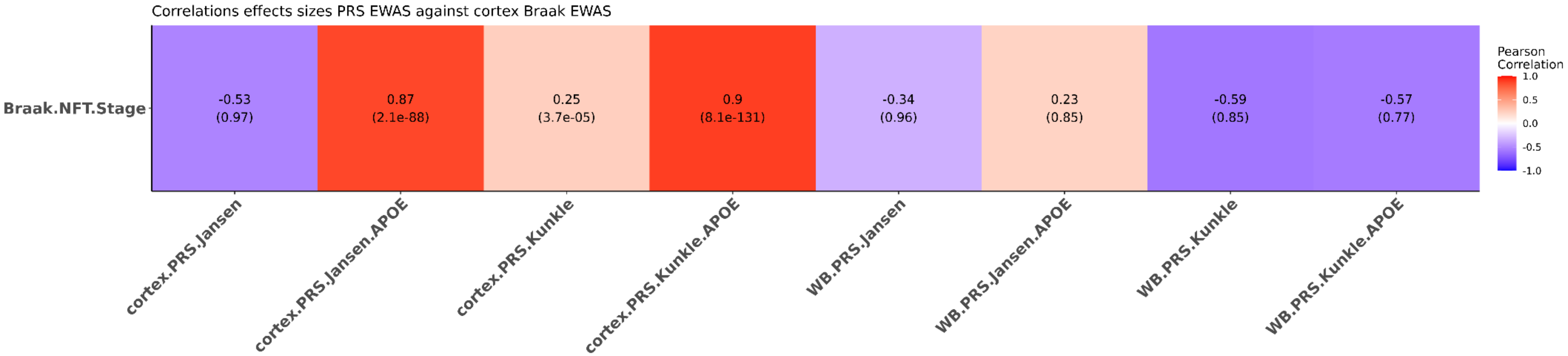


Figure 6.28: Comparing the effect sizes Of DMPs reaching suggestive significance (p<5e-05) in each LOAD-PRS EWAS to the cortex Braak neurofibrillary tangle stage EWAS results. *PRS.Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS excluding the APOE region. PRS.Kunkle.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS including the APOE region. PRS.Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region. PRS.Jansen.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS including the APOE region. WB = whole blood EWAS. Cortex = cortex EWAS. Colour represents the strength of the Pearson correlation. The numbers inside the squares represent the Pearson's correlation coefficient and the binomial sign test p-values which are within parentheses.*

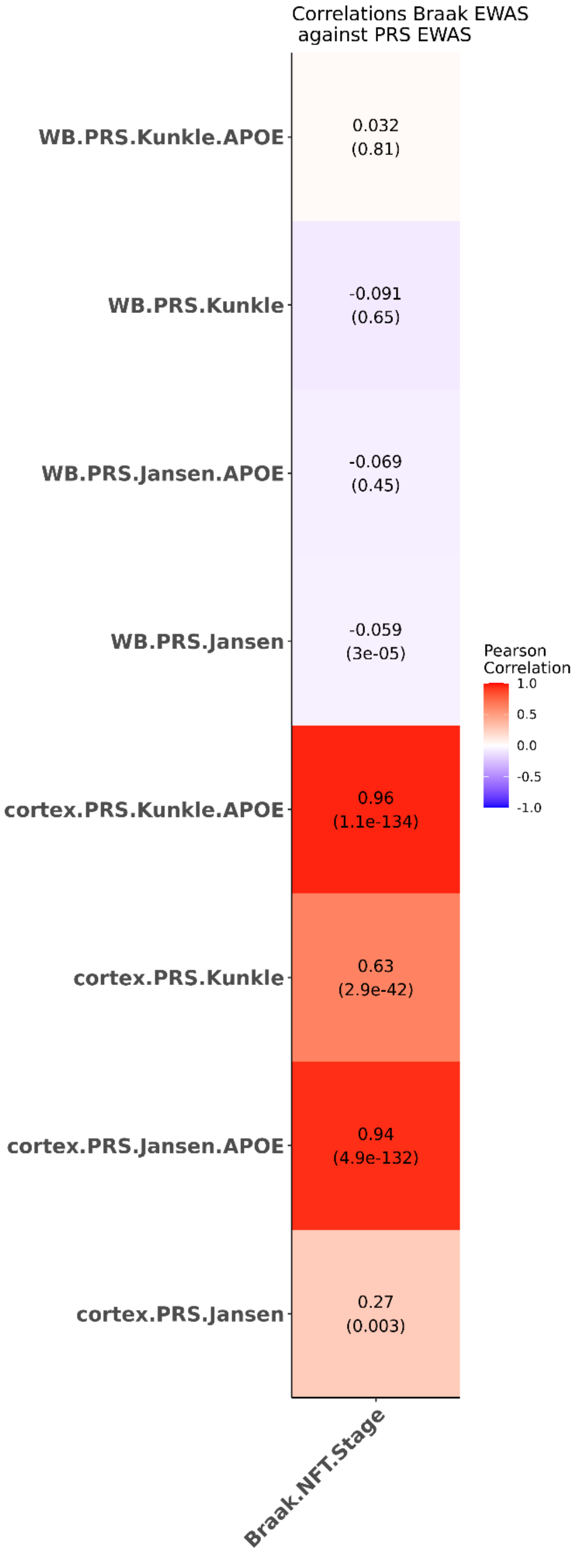


Figure 6.29: Comparing the effect sizes of the top 100 cortex Braak neurofibrillary tangle stage EWAS results to the effects sizes of those DMPS in each LOAD-PRS EWAS. *PRS.Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle et al. GWAS excluding the APOE region. PRS.Kunkle.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle et al. GWAS including the APOE region. PRS.Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen et al. GWAS excluding the APOE region. PRS.Jansen.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen et al. GWAS including the APOE region. WB = whole blood EWAS. Cortex = cortex EWAS. Colour represents the strength of the Pearson correlation. The numbers inside the squares represent the Pearson's correlation coefficient and the binomial sign test p-values which are within parentheses.*

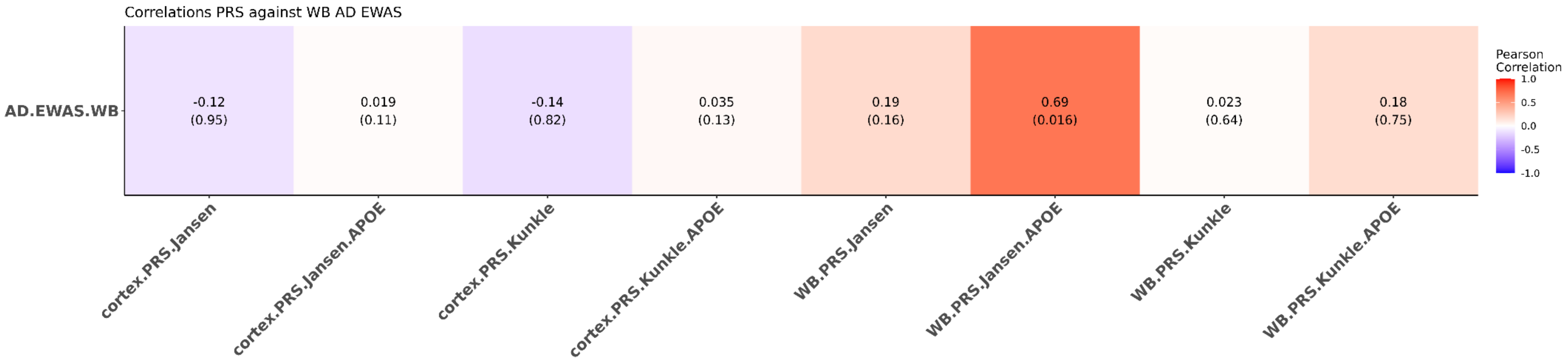


Figure 6.30: Comparing the effect sizes Of DMPs reaching suggestive significance (p<5e-05) in each LOAD-PRS EWAS to a whole blood AD case-control EWAS results. *PRS.Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS excluding the APOE region. PRS.Kunkle.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle at al. GWAS including the APOE region. PRS.Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS excluding the APOE region. PRS.Jansen.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen at al. GWAS including the APOE region. WB = whole blood EWAS. Cortex = cortex EWAS. Colour represents the strength of the Pearson correlation. The numbers inside the squares represent the Pearson's correlation coefficient and the binomial sign test p-values which are within parentheses*

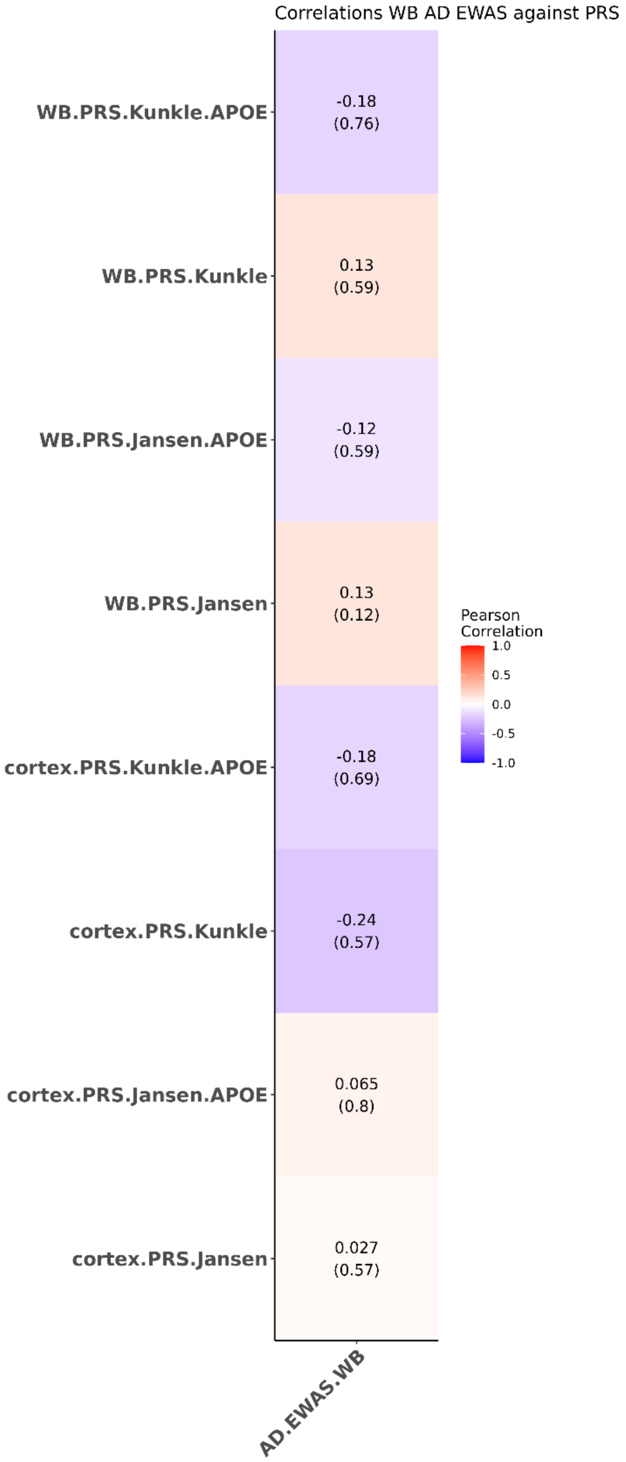


Figure 6.31: Comparing the effect sizes of the top 100 whole blood AD case-control EWAS results to the effects sizes of those DMPS in each LOAD-PRS EWAS. *PRS.Kunkle = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle et al. GWAS excluding the APOE region. PRS.Kunkle.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Kunkle et al. GWAS including the APOE region. PRS.Jansen = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen et al. GWAS excluding the APOE region. PRS.Jansen.$_{APOE}$ = Late onset Alzheimer's disease (LOAD) polygenic risk scores (PRS) calculated using the Jansen et al. GWAS including the APOE region. WB = whole blood EWAS. Cortex = cortex EWAS. Colour represents the strength of the Pearson correlation. The numbers inside the squares represent the Pearson's correlation coefficient and the binomial sign test p-values which are within parentheses.*

6.5 Discussion

6.5.1 Overview of results

In this chapter I aimed to investigate the association between polygenic burden for AD and variable DNAm in the cortex and whole blood. To my knowledge, this represents the largest EWAS of polygenic risk across tissues for any disease and is the first study to look at the effects of PRS on DNAm in LOAD. LOAD PRS were generated using two publicly available GWAS (Jansen *at al.*, 2019; Kunkle *at al.*, 2019). To assess the effects of the inclusion of *APOE* in a PRS, PRS were generated both including and excluding the *APOE* region. The optimal p-value thresholds for the PRS were based on definitive AD status derived from neuropathology. The variance explained by the PRS on AD was marginally higher than previous estimates (Escott-Price *at al.*, 2015). However, I utilised more recent AD GWAS which included additional variants and studies have demonstrated that AD-PRS performs best (i.e. correctly predicts cases/ controls) when applied to pathologically confirmed LOAD cases (Valentina Escott-Price, Myers, Huentelman, & Hardy, 2017). I identified multiple DMPs and DMRs associated with polygenic burden for AD across both tissues. Many of the PRS-associated loci are annotated to genes which have previously implicated in the aetiology of AD and other neurodegenerative diseases.

The genetic analysis against neuropathology indicates that *APOE* influences AD disease neuropathology via two independent pathways, one where Aβ accumulation correlates with the development of tauopathy, and a second pathway with direct effects on NFTs independent of β-amyloidosis. The relationship between common genetic variants associated with AD and neuropathology is more complex, with each individual variant potentially having a different effect on neuropathology and cognition. Taken together, these results provide insights into how the symptoms of AD manifest and how genetic risk factors influence the development of pathology.

In the PRS EWAS excluding the *APOE* locus, several loci in the HLA locus were identified to be differentially methylated across both tissues. The HLA locus is located on chromosome 6 and encodes several molecules that have a major role in innate immunity and has been implicated in genetic studies of AD (Jansen *at al.*, 2019; Karch *at al.*, 2016; Kunkle *at al.*, 2019). In addition, there were consistent signals across tissues on chromosome 11 around the MS4A and *SPI1* GWAS loci (Kunkle *at al.*, 2019), including DMPs annotated to *PSMC3* and *C1QTNF4*. Several genes within the *SPI1* locus are highly correlated with one another and have been associated with AD status (Karch *at al.*, 2016). Coupled with the HLA findings, these results implicate genes which are highly expressed in microglia, astrocytes or other myeloid cell types in AD pathogenesis, further supporting the involvement of immune related pathways in AD.

On the other hand, when *APOE* was included in the PRS analyses it diminished the associations on chromosome 6 and 11, with the predominant associations being driven by *cis* effects of the *APOE* region, which is supported by the associations of SNPs included in the PRS and the DMPs (mQTLs). Genome-wide Complex Trait Analysis (GCTA) (Yang, Lee, Goddard, & Visscher, 2011) - a software tool that can estimate the proportion of phenotypic variance explained by all genome-wide SNPs for complex traits – indicated that SNPs explain 53% of the variation in AD (Ridge *at al.*, 2016). 16% of the variation in AD was explained by known variants, and 13% is attributed to *APOE* (Ridge *at al.*, 2016); *APOE* is a major locus in the PRS so it is unsurprising that the effects differ drastically when it is included. Since the study by Ridge and colleagues (2016) was published several larger AD GWAS have been conducted (Bellenguez *at al.*, 2020; de Rojas *at al.*, 2020; Jansen *at al.*, 2019; Kunkle *at al.*, 2019; Wightman *at al.*, 2020), and therefore more of the variance in AD is likely explained by known GWAS SNPs and GCTA should be conducted on these datasets to validate this claim. Due to the large influence *APOE* has on DNAm when included the PRS, I recommend the *APOE* region is removed from PRS analyses, and where possible *APOE* genotype should be incorporated separately to assess the effect it has on methylomic variation independently of other common genetic variants.

Given my previous findings that there is an enrichment of mQTLs amongst genetic variants associated with AD (see Chapter 5 section **5.4.6**), I investigated if any of the PRS-associated DMPs resulted directly from these associations, utilising the results generated in **Chapter 5**. I only considered instances where mQTLs were included in the PRS or were in high LD with these SNPs. Several PRS-associated DMPs had a PRS-mQTL association, suggesting that direct genetic *cis* effects may be driving the methylomic changes at these sites. However, not all the DMPs were associated with a PRS-mQTL, and therefore our data suggest that PRS-associated variation at these loci may be a consequence of the combined effects of multiple genetic variants associated with AD. Of note, there may not have been the power to detect an mQTL at these sites and therefore we cannot definitively make this conclusion.

There was overlap in the genes annotated to DMPs identified in the analyses of cortex and whole blood (e.g. *APOE* and HLA), which suggests there are consistent effects of PRS on DNAm across tissue types. These data support the notion that whole blood may be a valid correlate of physiological processes in other tissues. However, there was also evidence of tissue heterogeneity and the associations were stronger in the cortex than whole blood. This might be due to tissue specific effects in areas primarily affected in AD (e.g. the cortex) in comparison to peripheral tissues (e.g. whole blood) or the fact a dementia cohort was used in the cortex; these samples are likely to have higher genetic loading than the whole blood non-AD cohorts. Together, these results suggest whole blood can be utilised to investigate LOAD PRS-associated DNAm variation, however there are limitations. Where possible multiple tissues should be considered to increase our understanding of disease pathogenesis.

Several of the DMPs and DMRs associated with genetic burden for AD were independent of the changes observed in the Braak NFT stage EWAS conducted in **Chapter 4**. However, there was an enrichment for the direction of effect with the cortex PRS EWAS results, indicating the results are not entirely independent from neuropathology. Of note, the cortex PRS EWAS and Braak NFT EWAS were both conducted utilising BDR, which is a dementia cohort, and this may partly explain the co-variation between these analyses.

6.5.2 Limitations

There are several limitations to consider within this Chapter. First, the use of bulk tissue is a potential confounder in DNAm studies (Guintivano, Aryee, & Kaminsky, 2013). Future studies could look at the associations between PRS and derived cell proportions and compare these to the bulk tissue results.

Although I used the two most recent publicly available LOAD-GWAS, several GWAS meta-analysis pre-prints have been recently published (Bellenguez *at al.*, 2020; de Rojas *at al.*, 2020; Schwartzentruber *at al.*, 2021; Wightman *at al.*, 2020). In addition, there is still a large proportion of the genetic component of AD which is unaccounted for, a concept known as “missing heritability” (Manolio *at al.*, 2009). Using larger and more comprehensive GWAS would increase the power of the PRS. However, it is worth noting all PRS used in this chapter significantly differentiated cases from controls, emphasising the robustness of the PRS. It is also worth considering that the majority of large AD GWAS have been conducted in European samples, and therefore all my analyses were limited to Europeans, reducing the population validity of my results as they may not be consistent for other ethnicities.

The PRS were not completely independent from the Braak NFT stage EWAS as they were conducted in the same sample. Therefore, there is a chance that the correlation between the two may confound the results. However, since genetic loading for AD is associated with Braak NFT stage this is to be expected and the overlap was small between the two, indicating there are independent effects.

6.5.3 Conclusion

This is the first study investigating DNAm changes associated with genetic burden for AD and is the first to incorporate non-disease specific blood cohorts, highlighting the utility of PRS for identifying molecular pathways associated with aetiological variation across multiple tissues. EWAS conducted using PRS are likely to be less confounded by environmental or disease driven factors such as smoking, medication or neuropathology since they only incorporate genetic factors which are unaffected by these. Although PRS in AD are currently not clinically useful due to the small differences between cases and controls, this Chapter indicates they can be used to aid our understanding of AD and identify potential therapeutic targets.

7 Discussion

In this PhD I have assessed the regulatory genomic processes (i.e. epigenomic and transcriptomic) involved in the aetiology of Alzheimer's disease (AD) across multiple tissues (two regions of the cortex and whole blood) utilising multi-omics methods. In this final discussion I present an overview of my primary findings and relate them to existing literature on genomic variation in AD. I also summarise the strengths, limitations and future directions of the research presented in my thesis.

7.1 Key findings

7.1.1 Recalibrating the epigenetic clock: Implications for assessing biological age in the human cortex

In **Chapter 3** I developed a novel epigenetic age model specifically for the human cortex - the cortical DNAm clock (DNAmClock$_{\text{Cortical}}$) - built using an extensive collection of DNA methylation (DNAm) data derived from >1000 human cortex samples. The model dramatically outperforms existing DNAm-based biomarkers for age prediction in data derived from the human cortex. I demonstrated that previous epigenetic clocks systematically underestimate age in older samples and do not perform optimally in human cortex tissue. I established that the age distribution and tissue type of samples included in training datasets for epigenetic clocks are important to consider when building and applying epigenetic clock algorithms to human epidemiological or disease cohorts. Additionally, the lack of association between accelerated ageing estimated using the DNAmClock$_{\text{Cortical}}$ and neuropathological measures in the BDR cohort indicate that although 1st generation epigenetic clocks are good predictors of age, they may not be directly associated with neuropathology. My findings suggest that previous associations between predicted DNAm age and neurodegenerative phenotypes may represent false positives resulting from suboptimal calibration of DNAm clocks for the tissue being tested and for phenotypes that manifest at older ages.

7.1.1 Epigenome wide association study of neuropathology in the Brains for Dementia Research cohort

In **Chapter 4** I examined the association between DNAm and five different neuropathology measures (Braak neurofibrillary tangle [NFT] stage, Thal phase, CERAD density, Braak Lewy body [LB] stage and TDP-43 status) utilising samples from two cortical regions from donors in the BDR study. I identified a number of differentially methylated positions (DMPs) and differentially methylated regions (DMRs) that were associated neuropathology. Many of these neuropathology associated-loci were annotated to genes which have previously been implicated in neurodegenerative diseases, including several annotated to the *HOXA* gene cluster (Gasparoni *at al.*, 2018; R. G. Smith *at al.*, 2018). Additionally, a number of novel loci which have not previously been implicated in dementia were identified, which warrant future exploration in neurodegenerative diseases. The *WGCNA* analysis identified several modules of co-methylated sites which were associated with neuropathology. These modules were enriched for genes involved in functional pathways implicated in AD such as in cholesterol metabolism (Picard *at al.*, 2018), gliogenesis (Rusznák, Henskens, Schofield, Kim, & Fu, 2016) oligodendrocyte differentiation (Desai *at al.*, 2010) and the immune response (Heppner, Ransohoff, & Becher, 2015). The findings from this Chapter indicate that methylomic variation associated with neuropathology may be a consequence of general neurodegeneration as opposed to being involved in specific neuropathological processes. This conclusion highlights the strength of the experimental design which incorporated numerous neuropathology measures, enabling us to better understand disease pathogenesis.

7.1.2 Methylation quantitative trait loci (mQTL) analysis and summary data-based Mendelian randomisation (SMR)

In **Chapter 5** I explored the genetic architecture of DNAm in both whole blood and the cortex. I identified associations between common genetic variants and DNAm sites (mQTLs), finding evidence of co-variation across tissues for a large proportion of sites. The mQTL databases were subsequently used to characterise the relationship between proximally located DNAm sites, and I identified numerous examples where neighbouring DNAm sites are genetically co-regulated with the same causal variant. This finding concurs with previous research which identified that multiple DNAm sites are influenced by overlapping variants (Hannon *at al.*, 2018; Liu *at al.*, 2014). The mQTLs were used for SMR analyses, which identified multiple situations where SNPs are pleiotropically associated with LOAD with evidence that this relationship is mediated by either DNAm, gene expression or both. Many of the prioritised genes were identified in the vicinity (within 250kb) of AD risk loci including several which were pleiotropically associated with both DNAm and gene expression such as *MS4A4A* (Deming *at al.*, 2019), *MS4A6A* (Deming *at al.*, 2019) and *CELF1* (Karch *at al.*, 2016), in whole blood and *C1QTNF4* (Rosenthal & Kamboh, 2014) in the cortex. A number of GO pathways were enriched among genes prioritised by SMR including lipid and cholesterol metabolism (Di Paolo & Kim, 2011; Jones *at al.*, 2010; Penke *at al.*, 2018), Aβ (Kunkle *at al.*, 2019; Sadigh-Eteghad *at al.*, 2015), tau (Kosik, Joachim, & Selkoe, 1986; Kunkle *at al.*, 2019), *APP* processing (Eggert, Thomas, Kins, & Hermey, 2018) and the innate immune response (Cao & Zheng, 2018; Jones *at al.*, 2010; Kunkle *at* *al.*, 2019). By integrating SNPs and DNAm with LOAD GWAS variants and gene expression I was able to explore the mechanisms underlying disease, advancing our understanding of the interaction between gene regulation and expression, and enabling the prioritisation of candidate genes involved in disease aetiology.

7.1.3 Methylomic variation associated with polygenic risk for Alzheimer's disease

In **Chapter 6** I examined the relationship between DNAm and polygenic risk burden for LOAD in both whole blood and the cortex. PRS were generated both including and excluding the *APOE* region using two recent GWAS datasets. When excluding *APOE* from the PRS, I identified a number of DMPs and DMRs which were associated PRS and many of these were annotated to genes which are relevant in the context of neurodegenerative disease including HLA (Jansen *at al.*, 2019; Karch *at al.*, 2016; Kunkle *at al.*, 2019), *PSMC3* (Karch *at al.*, 2016), *C1QTNF4* (Karch *at al.*, 2016), *MS4A3* (Villegas-Llerena, Phillips, Garcia-Reitboeck, Hardy, & Pocock, 2016) and *BIN1* (De Jager *at al.*, 2014). When *APOE* was included in the PRS it diminished the associations at these loci*,* with the predominant associations being driven by *cis* effects of the *APOE* locus itself. PRS-associated loci are largely independent of changes observed in the case-control EWAS of neuropathology presented in **Chapter 4**. However, there was an enrichment for consistent directions of effect with the cortex PRS EWAS results, indicating and overlap between differences associated with disease neuropathology. There was evidence for direct *cis* genetic influences on DNAm at several PRS-associated loci as identified by utilising the mQTL analyses presented in **Chapter 5**, although a number of associations were likely driven by the cumulative effect of the PRS. There was overlap in the genes annotated to DMPs identified in the analyses of cortex and whole blood (e.g. *APOE* and HLA), which suggests there are several consistent effects of PRS on DNAm across tissues. However, there was also evidence for tissue heterogeneity, potentially reflecting tissue specific effects between areas primarily affected in AD (e.g. the cortex) compared to peripheral tissues (e.g. whole blood). This research highlights the utility of PRS for identifying molecular pathways associated with aetiological variation.

7.2 Integration of AD findings

Over the past few decades, researchers have made progress in understanding the underlying biological mechanisms involved in the development of AD. Several pathways are hypothesised to be involved in the aetiology of AD including lipid and cholesterol metabolism (Di Paolo & Kim, 2011; Jones *at al.*, 2010; Penke *at al.*, 2018), Aβ (Kunkle *at al.*, 2019; Sadigh-Eteghad *at al.*, 2015), tau (Kosik *at al.*, 1986; Kunkle *at al.*, 2019) and *APP* processing (Eggert *at al.*, 2018), as well as an extensive role of the immune system (Cao & Zheng, 2018; Jones *at al.*, 2010; Kunkle *at al.*, 2019). Additionally, the genetic component of AD has been well established and genome wide-association studies (GWAS) have identified numerous variants which are robustly associated with disease (Bellenguez *at al.*, 2020; de Rojas *at al.*, 2020; Jansen *at al.*, 2019; Kunkle *at al.*, 2019; Lambert *at al.*, 2013; Schwartzentruber *at al.*, 2021; Wightman *at al.*, 2020). However, little is known about the functional mechanisms by which risk variants mediate disease susceptibility; as the majority of these variants do not index coding variants affecting protein structure they are hypothesised to influence gene regulation (Kikuchi *at al.*, 2019; Marzi *at al.*, 2018). To my knowledge, this thesis consists of the most comprehensive study of genomic, methylomic and transcriptomic variation across tissues in AD and represents an important contribution to the field by providing support for these hypotheses.

Across **Chapter 4-6** I identified a number of associations with genes and pathways implicated in immune regulation, lipid and cholesterol metabolism and Aβ, tau and *APP* processes. For example, in **Chapter 4** several DMPs and DMRs were annotated to key immune related genes (e.g. *TNFRSF1A* and *OSCAR*). In combination with the pathway analysis of the neuropathology-associated WGCNA modules highlighting an abundance of immune pathways (e.g. B and T cell processes and humoral immune response), these findings provide further evidence that immune regulation plays a role in the aetiology of AD and other dementias (Heppner *at al.*, 2015). Additionally, in **Chapter 5** and **6** several immune related genes were prioritised from the analyses including a number of loci annotated to HLA. The HLA locus encodes several molecules that have a major role in innate immunity and has been implicated in genetic studies of AD (Jansen *at al.*, 2019; Karch *at al.*, 2016; Kunkle *at al.*, 2019). Several other immune related genes were identified including *PSMC3* (Karch *at al.*, 2016), *C1QTNF4* (Karch *at al.*, 2016), *MS4A3* (Villegas-Llerena *at al.*, 2016), and *BIN1* (De Jager *at al.*, 2014). These results implicate genes which are highly expressed in microglia, astrocytes or other myeloid cell types in AD pathogenesis, further supporting the involvement of immune related pathways in AD. Several other AD associated biological pathways were implicated in **Chapter 5** including lipid related pathways such as cholesterol metabolism and transport. Lipids are involved in *APP* processing and trafficking and influence the formation of Aβ peptides which are involved in AD pathogenesis (Penke *at al.*, 2018). Furthermore, multiple amyloid and tau related pathways were identified including Aβ binding and tau binding pathways and a number loci were annotated to genes involved in these processes such as *MS4A4A*, *MS4A6A* and *CR1* (Cruchaga *at al.*, 2013; Deming *at al.*, 2019). My results provide further support for a role of lipid and cholesterol metabolism, Aβ, tau, *APP* processes and the immune system in the aetiology of AD.

Chapters 5 and 6 utilised different methods for understanding the relationship between genetics and methylomic variation, with both approaches supporting the hypothesis that GWAS variants are involved in gene regulation. Several prioritised loci were consistent across analyses including HLA, MS4A, *BIN1* and multiple loci on chromosome 11 where there are four AD GWAS associations (Bellenguez *at al.*, 2020; Kunkle *at al.*, 2019). These analyses indicate that both direct *cis* genetic effects and indirect cumulative polygenic effects are associated with methylomic variation and are involved in the aetiology of AD. The genetic effects were largely independent of the neuropathology EWAS results generated in **Chapter 4**. However, there was an enrichment for the direction of effect with the cortex PRS EWAS results, indicating genetically mediated methylomic changes in AD are not fully independent from neuropathology. Although the genetic-DNAm associations are unlikely to be confounded by disease course and medication intake, it is important to establish whether the neuropathology associated differences are causal and have a direct impact on disease aetiology. The breadth of associations between genetically mediated and neuropathology driven DNAm variation highlights the complexity of understanding the molecular pathways involved in the aetiology of AD and warrant future exploration.

The results from **Chapters 3, 5** and **6** highlight that some degree of inter-individual epigenetic variation is conserved across tissues. For example, in **Chapter 3** the DNAmClock$_{\text{Cortical}}$ outperformed existing DNAm-based biomarkers for age prediction in data derived from the human cortex. However, DNAm age predicted using the DNAmClock$_{\text{Cortical}}$ correlated with chronological age in whole blood, albeit with lesser accuracy. In **Chapter 5** the whole blood and cortex mQTL databases showed high overlap, suggesting that the majority of mQTL effects are concordant across tissues. This corroborates previous studies demonstrating that a relatively high proportion of mQTLs co-vary across tissues (Hannon *at al.*, 2016; Smith *at al.*, 2014). Additionally, a number of pleiotropic associations between DNAm, gene expression and LOAD identified by SMR were characterised by consistent signals at certain loci in both whole blood and the cortex. This pattern was also reflected in the results from **Chapter 6,** which provided evidence for consistent PRS-associated DNAm effects at some loci across tissues. Collectively, these results further support the notion that whole blood may be a valid correlate of physiological processes in other tissues for both genetically-mediated and age-associated sites. However, my analyses suggest that epigenetic clocks work optimally in the tissues that they are trained in, and additionally there was evidence for cortex-specific genetic effects at certain loci, indicating there is some regulatory heterogeneity between whole blood and cortex. This is accordance with research conducted by Hannon and colleagues (2015) suggesting that EWAS using whole blood for disorders where the brain is primarily affected may provide limited information relating to the underling pathological process. Since LOAD is a disease of the brain, this highlights the importance of utilising relevant tissue where possible, although it does not discount the utility of using a blood-based in epigenetic studies to identify potential biomarkers of neurodegenerative phenotypes.

7.3 Limitations and future directions

Several AD associated genes have been highlighted by the analyses presented within this thesis. Since the methods used throughout this thesis are hypothesis generating (i.e. providing statistical evidence for an association with different aspects of AD), it is unknown if the genes are biologically relevant. The next step would be to functionally validate these genes. This research could potentially be conducted using cell culture and genetic editing approaches such as CRISPR/cas9 (Hsu, Lander, & Zhang, 2014). This would be an interesting approach to take since CRISPR/cas9 has recently been applied to induce epigenetic changes, enabling the exploration of the effects variable DNAm has on the expression of a gene in specific cell lines (Liao *at al.*, 2017). Leading on from this, considering the dynamic and potentially reversible nature of DNAm, AD associated loci identified in this thesis are potential therapeutic targets. It may be possible to develop compounds which modify abnormally differentially methylated genes. Currently, there has been limited successes in AD drug trials and epigenetic mechanisms represent a promising target for future studies.

As discussed in individual chapters, there are several limitations of my work that need to be taken into consideration. First, the analyses conducted throughout this thesis were all based on ‘bulk’ tissue comprising of a complex mixture of cells types in both whole blood and cortex and therefore no cell-type-specific conclusions can be drawn. Cellular heterogeneity is important to consider in AD as pathology affects specific neural populations in the brain including loss of neurons, glial cell activation and hypertrophy of astrocytes. Areas of the cortex which are heavily impacted by neuropathology are highly susceptible to these changes. In order to account for cellular heterogeneity between individuals I included a measure of derived cell proportions in my analyses. However, there is no method to confirm that the proportions have been adequately controlled for. Moreover, it is possible some important disease associated variation is missed if one cell type negates the effects of a different cell type. Sorted cell populations should be incorporated into future analyses. For example, the isolation of purified nuclei populations using methods such as fluorescence-activated nuclei sorting (FANS) could be utilised to isolate nuclei populations which are enriched for cells involved in the aetiology of AD such as neurons, oligodendrocytes and microglia. Currently, there are limited studies which have utilised these methods in samples which cover the entire span of AD neuropathology. The BDR cohort could potentially be utilised for these analyses. EWAS of neuropathology could be conducted on the purified cell populations and comparisons between bulk and single cell profiles could be conducted. mQTLs could be identified in these specific cell populations enabling us to explore the extent to which AD risk loci are associated with these QTLs in different cell types. Following this, SMR analyses could be conducted to identify genetic variants which are pleiotropically associated with AD and molecular markers of regulation in specific cell types.

Within this thesis I focused exclusively on one epigenetic mechanism – DNAm. However, recent studies have indicated the importance of incorporating other regulatory marks. For example 5-hydroxymethyl cytosine (5hmC) - another modification which occurs to DNA - has been associated with active transcription and is associated with AD pathology in the cortex (A. R. Smith *at al.*, 2019). Of note, 5hmC is indistinguishable from DNAm when standard bisulfite approaches are used, and this could potentially confound the results identified in this thesis. Additionally, histone modifications such as lysine H3K27 acetylation (H3K27ac) - a mark of enhancer and promoter activation – has been associated with AD (Marzi *at al.*, 2018), where differentially acetylated peaks were identified near genes which have been implicated in AD pathology (e.g. *APP*, *PSEN1*, *PSEN1* and regions enriched for LOAD GWAS variants). Future studies should focus on incorporating a number of epigenetic mechanisms in order to increase our understanding of the regulatory genomic mechanisms involved in AD pathology.

Within the exception of the analyses of neuropathology, throughout this thesis I have focused on prioritising genes for AD. The methods used in **Chapters 5** and **6** could be applied to other neurodegenerative diseases including PD, DLB and FTD. Future studies should focus on identifying associations between genetic and methylomic variation utilising the SMR and PRS EWAS methods for these diseases. It would be particularly interesting to compare cross-disorder results since there is evidence of pleiotropy between these diseases. For example, SNPs annotated to HLA, MAPT and *APOE* all contribute to increased risk for FTD, AD and PD (Ferrari *at al.*, 2017). There have been several studies suggesting that PD, DLB and AD share underlying mechanisms and there are strong genetic correlations between these diseases (Desikan *at al.*, 2015; Guerreiro *at al.*, 2016). Previous EWAS have also identified methylomic similarities in neurodegenerative diseases (2016) and this hypothesis was further supported by the analyses in **Chapter 4**.

DNAm sites are generally annotated to genes based on proximity. However, the SMR analysis revealed that these are not necessarily the most functionally relevant. Future studies should focus on incorporating eQTL data in order to identify the most relevant genes. Additionally, the mQTL-eQTL information from this thesis could be utilised to create a novel gene annotation manifest which takes into consideration regulatory processes as opposed to proximity. Additionally, since only *cis* QTLs were considered, future studies should look at *trans* QTL effects and incorporate these results into the refined annotation manifest. Hi-C data could also be utilised to better annotate regulatory domains to genes. It has revealed there are contacts between distant genomic regions within the same or across different chromosomes, which has many implications for gene regulation. Hi-C assesses this 3D chromatin conformation and can be used to elucidate how the spatial organisation of DNA affects gene regulation by identifying interactions and topologically associating domains (Kikuchi *at al.*, 2019; Pombo & Dillon, 2015).

In future analyses, sequencing-based approaches could be used to give better coverage of variable DNAm across the genome. Most of the existing genomic, methylomic and transcriptomic studies of AD have been conducted using microarray technologies as they are cost-effective and high throughput. However, the content on these arrays is constrained. The Illumina EPIC array for example, contains a small proportion of the total number of genome-wide DNAm sites and has sparse coverage of certain regulatory features which are often represented by one DNAm probe (Pidsley *at al.*, 2016). Whole genome bisulfite sequencing (WGBS) provides a more comprehensive method of DNAm quantification, covering the entire genome (Stirzaker, Taberlay, Statham, & Clark, 2014). However, there are limitations to this methodology; the high costs and technical expertise required to generate WGBS has limited its application in the context of large cohort studies (Ziller, Hansen, Meissner, & Aryee, 2015). In the future, long-read sequencing technologies which are currently under active development will enable the identification of both genetic and regulatory variation in the same sample and could be utilised to gain deeper insights into the functional regulatory processes involved in the pathogenesis of AD (Amarasinghe *at al.*, 2020).

7.4 Conclusion

In conclusion, the work presented throughout my thesis represents a comprehensive assessment of genomic, methylomic and transcriptomic variation in AD across two tissues: whole blood and the cortex. I have prioritised numerous loci that could be targeted for future functional studies which could confirm their biological relevance in relation to AD. I provide further support for several pathways which are hypothesised to be involved in AD pathogenesis such as lipid and cholesterol metabolism, Aβ, tau and *APP* processing, as well as a role for the immune system. The analyses incorporating genetic and methylomic variation suggests that there are both direct *cis* genetic effects and indirect polygenic effects on regulatory processes which are involved in the aetiology of AD. A high proportion of mQTLs were conserved across tissues and there were consistent findings at a number of loci across both the whole blood and cortex PRS EWAS and SMR analyses. However, there was also evidence for heterogeneity across tissues suggesting that some tissue specific effects occur in areas primarily affected in AD (e.g. the cortex) in comparison to peripheral tissues. The data in this thesis provides a foundation for future work which could focus on validating the prioritised genes. Overall, my work further advances our understanding of the regulatory genomic process involved in the aetiology of AD.

Appendix A – Additional publications

BRAIN COMMUNICATIONS

Genetic risk for Alzheimer's disease influences neuropathology via multiple biological pathways

Eilis Hannon,[1] Gemma L. Shireby,[1] Keeley Brookes,[2] Johannes Attems,[3] Rebecca Sims,[4] Nigel J. Cairns,[1] Seth Love,[5] Alan J. Thomas,[3] Kevin Morgan,[6] Paul T. Francis[1,7] and Jonathan Mill[1]

Alzheimer's disease is a highly heritable, common neurodegenerative disease characterized neuropathologically by the accumulation of β-amyloid plaques and tau-containing neurofibrillary tangles. In addition to the well-established risk associated with the *APOE* locus, there has been considerable success in identifying additional genetic variants associated with Alzheimer's disease. Major challenges in understanding how genetic risk influences the development of Alzheimer's disease are clinical and neuropathological heterogeneity, and the high level of accompanying comorbidities. We report a multimodal analysis integrating longitudinal clinical and cognitive assessment with neuropathological data collected as part of the Brains for Dementia Research study to understand how genetic risk factors for Alzheimer's disease influence the development of neuropathology and clinical performance. Six hundred and ninety-three donors in the Brains for Dementia Research cohort with genetic data, semi-quantitative neuropathology measurements, cognitive assessments and established diagnostic criteria were included in this study. We tested the association of *APOE* genotype and Alzheimer's disease polygenic risk score—a quantitative measure of genetic burden—with survival, four common neuropathological features in Alzheimer's disease brains (neurofibrillary tangles, β-amyloid plaques, Lewy bodies and transactive response DNA-binding protein 43 proteinopathy), clinical status (clinical dementia rating) and cognitive performance (Mini-Mental State Exam, Montreal Cognitive Assessment). The *APOE* ε4 allele was significantly associated with younger age of death in the Brains for Dementia Research cohort. Our analyses of neuropathology highlighted two independent pathways from *APOE* ε4, one where β-amyloid accumulation co-occurs with the development of tauopathy, and a second characterized by direct effects on tauopathy independent of β-amyloidosis. Although we also detected association between *APOE* ε4 and dementia status and cognitive performance, these were all mediated by tauopathy, highlighting that they are a consequence of the neuropathological changes. Analyses of polygenic risk score identified associations with tauopathy and β-amyloidosis, which appeared to have both shared and unique contributions, suggesting that different genetic variants associated with Alzheimer's disease affect different features of neuropathology to different degrees. Taken together, our results provide insight into how genetic risk for Alzheimer's disease influences both the clinical and pathological features of dementia, increasing our understanding about the interplay between *APOE* genotype and other genetic risk factors.

1 College of Medicine and Health, University of Exeter, Exeter, Devon, EX2 5DW, UK
2 School of Science & Technology, Nottingham Trent University, Nottingham, NG11 8NF, UK
3 Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK
4 Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, CF24 4HQ, UK
5 Bristol Medical School (THS), University of Bristol, Bristol, BS2 8DZ, UK
6 Human Genetics Group, School of Life Sciences, University of Nottingham, Nottingham, NG7 2RD, UK
7 Wolfson Centre for Age-Related Diseases, King's College London, London, SE1 1UL, UK

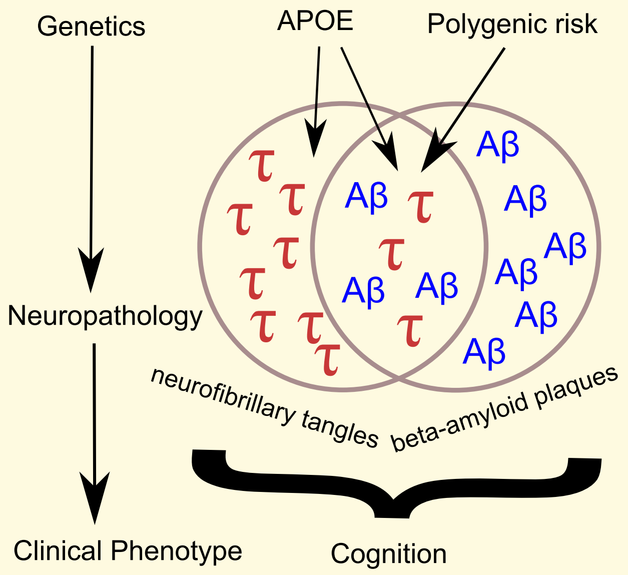
Correspondence to: Eilis Hannon PhD, RILD Level 3, RD&E Hospital, Barrack Road, Exeter EX2 5DW, UK
E-mail: e.j.hannon@exeter.ac.uk

Received August 24, 2020. Revised September 11, 2020. Accepted September 14, 2020. Advance Access publication October 12, 2020

Keywords: Alzheimer's disease; dementia; APOE; beta-amyloid; neurofibrillary tangles

Abbreviations: BDR = Brains for Dementia Research;CDR = clinical dementia rating;CERAD = Consortium to Establish a Registry for Alzheimer's disease;GWAS = genome-wide association studies;MMSE = Mini-Mental State Examination;MoCA = Montreal Cognitive Assessment;NFT = neurofibrillary tangle;PRS = polygenic risk score;SD = standard deviation;SNP = single nucleotide polymorphism;TDP-43 = transactive response DNA-binding protein 43.

Graphical Abstract



Introduction

Alzheimer's disease is a common neurodegenerative disease characterized clinically by progressive memory and cognitive decline leading to dementia and neuropathologically by β-amyloid plaques and tau-containing neurofibrillary tangles (NFTs). The most frequent manifestation of Alzheimer's disease is late onset Alzheimer's disease where onset occurs after the age of 65. Late onset Alzheimer's disease is highly heritable (Gatz *et al.*, 2006) with the most established genetic risk factor being variants of the *APOE* gene. Relative to the most common genotype (ε3/ε3), the ε4 allele increases the risk of Alzheimer's disease, with ε4 homozygosity associated with ∼20-fold increase in risk (Farrer *et al.*, 1997). In contrast, the ε2 allele of *APOE* has strong protective effects (Reiman *et al.*, 2020). Genome-wide association studies (GWAS) in large sample cohorts (Lambert *et al.*, 2013; Marioni *et al.*, 2018; Jansen *et al.*, 2019; Kunkle *et al.*, 2019) have identified additional variants in more than 40 regions of the genome which individually confer subtler effects on risk, but cumulatively account for a large proportion of genetic risk. To index an individuals' genetic risk profile, disease-associated variants—typically including those below genome-wide significance—can be combined into a 'polygenic risk score' (PRS). PRSs quantify the number of genetic risk variants an individual has, weighted by their effect size, and have been shown to improve prediction models of Alzheimer's disease (Escott-Price *et al.*, 2015, 2019; Cruchaga *et al.*, 2018). Of note, the Alzheimer's disease PRS has greatest predictive power where disease status has been defined by standardized neuropathological assessment (Escott-Price *et al.*, 2017), and is most elevated in sporadic early-onset cases (Cruchaga *et al.*, 2018).

In addition to genetic prediction, PRSs provide a powerful mechanism to investigate how genetic risk mediates the development of symptoms, and can potentially be used to disentangle the primary causal features from the secondary consequences of disease. As well as being associated with dementia status, the Alzheimer's disease PRS has been shown to correlate with mild cognitive impairment (Adams *et al.*, 2015; Chaudhury *et al.*, 2019), cognitive decline (Mormino *et al.*, 2016; Marioni *et al.*, 2017; Felsky *et al.*, 2018), memory impairments (Mormino *et al.*, 2016; Marioni *et al.*, 2017), cortical thickness (Sabuncu *et al.*, 2012; Corlier *et al.*, 2018), hippocampal volume (Lupton *et al.*, 2016; Mormino *et al.*, 2016), cerebrospinal biomarkers (Martiskainen *et al.*, 2014; Louwersheimer *et al.*, 2016; Desikan *et al.*, 2017) and neuropathology (Desikan *et al.*, 2017; Felsky *et al.*, 2018; Tasaki *et al.*, 2018). The breadth of associations highlights the complexity of understanding the pathways from genetic risk to symptomatic disease. Furthermore, many of these analyses have included the *APOE* locus within the PRS, meaning their results may reflect *APOE*-specific effects rather than the consequences of a broader polygenic risk burden. To truly understand how multiple genetic risk factors combine to influence the interplay of the clinical, cognitive and neuropathological characteristics of Alzheimer's disease, we need large, longitudinal cohorts with post-mortem tissue that can align genetics, clinical data and standardized neuropathological assessments.

A major challenge in understanding how genetic risk influences the development of Alzheimer's disease relates to clinical and neuropathological heterogeneity, and the high level of accompanying comorbidities associated with a diagnosis of Alzheimer's disease. The presence of the neuropathological hallmarks of Alzheimer's disease can only be confirmed following post-mortem brain examination. Standardized sampling and staining methods, along with the introduction of a number of semi-quantitative classification schemes, each focused on a single neuropathological feature (Thal *et al.*, 2002; Braak *et al.*, 2003, 2006), promote consistency making it easier to harmonize data across brain banks and ultimately the reproducibility of findings across studies. It is now recognized that sporadic dementia in older people is predominantly due to multiple pathologies (Robinson *et al.*, 2018). The most frequent comorbidity is Lewy body pathology affecting up to 50% of sporadic Alzheimer's disease cases (Toledo *et al.*, 2013). Another common comorbidity is the presence of inclusion bodies containing aggregates of transactive response DNA-binding protein 43 (TDP-43), particularly in the oldest old (Amador-Ortiz *et al.*, 2007; Uryu *et al.*, 2008; James *et al.*, 2016). As well as influencing cognitive impairment in non-Alzheimer's disease cases (Nag *et al.*, 2017), these comorbidities contribute to the cognitive decline observed in Alzheimer's disease cases beyond that associated with β-amyloid and NFT pathology (Wilson *et al.*, 2013; Nelson *et al.*, 2019), hence it is important to consider multiple neuropathological features simultaneously, to understand the processes that underlie cognitive performance in old age.

The paucity of comprehensive neuropathological data in large sample cohorts has limited previous genetic studies of Alzheimer's disease-associated neuropathology. To address this gap, the Brains for Dementia Research (BDR) cohort was established in 2007 recruiting both dementia patients and unaffected controls over the age of 65 to partake in routine longitudinal assessments collecting cognitive, clinical, lifestyle and psychometric data, prior to post-mortem brain donation (Francis *et al.*, 2018). The inclusion of standardized semi-quantitative data for a range of neuropathological features facilitates analyses into the specificity of genetic risk factors for the different abnormalities, and an assessment of their clinical contributions. In this study we report the first multimodal analysis of the BDR cohort, integrating longitudinal clinical and cognitive assessment with neuropathological data to explore how known genetic risk factors for Alzheimer's disease influence the development of different aspects of neuropathology and cognitive performance in old age. We focus on four common neuropathological features observed in Alzheimer's disease brain tissue: NFTs, β-amyloid plaques, Lewy bodies and TDP-43 proteinopathy. The results of this study provide insights into the neurobiological pathways to cognitive decline by refining our understanding of the complex interplay of genetic risk, clinical presentation and neuropathological burden.

Materials and methods

BDR cohort description

BDR was established in 2007 and consists of a network of six dementia research centres in England and Wales (King's College London, Bristol, Manchester, Oxford, Cardiff and Newcastle Universities) and the associated university brain banks handling the donations (Cardiff brain donations were banked in London). Participants over the age of 65 were recruited using national and local press, TV and radio coverage, articles in charity newsletters, national magazines with an older following, BDR posters, leaflets, memory clinics, talks at carer/support groups, Women's Institute and the University of the Third Age. There was no screening to exclude or include individuals with particular diagnoses or those carrying genetic variants associated with neurodegenerative diseases. The cohort includes individuals with and without dementia, spanning the full spectrum of dementia diagnoses. Participants underwent a series of longitudinal cognitive and psychometric assessments and registered for brain donation. An extensive description of the recruitment strategy, demographics, assessment protocols and neuropathic assessment procedures can be found in (Francis *et al.*, 2018).

Longitudinal cognitive and clinical assessments

All assessments were conducted by a trained psychologist or research nurse. Exclusion criteria to undergo assessments included: (i) factors precluding brain donation (e.g. brain injury/trauma, major stroke), (ii) being younger than 65 for healthy controls (except where they were spouses/partners of participants with dementia), (iii) having insufficient English language skills for completing assessments and (iv) being geographically too remote from an assessment centre. Baseline assessments were conducted face-to-face (in the participant's place of residence or a BDR centre), follow-up assessments were usually face-to-face but telephone interviews were also used for some healthy control participants. Follow-up interviews were annual for participants with cognitive impairment, and every 1–5 years (depending on age) for cognitively healthy participants. Clinical assessment was performed using the clinical dementia rating (CDR) (Morris, 1993). Cognitive assessment measures relevant to this study included the Mini-Mental State Examination (MMSE) (Folstein *et al.*, 1975) and Montreal Cognitive Assessment (MoCA) (Nasreddine *et al.*, 2005).

Post-mortem neuropathological assessment

After removal, the brain was examined macroscopically and digitally recorded. After slicing, the brain was comprehensively sampled according to the BDR protocol by experienced neuropathologists in each of the five network brain banks. This protocol, arrived at by consensus across the BDR network and based on the BrainNet Europe initiative (Bell *et al.*, 2008), was used to generate a description of the regional pathology within the brain together with standardized scoring. In this study we considered five variables representing four neuropathological features: (i) Braak tangle stage which captures the progression of NFT pathology (Braak and Braak, 1991; Braak *et al.*, 2006), (ii) Thal β-amyloid phase which captures the regional distribution of plaques (Thal *et al.*, 2002), (iii) Consortium to Establish a Registry for Alzheimer's disease (CERAD) stage which profiles neuritic plaque density (Mirra *et al.*, 1991; Montine *et al.*, 2012), (iv) Braak Lewy body stage (Braak *et al.*, 2003) and (v) TDP-43 status (a binary indicator of the absence/presence of TDP-43 inclusions, as assessed by immunohistochemistry of the amygdala and the hippocampus and adjacent temporal cortex for phosphorylated TDP-43). All variables apart from TDP-43 were analysed as continuous variables, using their semi-quantitative nature to capture dose-dependent relationships of increasing neuropathological burden.

Genetic data

DNA extraction was performed using a standard phenol chloroform method on 100 mg of brain tissue. DNA quality was assessed using the Agilent 2200 TapeStation DNA integrity number and quantified using NanoDrop 3300 spectrometry. Genotyping was performed on the NeuroChip array which is a custom Illumina genotyping array with an extensive genome-wide backbone ($n = 306\ 670$ variants) and custom content covering 179 467 variants specific to neurological diseases (Blauwendraat *et al.*, 2017). Genotype calling was performed using GenomeStudio (v2.0, Illumina) and quality control was completed using PLINK1.9 (Chang *et al.*, 2015). Individuals were excluded if either (i) they had > 5% missing data, (ii) their genotype predicted sex using X chromosome homozygosity was discordant with their reported sex (excluding females with an F value > 0.2 and males with an F value < 0.8), (iii) they had excess heterozygosity [>3 standard deviation (SD) from the mean], (iv) they were related to another individual in the sample (pi hat > 0.2), where one individual from each pair of related samples was excluded considering data quality and phenotype or (v) they were classed as non-European, determined by merging the BDR genotypes with data from HapMap Phase 3 (http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html), linkage disequilibrium pruning the overlapping single nucleotide polymorphisms (SNPs) such that no pair of SNPs within 1500 bp had $r^2 > 0.20$ and visually inspecting the first two genetic principal components along with the known ethnicities of the HapMap sample to define European samples (Supplementary Fig. 1). Prior to imputation SNPs with high levels of missing data (>5%), Hardy-Weinberg equilibrium $P < 0.001$ or minor allele frequency <1% were excluded. The genetic data were then recoded as vcf files before uploading to the Michigan Imputation Server (Das *et al.*, 2016) (https://imputationserver.sph.umich.edu/index.html#!) which uses Eagle2 (Loh *et al.*, 2016) to phase haplotypes, and Minimac4 (https://genome.sph.umich.edu/wiki/Minimac4) with the most recent 1000 Genomes reference panel (phase 3, version 5). Imputed genotypes were then filtered with PLINK2.0alpha, excluding SNPs with an R^2 INFO score < 0.5 and recoded as binary PLINK format. Proceeding with PLINK1.9, samples with >5% missing values, and SNPs with >2 alleles, >5% missing values, Hardy-Weinberg equilibrium $P < 0.001$ or a minor allele frequency of <5% were excluded. The final quality controlled imputed set of genotypes contained 6 607 832 variants.

Polygenic risk scores

GWAS results from Kunkle *et al.* (2019) were used to calculate an Alzheimer's disease PRS for each individual. We choose this GWAS as it is based on clinically defined cases compared to controls. To separate the effects of *APOE* from other genetic variants associated with Alzheimer's disease, we excluded the *APOE* region (chr19:45 116 911–46 318 605) (Kunkle *et al.*, 2019) from the PRS calculations. We generated PRS using PRSice (v2.0) (Choi and O'Reilly, 2019) which 'clumps' the Alzheimer's disease GWAS summary statistics using the BDR genotype data such that the most significant variant in each linkage disequilibrium block was retained. The PRS was then calculated in the target (BDR) dataset for each individual, as the number of reference alleles multiplied by the log odds ratio for that SNP (taken from the Kunkle *et al.* Alzheimer's disease GWAS), and then summed across all retained clumped variants with an Alzheimer's disease GWAS P-value < $P_{T.}$ A range of P-value thresholds (P_T) were used initially, to generate multiple possible PRS, where the optimal PRS was selected as the score that explained the highest proportion of variance (Nagelkerke's pseudo R^2) in Alzheimer's disease case control status. In this analysis, Alzheimer's disease cases and controls were defined as Braak high (Braak tangle stages V and VI) and low (Braak tangle stages 0–II) respectively, and PRS was tested using a logistic regression model with the first 8 genetic principal components as covariates. In the BDR cohort the optimal threshold for selecting SNPs for the PRS was $P < 5 \times 10^{-8}$ (Supplementary Fig. 2). Prior to analysis the PRS calculated at this threshold was standardized to have a mean of 0 and SD of 1; therefore the interpretation is in units of SDs.

APOE genotyping

The *APOE* SNPs rs7412 and rs429358 were genotyped with TaqMan assays using standard protocols. Where *APOE* genotype by TaqMan assay was not available, it was generated from the NeuroChip data ($n = 44$). The NeuroChip array includes multiple probes to assay the two *APOE* SNPs; based on the optimal concordance with the TaqMan assay (91% concordant across assays) we used the probes rs7412.B3 and rs429358.T2 to determine *APOE* status. In all statistical analyses, *APOE* status was modelled as two numeric variables counting the number ε2 alleles and number of ε4 alleles an individual had. Given the rarity of ε2/ε2 genotype [only four occurrences (0.58%) in this sample], the ε2/ε2 individuals were combined with the individuals with one ε2 allele.

Statistical analysis

All statistical analyses were performed in R version 3.5.2. All analytical code is available via GitHub (https://github.com/ejh243/BDR-Genetic-Analyses).

Survival analysis

To test whether *APOE* and Alzheimer's disease PRS were associated with younger age at death, we fitted Cox's proportional hazards models using the R package survival. Three models were fitted with age at death as the outcome to test (i) *APOE* genotype modelled as two variables, (ii) Alzheimer's disease PRS and (iii) *APOE* genotype and Alzheimer's disease PRS simultaneously. All models included covariates for sex, BDR centre and eight genetic principal components.

Genetic analysis of neuropathology and clinical/cognitive status at death

Genetic associations between either *APOE* status or Alzheimer's disease PRS and any of the continuous neuropathology variables (Braak tangle stage, Thal β-amyloid stage, CERAD stage, Braak Lewy body stage), clinical (CDR global rating) or cognitive status at death (MMSE, MoCA) were tested using a linear regression model. TDP-43 proteinopathy as a binary variable was analysed with logistic regression, but the model framework was the same. Up to four regression models were fitted for each variable. First, the effects of *APOE* status and Alzheimer's disease PRS were estimated separately using Model 1 and Model 2 below.

Model 1:

$$\text{variable} \sim APOE_{\varepsilon2} + APOE_{\varepsilon4} + \text{covariates} + \text{genetic PCs}_{1-8}.$$

Model 2:

$$\text{variable} \sim \text{PRS} + \text{covariates} + \text{genetic PCs}_{1-8}.$$

If *APOE* (either variable) and PRS were significantly associated with an outcome, then a multiple regression analysis was additionally fitted testing *APOE* and PRS simultaneously to confirm these were independent associations (Model 3).

Model 3:

$$variable \sim APOE_{\varepsilon2} + APOE_{\varepsilon4} + \text{PRS} + \text{covariates} + \text{genetic PCs}_{1-8}.$$

Finally, an interaction model (Model 4) between *APOE* and PRS was fitted to test if PRS associations differed depending on *APOE* genotype.

Model 4:

$$\text{variable} \sim APOE_{\varepsilon2} + APOE_{\varepsilon4} + \text{PRS} + \text{APOE}_{\varepsilon2}{*}\text{PRS} + APOE_{\varepsilon4}{*}\text{PRS} + \text{covariates} + \text{genetic PCs}_{1-8}.$$

All analyses included age at death, sex and BDR centre as covariates and the first eight genetic principal components. Analyses for clinical or cognition measures also included a covariate that measured the time lapse between the last assessment and death.

Longitudinal clinical and cognition analyses

To test how *APOE* and Alzheimer's disease PRS affected clinical status and cognitive trajectories, we fitted multilevel regression models using all available pre-mortem assessment data. A time variable was created which measured the number of days after the first visit that an assessment took place. Each cognitive variable was then tested as the dependent variable against this time variable included as a fixed effect along with covariates for age, sex, BDR centre and the first eight genetic principal components and a random effect for individual. To test for genetic effects on the cognitive trajectory, either *APOE* (coded as two variables) or Alzheimer's disease PRS, was included in the model as a main effect and as an interaction with time. Models were fitted using the R packages lme4 and lmTest.

Multiple testing

In total, we tested 12 outcomes against 3 genetic variables. Our outcomes comprised five neuropathological variables, one clinical variable at death, two cognitive measures, one longitudinal clinical, two longitudinal cognitive measures and a survival analysis of age at death. Against these 12 outcomes, we tested 3 genetic variables (Alzheimer's disease PRS and two variables to model *APOE* genotype). Therefore, we performed a multiple testing correction for 36 tests, reporting significant associations as those with $P < 0.0014$. Given the correlations between the neuropathological, clinical and cognitive variables this is likely to be a conservative approach.

Data availability

Genetic, clinical and cognitive data are available through the Dementia's Platform UK (DPUK; https://www.dementiasplatform.uk/) platform upon application.

Results

Both tauopathy and β-amyloidosis are present at high frequencies in the BDR cohort

To profile the effects of both *APOE* genotype and Alzheimer's disease PRS, our analyses were limited to BDR donors who had undergone neuropathological assessment and had NeuroChip array data (n = 693, Table 1). The participants had a mean age at death of 83.5 years (SD = 9.34 years) and 52.8% were male. Consistent with epidemiological reports, females were significantly older at death than males (mean difference = 3.84 years; $P = 4.87 \times 10^{-8}$). Within this cohort, 57.3% of individuals had dementia at their first assessment (i.e. at baseline), with 63.3% of the cohort affected by dementia at death. At recruitment, individuals had a mean CDR of 1.42 (SD = 1.36), a mean MMSE score of 22.3 (SD = 8.81) and a mean MoCA score of 17.2 (SD = 10.6). These scores indicate that the majority of participants only suffered mild cognitive impairment, although the full range of cognitive performance was represented in the cohort. Participants underwent a mean of 2.85 assessments (SD = 1.71) prior to death. Individuals who had at least two assessments (N = 486) were followed for a mean of 3.40 years (SD = 2.00 years) with a mean of 1.42 years between assessments (SD = 0.67 years). Our genetic analyses focused on four semi-quantitative and one indicator neuropathology variable. In 672 samples NFT pathology was quantified using Braak NFT stage (Braak and Braak, 1991; Braak *et al.*, 2006) with a mean of 3.76 (SD = 1.90). Two variables reflecting the extent of β-amyloidosis were considered: β-amyloid distribution was measured by Thal β-amyloid phase (Thal *et al.*, 2002) with a mean value of 3.14 (SD = 1.78) across 612 individuals and neuritic plaque density was scored using the CERAD classification (Mirra *et al.*, 1991; Montine *et al.*, 2012) with a mean value of 1.72 (SD = 1.26) across 634 individuals. α-Synuclein pathology was quantified using Braak Lewy body stage, where across 634 individuals the mean was 1.36 (SD = 2.26). TDP-43 status was available for 658 individuals, with 150 (22.8%) individuals classed as being TDP-43 positive.

Genetic risk factors for Alzheimer's disease are associated with increased mortality

To determine whether higher genetic risk for Alzheimer's disease was associated with increased mortality we

Table 1 Summary of BDR cohort

		%	Mean	SD	*N*
Demographics	Sex (male)	52.8			693
	Age		83.5	9.34	693
Clinical assessments	Number of assessments		2.85	1.71	693
	Time in study (years)		3.40	2.00	486
	Time between assessments (years)		1.42	0.67	486
Dementia status at first assessment	Dementia	57.3			693
	MCI/inconclusive	13.3			693
	No dementia	29.3			693
Dementia status at last assessment	Dementia	63.3			693
	MCI/inconclusive	14.2			693
	No dementia	22.5			693
Neuropathology	Braak stage tangle		3.76	1.9	672
	Thal amyloid stage		3.14	1.78	612
	CERAD stage		1.72	1.26	634
	Braak Lewy body stage		1.36	2.26	597
	TDP-43	22.8			658
Cognitive scores at first assessment	CDR		1.42	1.36	639
	MMSE		22.3	8.81	469
	MoCA		17.2	10.6	270
Cognitive scores at last assessment	CDR		1.79	1.3	639
	MMSE		19.1	10.3	469
	MoCA		16.1	11	270

Table 2 ***APOE*** **is associated with increased mortality**

Analytical model	*APOE*						PRS		
	Number of ε2 alleles			Number of ε4 alleles					
	Hazard ratio	SE	*P*-value	Hazard ratio	SE	*P*-value	Hazard ratio	SE	*P*-value
Model 1	0.835	0.123	0.142	1.293	0.066	9.66E−05			
Model 2							1.105	0.038	8.97E−03
Model 3	0.839	0.124	0.155	1.292	0.066	1.00E−04	1.106	0.038	8.41E−03

Table 3 **Common genetic risk factors for Alzheimer's disease are associated with multiple aspects of neuropathology**

Analytical model	Neuropathological variable	*APOE*						PRS		
		Number of ε 2 alleles			Number of ε 4 alleles					
		P-value	Coefficient	%VarExp	*P*-value	Coefficient	%VarExp	*P*-value	Coefficient	%VarExp
Model 1	Braak stage tangle	0.0877	−0.357	0.958	4.16E−24	1.16	15.1			
	Thal amyloid stage	0.00333	−0.562	1.54	3.96E−20	0.981	13.5			
	CERAD stage	0.0224	−0.329	1.99	1.03E−19	0.713	13.4			
	Braak Lewy body stage	0.988	−0.00439	0.0809	0.000264	0.555	2.59			
	TDP-43	0.859	−0.0574	0.00821	0.00158	0.537	2.58			
Model 2	Braak stage tangle							1.36E−06	3.4	0.354
	Thal amyloid stage							0.00288	1.1	0.201
	CERAD stage							5.27E−05	2.95	0.202
	Braak Lewy body stage							0.267	0.167	0.105
	TDP-43							0.315	0.26	0.104
Model 3	Braak stage tangle	0.0885	−0.3505	0.9580	9.40E−24	1.132	15.119	4.97E−06	0.309	2.465
	CERAD stage	0.0224	−0.3254	1.9865	2.02E−19	0.700	13.402	1.30E−04	0.179	2.192

analysed survival with Cox's proportional hazard models (Table 2). *APOE* genotype was modelled as two variables—the number of ε4 alleles and the number of ε2 alleles, to distinguish the hypothesized risk effects of ε4 (Corder *et al.*, 1993; Farrer *et al.*, 1997) from the protective effects of ε2 (Reiman *et al.*, 2020). Analysis of *APOE* genetic risk found that *APOE* ε4 status was significantly associated with younger age at death, with each additional ε4 allele associated with 29% increased risk of death (hazard ratio = 1.29; $P = 9.66 \times 10^{-5}$). Alzheimer's disease PRS was nominally associated with an increased mortality (hazard ratio = 1.11; $P = 8.97 \times 10^{-3}$), although this was not significant after correcting for multiple testing.

APOE and Alzheimer's disease PRS independently influence tauopathy and β-amyloidosis

The number of *APOE* ε4 alleles was positively associated ($P < 0.00014$) with all four semi-quantitative neuropathology measures (Table 3). The most significant association was with Braak NFT stage: each ε4 allele was associated with an increase in 1.16 Braak NFT stages ($P = 4.16 \times 10^{-24}$). Associations were also found between ε4 status and Thal *β*-amyloid phase (mean difference per ε4 allele = 0.981 phases; $P = 3.96 \times 10^{-20}$), neuritic plaque density (mean difference per ε4 allele = 0.713 stages; $P = 1.03 \times 10^{-19}$) and Braak Lewy body stage (mean difference per ε4 allele = 0.555 stages; $P = 2.64 \times 10^{-4}$). Alzheimer's disease PRS was associated with two measures of neuropathology (Table 3): a higher polygenic burden was associated with Braak NFT stage (mean difference per SD of PRS = 0.354 stages; $P = 1.36 \times 10^{-6}$) and neuritic plaque density (mean difference per SD of PRS = 0.202 stages; $P = 5.27 \times 10^{-5}$). TDP-43 was not associated with either *APOE* genotype or Alzheimer's disease PRS. Although variants in the *APOE* region were excluded from the PRS, we tested both *APOE* and PRS against Braak NFT stage and neuritic plaque density simultaneously to confirm that the identified associations were independent. The estimated effects of ε4 on both Braak NFT stage and neuritic plaque density were unaffected, while the Alzheimer's disease PRS associations were slightly attenuated (Table 3) but remained significant. In addition to an additive model, we tested whether there was evidence for a multiplicative effect between Alzheimer's disease PRS and *APOE* genotype on neuropathological burden to explore the hypothesis that in individuals with protective *APOE* genotypes, Alzheimer's disease PRS is more important (i.e. has a larger effect on neuropathology). In this analysis, none of the five neuropathological variables had statistically significant differences across *APOE* genotype groups ($P > 0.05$) (Supplementary Table 1). Taken together, these results suggest that *APOE* status and Alzheimer's disease PRS are independently associated with neuropathology, combining in an additive manner to influence an individual's accumulation of tauopathy (NFTs) and β-amyloid plaques.

Given that the two distinct molecular pathologies—tauopathy and β-amyloidosis—that define Alzheimer's disease are highly correlated (Supplementary Fig. 3), we wanted to establish whether *APOE* or Alzheimer's disease PRS had a specific (or primary) effect on a particular aspect of neuropathology. To this end, we repeated the analysis of how Alzheimer's disease PRS and *APOE* influence pathology, sequentially controlling for other neuropathology variables. This analysis revealed some interesting patterns. First, after controlling for any of the other three quantitative neuropathological variables, Braak Lewy body stage was not significantly associated with *APOE* ε4 (Supplementary Table 2) suggesting that the association we detected was largely driven by the fact that individuals with Lewy bodies have also NFTs and β-amyloid plaques. Second, after we controlled for Braak NFT stage, neither of the plaque measures remained significantly associated with *APOE* ε4 (Supplementary Table 2). In contrast, Braak NFT stage remained significantly associated with *APOE* ε4 status after controlling for plaque variable (adjusted for Thal phase, mean difference per *APOE* ε4 allele = 0.468; $P = 6.44 \times 10^{-7}$; adjusted for neuritic plaque density, mean difference per ε4 allele = 0.238; $P = 1.82 \times 10^{-4}$), albeit with an attenuated magnitude of effect. Considering the two measures of plaque burden, only Thal β-amyloid phase remained significantly associated with ε4 after controlling for neuritic plaque density (mean difference per ε4 allele = 0.265; $P = 3.42 \times 10^{-4}$). Neither Braak NFT stage nor neuritic plaque density remained significantly associated with Alzheimer's disease PRS after controlling for the other measure of pathology (Supplementary Table 2). These results indicate that *APOE* ε4 has a specific influence on tauopathy (NFTs) as well as a shared effect on both plaque and NFT development, whereas the PRS is more generally associated with an increased burden of Alzheimer's disease neuropathology.

Association between *APOE* and cognitive performance is confounded by neuropathology

We determined clinical and cognitive status at death from the final pre-mortem assessment (Table 1). Data were available from 639 individuals who had had at least one CDR assessment with a mean final score of 1.79 (SD = 1.30) measured a mean of 353 days (SD = 374 days) prior to death. In addition, 469 individuals had had at least one MMSE assessment with a mean final score of 19.1 (SD = 10.3) measured a mean of 594 days (SD = 521 days) prior to death and 270 individuals had a MoCA assessment (mean = 16.1; SD = 11.0) measured a mean of 617 days (SD = 590 days) prior to death. *APOE* was significantly associated with dementia severity with each ε4 allele associated with an increase in 0.492 ($P = 2.14 \times 10^{-9}$) in pre-mortem CDR score (Table 4). *APOE* was also significantly associated with lower cognitive performance in MMSE prior to death (Table 4) with each ε4 allele being associated with a decrease in 4.86 ($P = 1.30 \times 10^{-8}$). In contrast, Alzheimer's disease PRS was not significantly associated with any of the measures of clinical or cognitive status prior to death. To test whether the association between *APOE* and clinical measures was mediated by neuropathology we repeated these analyses including Braak NFT stage as an additional covariate; this variable had the largest effect in the genetic analyses described above, and its effect additionally captured associations with plaque pathology. In this model, the associations between *APOE* ε4 and CDR or MMSE were attenuated and neither remained significant (Supplementary Table 3). In contrast, on retesting Braak NFT stage whilst controlling for the clinical variables in turn, we observed that *APOE* ε4 remained significantly associated (Supplementary Table 3). This indicates that the association between *APOE* and clinical variables is a consequence of an increased burden of neuropathology.

APOE ε4 is associated with faster cognitive decline in old age, but this is driven by Alzheimer's disease neuropathological burden

Participants had a mean of 2.85 (SD = 1.71 visits) clinical assessment visits spread over a mean of 3.40 years (SD = 2.00 years) with a mean time between visits of 1.42 years (SD = 0.67 years). Over the course of all participants' involvement in the BDR study, there was an overall decline in clinical status and cognitive performance. On average the CDR increased by a mean of 0.139 per year ($P = 2.02 \times 10^{-31}$), while MMSE declined by a mean of 1.07 per year ($P = 3.00 \times 10^{-29}$). *APOE* genotype was associated with worse cognitive scores at the start of the study and faster rates of decline as the study progressed (Table 5). For every ε4 allele, MMSE score was 3.19 points lower ($P = 4.92 \times 10^{-5}$) at the start of the study, and individuals then accumulated an additional decrease in 0.803 in their score per allele per year ($P = 1.58 \times 10^{-8}$). In contrast, although *APOE* was associated with a higher CDR score at the start of the study (mean difference per ε4 allele = 0.468; $P = 4.34 \times 10^{-8}$), there was no significant difference in the change in clinical status related to *APOE* as the study progressed. There was no significant association with MoCA scores and *APOE* genotype. There was no significant association between Alzheimer's disease PRS and longitudinal clinical or cognitive profiles or clinical or cognitive status at study entry. On repeating these analyses using the

Table 4 ***APOE* is associated with clinical and cognitive status at death**

Analytical model	Cognitive variable	*APOE*						PRS		
		Number of ε 2 alleles			Number of ε 4 alleles					
		P-value	Coefficient	%VarExp	*P*-value	Coefficient	%VarExp	*P*-value	Coefficient	%VarExp
Model 1	CDR	0.706	−0.058	0.336	2.14E−09	0.492	9.83			
	MMSE	0.693	0.574	0.310	1.30E−08	−4.859	10.05			
	MoCA	0.876	−0.299	0.157	5.00E−03	−3.403	3.19			
Model 2	CDR							0.034	0.109	1.82
	MMSE							0.025	−1.136	2.03
	MoCA							0.785	0.191	0.089

Table 5 ***APOE* ε4 is associated with steeper cognitive decline prior to death**

Time variable	Cognitive variable	Time		*APOE*				Interaction (time × *APOE*)			
				Number of ε2 alleles		Number of ε4 alleles		Time × ε2		Time × ε4	
		Coefficient	*P*-value	Coefficient	*P*-value	Coefficient	*P*-value	Coefficient	*P*-value	Coefficient	*P*-value
Time since study entry (days)	CDR	2.78E−04	2.45E−08	0.075	0.642	0.468	4.34E−08	−1.34E−04	0.121	1.28E−04	9.83E−03
	MMSE	−1.39E−03	7.50E−05	1.209	0.371	−3.195	4.92E−05	−3.73E−04	0.529	−2.20E−03	1.58E−08
	MoCA	−1.33E−03	0.146	−0.663	0.710	−2.335	0.040	1.98E−03	0.151	−3.18E−03	0.024
Age (years)	CDR	2.01E−03	0.797	−0.058	0.965	−0.868	0.216	4.99E−04	0.974	0.017	0.042
	MMSE	−0.258	2.20E−04	4.759	0.675	2.574	0.689	−0.040	0.766	−0.086	0.268
	MoCA	0.011	0.904	−4.510	0.783	−5.281	0.616	5.21E−02	0.785	0.031	0.809

participant's age rather than time in the study, we found no significant linear associations with either cognitive status at study entry or performance as the study progressed (Table 5).

Given our previous observation that genetic associations with clinical status and cognition are mediated by neuropathology, we wanted to confirm whether the longitudinal analyses were similarly affected. First, we tested whether change in clinical status was associated with neuropathology measured by Braak NFT stage, independent of genetic status (Supplementary Table 5). As expected, those with higher levels of tangle pathology at death had a more severe clinical rating, even at the start of the study (mean difference in CDR per Braak NFT stage = 0.355; $P = 7.30 \times 10^{-42}$) and declined quicker; each additional Braak NFT stage was associated with an additional increase in 0.0247 in CDR per year ($P = 3.99 \times 10^{-5}$). We observed similar results for cognitive performance measured by MMSE; at study entry, each additional Braak NFT stage was associated with a decrease in 2.58 in MMSE score ($P = 7.27 \times 10^{-26}$) and participants accumulated an additional decrease in 0.384 in MMSE per Braak NFT stage per year ($P = 3.90 \times 10^{-15}$). Repeating the *APOE* analysis with a covariate for the potential confounder of neuropathology found that in line with the cross-sectional analyses, the associations with both clinical severity and cognition were no longer significant after adjusting for Braak NFT stage (Supplementary Table 6). These results suggest that cognitive performance prior to death, and even many years before death, is a consequence of accumulating Alzheimer's disease neuropathology.

Discussion

In this study, we used the longitudinal cognitive and neuropathological assessment data in the BDR cohort to investigate how genetic risk factors for Alzheimer's disease influence the accumulation of β-amyloid plaques, tauopathy, synucleinopathy and TDP-43 proteinopathy, and progressive decline in clinical status and cognitive performance. Our results indicate that *APOE* ε4 status has the most dramatic influence on tauopathy (NFT burden) and that although *APOE* genotype is also associated with β-amyloidosis, synucleinopathy and cognition, these relationships are largely confounded by their correlation with tangle burden. Furthermore, our results indicate that *APOE* has a specific direct effect on NFT independent of other neuropathologies. Although this finding contradicts the predictions of the 'amyloid cascade hypothesis' in which tau tangle formation is considered secondary to β-amyloid pathology (Hardy and Allsop, 1991; Selkoe, 1991), it is consistent with careful neuropathologic studies that show that tauopathy can precede beta-amyloidosis, at least in some brain areas (Duyckaerts, 2011). Our results also agree with previous research showing that although the influence of *APOE* on tau tangles is largely mediated indirectly through neurobiological pathways associated with β-amyloid, approximately one-third of its influence on tangle development is via an alternative non-amyloid pathway (Yu *et al.*, 2014). Our findings also support the 2-process model proposed by Mungas *et al.* (2014), according to which neocortical NFTs are mediated by *β*-amyloid deposition and medial temporal lobe NFTs and may be the consequence of a separate age-associated process.

In our analysis of pathologies that frequently co-occur with the accumulation of *β*-amyloid and tauopathy, we replicated the positive association between Lewy body burden and the *APOE* ε4 allele (Tsuang *et al.*, 2013; Beecham *et al.*, 2014). However, when we adjusted for either *β*-amyloid or NFTs, this association was attenuated, indicating that in our sample, the association may be a consequence of the higher levels of tau and *β*-amyloid in individuals with Lewy bodies. It should be noted that the majority of participants in our study were free of any Lewy body pathology, with 423 individuals (70.8%) having a Braak Lewy body stage of 0. Therefore these analyses may be underpowered, particularly in the context of disentangling the effects on multiple correlated neuropathology variables. In addition, we were not able to replicate associations between *APOE* genotype and the presence of TDP-43 proteinopathy (Josephs *et al.*, 2014; Yang *et al.*, 2018), although the direction of effect was consistent with previous reports. Although TDP-43 proteinopathy was not infrequent in the BDR cohort, with 22.8% participants classed as positive, our simple binary classification may have decreased our power to detect an effect. Although BDR is not limited to a particular dementia subtype, and includes unaffected controls, Alzheimer's disease is the most common form of dementia and therefore the sample is enriched for NFT and *β*-amyloid pathology. To truly establish whether *APOE* genotype has an independent, direct effect on the common comorbidities associated with Alzheimer's disease, such as Lewy bodies and TDP-43 proteinopathy, we will likely require a larger number of samples to detect residual effects after accounting for correlations between neuropathological variables.

As well as our examination of associations with *APOE*, we tested the cumulative effect of common Alzheimer's disease-associated genetic variants on neuropathology, clinical status and cognition. Given that individual variants only confer a small amount of additional risk, we used a combined PRS to improve power. Although Alzheimer's disease PRS was associated with both tauopathy (NFTs) and *β*-amyloidosis, there was no evidence of independent effects on either, suggesting that, in combination, common genetic variants have a broader, more general effect on the neuropathological burden present in Alzheimer's disease. This contrasts with findings from a previous study testing the consequences of an Alzheimer's disease PRS without *APOE*, which only reported a significant association with NFTs and not *β*-amyloid plaques (Felsky *et al.*, 2018). Of note, in that study the PRS was based on an older GWAS with fewer significant association signals, and therefore our study might highlight the additional power derived using variants from the latest GWAS for Alzheimer's disease. While leveraging multiple genetic variants into a single PRS is a powerful approach, particularly where sample sizes are small, it can be challenging to interpret shared associations. As the PRS is a harmonized variable generated in our case from seventeen genetic variants, our results could be explained by different subsets of variants being causally associated with the distinct pathologies. This explanation fits with results from previous studies that have tested individual SNPs associated with Alzheimer's disease against multiple measures of neuropathology reporting some variants having specific effects, while others were associated with multiple aspects (Beecham *et al.*, 2014; Mäkelä *et al.*, 2018). Furthermore, it is likely that some genetic risk factors do not act via either plaques or tauopathy (NFTs), possibly affecting other aspects of neuropathology such as vascular disease which was not included in this study.

We found that clinical and cognitive status at study recruitment and prior to death, in addition to decline over the course of the study, are not directly associated with *APOE* genotype but are likely to be a consequence of neuropathological burden and in particular the accumulation of NFTs. This concurs with results from a previous study in a slightly larger cohort that focused specifically on episodic memory and non-episodic cognition (Yu *et al.*, 2014). Alzheimer's disease-associated cognitive decline is hypothesized to start as much as 17 years prior to death, with the rate of decline fastest in those with the most extensive neuropathology; tauopathy, *β*-amyloidosis, TDP-43 proteinopathy and synucleinopathy are all positively associated with decline (Boyle *et al.*, 2017). While a strength of our study is the availability of longitudinal cognitive data, clinical data was only available for up to three years before death, limiting our ability to characterize the effects of neuropathology on cognitive trajectories. Furthermore, multiple aspects of neuropathology have been independently negatively associated with cognitive performance (Boyle *et al.*, 2013). Although Alzheimer's disease is characterized by *β*-amyloidosis and tauopathy, it is increasingly apparent that in older cohorts, there may be additional comorbidities which potentially confound this relationship (Schneider *et al.*, 2009; James *et al.*, 2012, 2016; Robinson *et al.*, 2018). At present, the presence of multiple comorbidities makes it difficult to resolve cause from effect as each comorbidity may affect different domains of cognition at different times during pathogenesis. When considering the regional presence and global burden of different pathologies, there is extensive variation in the specific combination of neuropathological features that an individual develops ultimately having a unique effect on their individual cognitive performance over time (Boyle *et al.*, 2018). The strengths of the BDR study design, collating repeated measures of cognitive performance in addition to standardized protocols for high quality neuropathological assessments in a large sample size make it an ideal dataset to ultimately disentangle the role of mixed pathologies on cognition and dementia and more extensive analyses will be possible in the future.

Our results should be considered in light of a number of limitations. First, the participants were self-selecting, which in line with many other observational cohorts introduces bias into the sample; they are from less deprived socio-economic areas and have higher levels of education than the general population. Second, consistent with the majority of genetic studies, our analysis was limited to participants of European ancestry to remove the biases associated with population stratification. Third, we only included a subset of Alzheimer's disease and related neuropathology phenotypes, which were selected for practical reasons in that they were observed with sufficient frequency in the current sample. Analyses of rarer phenotypes will be possible with subsequent waves of the data as the overall sample size and number of cases increases. Fourth, our measures were of global cognition, rather than specific domains. As previous studies have found that different pathologies have specific effects of different cognitive domains (Yu *et al.*, 2014), this may mean we miss some of the nuances of the relationship between neuropathology and cognition. Fifth, to aid interpretation of the analytical models we converted semi-quantitative neuropathological variables into continuous variables which assume an equal effect between all pairs of consecutive stages. This simplification may obscure some more complex patterns in the data but should enable us to pick up general correlations which were our primary interest. Finally, we did not control for severity of ischaemic brain damage or any vascular risk factors, which are common in Alzheimer's disease cases and negatively influence cognition.

In summary, our data indicate that *APOE* influences Alzheimer's disease neuropathology via two independent pathways, one where β-amyloid accumulation correlates with the development of tauopathy (NFTs), and a second pathway with direct effects on NFTs independent of β-amyloidosis. It is as a consequence of these neuropathological changes that cognitive performance is then impaired. The relationship between common genetic variants associated with Alzheimer's disease and neuropathology is more complex, with each individual variant potentially having a different effect on neuropathology and cognition. Taken together, these results provide insights into how the symptoms of Alzheimer's disease dementia manifest and how genetic risk factors influence the development of pathology.

Acknowledgements

We would like to gratefully acknowledge all donors and their families for the tissue provided for this study. Human post-mortem tissue was obtained from the South West Dementia Brain Bank, London Neurodegenerative Diseases Brain Bank, Manchester Brain Bank, Newcastle Brain Tissue Resource and Oxford Brain Bank, members of the Brains for Dementia Research (BDR) Network. We wish to acknowledge the neuropathologists at each centre and BDR Brain Bank staff for the collection and classification of the samples.

Funding

E.H. and J.M. were supported by Medical Research Council grant K013807. G.S. was supported by a PhD studentship from the Alzheimer's Society. The BDR is jointly funded by Alzheimer's Research UK and the Alzheimer's Society in association with the Medical Research Council. The South West Dementia Brain Bank is part of the Brains for Dementia Research program, jointly funded by Alzheimer's Research UK and Alzheimer's Society, and is also supported by BRACE (Bristol Research into Alzheimer's and Care of the Elderly) and the Medical Research Council.

Supplementary material

Supplementary material is available at *Brain Communications* online.

Competing interests

The authors report no competing interests.

References

Adams HH, de Bruijn RF, Hofman A, Uitterlinden AG, van Duijn CM, Vernooij MW, et al. Genetic risk of neurodegenerative diseases is associated with mild cognitive impairment and conversion to dementia. Alzheimers Dement 2015; 11: 1277–85.

Amador-Ortiz C, Lin WL, Ahmed Z, Personett D, Davies P, Duara R, et al. TDP-43 immunoreactivity in hippocampal sclerosis and Alzheimer's disease. Ann Neurol 2007; 61: 435–45.

Beecham GW, Hamilton K, Naj AC, Martin ER, Huentelman M, Myers AJ, et al. Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. PLoS Genet 2014; 10: e1004606.

Bell JE, Alafuzoff I, Al-Sarraj S, Arzberger T, Bogdanovic N, Budka H, et al. Management of a twenty-first century brain bank: experience in the BrainNet Europe consortium. Acta Neuropathol 2008; 115: 497–507.

Blauwendraat C, Faghri F, Pihlstrom L, Geiger JT, Elbaz A, Lesage S, et al. NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases. Neurobiol Aging 2017; 57: 247.e9–13.

Boyle PA, Wilson RS, Yu L, Barr AM, Honer WG, Schneider JA, et al. Much of late life cognitive decline is not due to common neurodegenerative pathologies. Ann Neurol 2013; 74: 478–89.

Boyle PA, Yang J, Yu L, Leurgans SE, Capuano AW, Schneider JA, et al. Varied effects of age-related neuropathologies on the trajectory of late life cognitive decline. Brain 2017; 140: 804–12.

Boyle PA, Yu L, Wilson RS, Leurgans SE, Schneider JA, Bennett DA. Person-specific contribution of neuropathologies to cognitive loss in old age. Ann Neurol 2018; 83: 74–83.

Braak H, Alafuzoff I, Arzberger T, Kretzschmar H, Del Tredici K. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. Acta Neuropathol 2006; 112: 389–404.

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

Braak H, Del Tredici K, Rüb U, de Vos RA, Jansen Steur EN, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol Aging 2003; 24: 197–211.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 2015; 4: 7.

Chaudhury S, Brookes KJ, Patel T, Fallows A, Guetta-Baranes T, Turton JC, et al. Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment. Transl Psychiatry 2019; 9: 154.

Choi SW, O'Reilly PF. PRSice-2: polygenic Risk Score software for biobank-scale data. GigaScience 2019; 8: giz082.

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

Corlier F, Hafzalla G, Faskowitz J, Kuller LH, Becker JT, Lopez OL, et al. Systemic inflammation as a predictor of brain aging: contributions of physical activity, metabolic risk, and genetic risk. NeuroImage 2018; 172: 118–29.

Cruchaga C, Del-Aguila JL, Saef B, Black K, Fernandez MV, Budde J, et al. Polygenic risk score of sporadic late-onset Alzheimer's disease reveals a shared architecture with the familial and early-onset forms. Alzheimers Dement 2018; 14: 205–14.

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet 2016; 48: 1284–7.

Desikan RS, Fan CC, Wang Y, Schork AJ, Cabral HJ, Cupples LA, et al. Genetic assessment of age-associated Alzheimer disease risk: development and validation of a polygenic hazard score. PLoS Med 2017; 14: e1002258.

Duyckaerts C. Tau pathology in children and young adults: can you still be unconditionally baptist? Acta Neuropathol 2011; 121: 145–7.

Escott-Price V, Myers AJ, Huentelman M, Hardy J. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. Ann Neurol 2017; 82: 311–14.

Escott-Price V, Myers AJ, Huentelman M, Shoai M, Hardy J. Polygenic risk score analysis of Alzheimer's disease in cases without APOE4 or APOE2 alleles. J Prev Alzheimers Dis 2019; 6: 16–9.

Escott-Price V, Sims R, Bannister C, Harold D, Vronskaya M, Majounie E, et al. Common polygenic variation enhances risk prediction for Alzheimer's disease. Brain 2015; 138: 3673–84.

Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer disease meta analysis consortium. JAMA 1997; 278: 1349–56.

Felsky D, Patrick E, Schneider JA, Mostafavi S, Gaiteri C, Patsopoulos N, et al. Polygenic analysis of inflammatory disease variants and effects on microglia in the aging brain. Mol Neurodegen 2018; 13: 38.

Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975; 12: 189–98.

Francis PT, Costello H, Hayes GM. Brains for Dementia Research: evolution in a longitudinal brain donation cohort to maximize current and future value. JAD 2018; 66: 1635–44.

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

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

James BD, Bennett DA, Boyle PA, Leurgans S, Schneider JA. Dementia from Alzheimer disease and mixed pathologies in the oldest old. JAMA 2012; 307: 1798–800.

James BD, Wilson RS, Boyle PA, Trojanowski JQ, Bennett DA, Schneider JA. TDP-43 stage, mixed pathologies, and clinical Alzheimer's-type dementia. Brain 2016; 139: 2983–93.

Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet 2019; 51: 404–13.

Josephs KA, Whitwell JL, Weigand SD, Murray ME, Tosakulwong N, Liesinger AM, et al. TDP-43 is a key player in the clinical features associated with Alzheimer's disease. Acta Neuropathol 2014; 127: 811–24.

Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat Genet 2019; 51: 414–30.

Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet 2013; 45: 1452–8.

Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 2016; 48: 1443–8.

Louwersheimer E, Wolfsgruber S, Espinosa A, Lacour A, Heilmann-Heimbach S, Alegret M, et al. Alzheimer's disease risk variants modulate endophenotypes in mild cognitive impairment. Alzheimers Dement 2016; 12: 872–81.

Lupton MK, Strike L, Hansell NK, Wen W, Mather KA, Armstrong NJ, et al. The effect of increased genetic risk for Alzheimer's disease on hippocampal and amygdala volume. Neurobiol Aging 2016; 40: 68–77.

Mäkelä M, Kaivola K, Valori M, Paetau A, Polvikoski T, Singleton AB, et al. Alzheimer risk loci and associated neuropathology in a population-based study (Vantaa 85+). Neurol Genet 2018; 4: e211.

Marioni RE, Campbell A, Hagenaars SP, Nagy R, Amador C, Hayward C, et al. Genetic stratification to identify risk groups for Alzheimer's disease. JAD 2017; 57: 275–83.

Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, et al. GWAS on family history of Alzheimer's disease. Transl Psychiatry 2018; 8: 99.

Martiskainen H, Helisalmi S, Viswanathan J, Kurki M, Hall A, Herukka SK, et al. Effects of Alzheimer's disease-associated risk loci on cerebrospinal fluid biomarkers and disease progression: a polygenic risk score approach. JAD 2014; 43: 565–73.

Mirra SS, Heyman A, McKeel D, Sumi SM, Crain BJ, Brownlee LM, et al. The Consortium to Establish a Registry for Alzheimer's disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. Neurology 1991; 41: 479–86.

Montine TJ, Phelps CH, Beach TG, Bigio EH, Cairns NJ, Dickson DW, et al. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. Acta Neuropathol 2012; 123: 1–11.

Mormino EC, Sperling RA, Holmes AJ, Buckner RL, De Jager PL, Smoller JW, et al. Polygenic risk of Alzheimer disease is associated with early- and late-life processes. Neurology 2016; 87: 481–8.

Morris JC. The clinical dementia rating (CDR): current version and scoring rules. Neurology 1993; 43: 2412–14.

Mungas D, Tractenberg R, Schneider JA, Crane PK, Bennett DA. A 2-process model for neuropathology of Alzheimer's disease. Neurobiol Aging 2014; 35: 301–8.

Nag S, Yu L, Wilson RS, Chen EY, Bennett DA, Schneider JA. TDP-43 pathology and memory impairment in elders without pathologic diagnoses of Alzheimer's disease or FTLD. Neurology 2017; 88: 653–60.

Nasreddine ZS, Phillips NA, BÃ©Dirian VÃ©R, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc. 2005; 53: 695–9.

Nelson PT, Dickson DW, Trojanowski JQ, Jack CR, Boyle PA, Arfanakis K, et al. Limbic-predominant age-related TDP-43 encephalopathy (LATE): consensus working group report. Brain 2019; 142: 1503–27.

Reiman EM, Arboleda-Velasquez JF, Quiroz YT, Huentelman MJ, Beach TG, Caselli RJ, et al. Exceptionally low likelihood of Alzheimer's dementia in APOE2 homozygotes from a 5,000-person neuropathological study. Nat Commun 2020; 11: 667.

Robinson JL, Lee EB, Xie SX, Rennert L, Suh E, Bredenberg C, et al. Neurodegenerative disease concomitant proteinopathies are prevalent, age-related and APOE4-associated. Brain 2018; 141: 2181–93.

Sabuncu MR, Buckner RL, Smoller JW, Lee PH, Fischl B, Sperling RA. The association between a polygenic Alzheimer score and cortical thickness in clinically normal subjects. Cereb Cortex 2012; 22: 2653–61.

Schneider JA, Arvanitakis Z, Leurgans SE, Bennett DA. The neuropathology of probable Alzheimer disease and mild cognitive impairment. Ann Neurol 2009; 66: 200–8.

Selkoe DJ. The molecular pathology of Alzheimer's disease. Neuron 1991; 6: 487–98.

Tasaki S, Gaiteri C, Mostafavi S, De Jager PL, Bennett DA. The molecular and neuropathological consequences of genetic risk for Alzheimer's dementia. Front Neurosci 2018; 12: 699.

Thal DR, Rüb U, Orantes M, Braak H. Phases of A beta-deposition in the human brain and its relevance for the development of AD. Neurology 2002; 58: 1791–800.

Toledo JB, Cairns NJ, Da X, Chen K, Carter D, Fleisher A, et al. Clinical and multimodal biomarker correlates of ADNI neuropathological findings. Acta Neuropathol Commun 2013; 1: 65.

Tsuang D, Leverenz JB, Lopez OL, Hamilton RL, Bennett DA, Schneider JA, et al. APOE ε4 increases risk for dementia in pure synucleinopathies. JAMA Neurol 2013; 70: 223–8.

Uryu K, Nakashima-Yasuda H, Forman MS, Kwong LK, Clark CM, Grossman M, et al. Concomitant TAR-DNA-binding protein 43 pathology is present in Alzheimer disease and corticobasal degeneration but not in other tauopathies. J Neuropathol Exp Neurol 2008; 67: 555–64.

Wilson RS, Yu L, Trojanowski JQ, Chen EY, Boyle PA, Bennett DA, et al. TDP-43 pathology, cognitive decline, and dementia in old age. JAMA Neurol 2013; 70: 1418–24.

Yang HS, Yu L, White CC, Chibnik LB, Chhatwal JP, Sperling RA, et al. Evaluation of TDP-43 proteinopathy and hippocampal sclerosis in relation to APOE ε4 haplotype status: a community-based cohort study. Lancet Neurol 2018; 17: 773–81.

Yu L, Boyle PA, Leurgans S, Schneider JA, Bennett DA. Disentangling the effects of age and APOE on neuropathology and late life cognitive decline. Neurobiol Aging 2014; 35: 819–26.

ARTICLE

Meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex

Rebecca G. Smith[1,ɸ], Ehsan Pishva[1,2,ɸ], Gemma Shireby[1], Adam R. Smith[1], Janou A.Y. Roubroeks[1,2], Eilis Hannon[1], Gregory Wheildon[1], Diego Mastroeni[3], Gilles Gasparoni[4], Matthias Riemenschneider[5], Armin Giese[6], Andrew J. Sharp[7], Leonard Schalkwyk[8], Vahram Haroutunian[9,10,11], Wolfgang Viechtbauer[2], Daniel L.A. van den Hove[2,12], Michael Weedon[1], Danielle Brokaw[3], Paul T. Francis[1], Alan J Thomas[13], Seth Love[14], Kevin Morgan[15] Jörn Walter[4], Paul D. Coleman[3], David A. Bennett[16], Philip L. De Jager[17,18], Jonathan Mill[1], Katie Lunnon[1,*]

[1] University of Exeter Medical School, College of Medicine and Health, University of Exeter, Exeter, UK.

[2] Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience (MHeNS), Maastricht University, Maastricht, The Netherlands.

[3] Banner ASU Neurodegenerative Research Center, Biodesign Institute, Arizona State University, Tempe, Arizona, USA.

[4] Department of Genetics, University of Saarland (UdS), Saarbruecken, Germany

[5] Department of Psychiatry and Psychotherapy, Saarland University Hospital (UKS), Homburg, Germany

[6] Center for Neuropathology and Prion Research, Ludwig-Maximilians-University (LMU), Munich, Germany

[7] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA.

[8] School of Biological Sciences, University of Essex, Colchester, UK.

[9] Department of Psychiatry, The Icahn School of Medicine at Mount Sinai, New York, USA.

[10] Department of Neuroscience, The Icahn School of Medicine at Mount Sinai, New York, USA.

[11] JJ Peters VA Medical Center, Bronx, New York, USA.

[12] Laboratory of Translational Neuroscience, Department of Psychiatry, Psychosomatics and Psychotherapy, University of Wuerzburg, Würzburg, Germany.

[13] Institute of Neuroscience, Newcastle University, Newcastle Upon Tyne, UK

[14] Dementia Research Group, Institute of Clinical Neurosciences, School of Clinical Sciences, University of Bristol, Bristol, UK

[15] Human Genetics Group, University of Nottingham, Nottingham, UK

[16] Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA

[17] Center for Translational and Computational Neuroimmunology, Department of Neurology and Taub Institute, Columbia University Medical Center, New York, USA.

[18] The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

[Φ] These authors contributed equally to the study

* Corresponding author: Katie Lunnon, University of Exeter Medical School, RILD Building Level 3 South, Royal Devon and Exeter Hospital, Barrack Rd, Exeter. EX2 5DW. UK. E-mail: k.lunnon@exeter.ac.uk

ABSTRACT

Epigenome-wide association studies of Alzheimer's disease have highlighted neuropathology-associated DNA methylation differences, although existing studies have been limited in sample size and utilized different brain regions. Here, we combine data from six DNA methylomic studies of Alzheimer's disease (N=1,453 unique individuals) to identify differential methylation associated with Braak stage in different brain regions and across cortex. We identified 236 CpGs in the prefrontal cortex, 95 CpGs in the temporal gyrus and ten CpGs in the entorhinal cortex at Bonferroni significance, with none in the cerebellum. Our cross-cortex meta-analysis (N=1,408 donors) identified 220 CpGs associated with neuropathology, annotated to 121 genes, of which 84 genes had not been previously reported at this significance threshold. We have replicated our findings using two further DNA methylomic datasets consisting of a > 600 further unique donors. The meta-analysis summary statistics are available in our online data resource (www.epigenomicslab.com/ad-meta-analysis/).

INTRODUCTION

Alzheimer's disease (AD) is a chronic neurodegenerative disease that is accompanied by memory problems, confusion and changes in mood, behavior and personality. AD accounts for ~60% of dementia cases, which affected 43.8 million people worldwide in 2016[1]. The disease is defined by two key pathological hallmarks in the brain: extracellular plaques comprised of amyloid-beta protein and intracellular neurofibrillary tangles of hyperphosphorylated tau protein[2-4]. These neuropathological changes are thought to occur perhaps decades before clinical symptoms manifest and the disease is diagnosed[4]. AD is a multi-factorial and complex disease, with the risk of developing disease still largely unknown despite numerous genetic and epidemiological studies over recent years.

Several studies have suggested that epigenetic mechanisms may play a role in disease etiology. In recent years a number of epigenome-wide association studies (EWAS) have been performed in AD brain samples, which have predominantly utilized the Illumina Infinium HumanMethylation450K BeadChip (450K array) in conjunction with bisulfite-treated DNA to assess levels of DNA methylation in cortical brain tissue from donors with varying degrees of AD pathology[5-12]. Independently these studies have identified a number of loci that show robust differential DNA methylation in disease, and many of these overlap between studies, for example loci annotated to *ANK1*, *RHBDF2*, *HOXA3*, *CDH23* and *RPL13* have been consistently reported. Here we have performed a meta-analysis of six independent existing EWAS of AD brain[5-8,10,12], totalling 1,453 independent donors, to identify robust and consistent differentially methylated loci associated with Braak stage, used as a measure of neurofibrillary tangle spread through the brain, before replicating these signatures in two further independent DNA methylation datasets. Our meta-analysis approach provides additional power to detect DNA methylomic variation associated with AD pathology at novel loci, in addition to providing further replication of loci that have been previously identified in the smaller independent EWAS.

RESULTS

Pathology-associated DNA methylation signatures in discrete cortical brain regions

We identified six EWAS of DNA methylation in AD that had been generated using the 450K array and had a cohort size of > 50 unique donors. All had data on Braak stage available, which we used as a standardized measure of tau pathology spread through the brain (**Table 1**). We were interested in identifying epigenomic profiles associated with Braak stage in specific brain regions, leveraging additional power by meta-analysing multiple studies to identify novel loci. To this end, we performed an EWAS in each available tissue and cohort separately, looking for an association between DNA methylation and Braak stage, whilst controlling for age and sex (all tissues) and neuron/glia proportion (cortical bulk tissues only), with surrogate variables added as appropriate to reduce inflation. For discovery, we then used the estimated effect size (ES) and standard errors (SEs) from these six studies (N = 1,453 unique donors) for a fixed-effect inverse variance weighted meta-analysis separately for each tissue (prefrontal cortex: three cohorts, N = 959; temporal gyrus: four cohorts, N = 608, entorhinal cortex: two cohorts, N = 189 cerebellum: four cohorts, N = 533) (**Supplementary Figure 1**).

The prefrontal cortex represented our largest dataset (N = 959 samples) and we identified 236 Bonferroni significant differentially methylated positions (DMPs) ($P < 1.238 \times 10^{-7}$ to account for 403,763 probes), of which 193 were annotated to 137 genes, with 43 unannotated loci based on Illumina UCSC annotation (**Figure 1a**, **Supplementary Figure 2, Supplementary Table 1**). Previous EWAS of the prefrontal cortex have consistently reported the *HOXA* gene cluster as a region that is hypermethylated in AD[6,7], with a cell-type specific EWAS demonstrating this is neuronal-derived[11]. Indeed, the most significant DMP in the prefrontal cortex in our meta-analysis resided in *HOXA3* (cg22962123: ES [defined as the methylation difference between Braak 0 and Braak VI] = 0.042, $P = 5.97 \times 10^{-15}$), with a further 16 of the Bonferroni significant DMPs also annotated to this gene. This locus appeared to be particularly hypermethylated with higher Braak stage in the prefrontal cortex, and to a slightly lesser extent in the temporal gyrus (**Supplementary Figure 3**). There was no significant difference in methylation at this locus in the entorhinal cortex (P = 0.864), which is interesting given that the entorhinal cortex may succumb to pathology early in the disease process (Braak stage III). Of the 236 prefrontal cortex DMPs, 92% (217 probes) were nominally significant ($P < 0.05$) in the temporal gyrus, of which 12% (28 probes) were Bonferroni significant, whilst 9% (22 probes) were nominally significant in the entorhinal cortex, with 1% (3 probes) reaching Bonferroni significance (**Figure 1b**). The effect sizes for the 236 Bonferroni significant prefrontal cortex DMPs were correlated with the effect sizes for the same probes in both the temporal gyrus (Pearson's correlation coefficient (r) = 0.94, $P = 6.17 \times 10^{-112}$) and entorhinal cortex (r = 0.58, $P = 1.80 \times 10^{-22}$) and were enriched for probes with the same direction of effect (sign test: temporal gyrus $P = 5.07 \times 10^{-67}$, entorhinal cortex $P = 6.88 \times 10^{-26}$) (**Supplementary Figure 4).** For the 236 Bonferroni significant prefrontal cortex DMPs these had the largest effect sizes in the prefrontal cortex, with a smaller effect size in the temporal gyrus and entorhinal cortex (**Figure 1c**). Of these 236 DMPs, 29 of these had being previously reported at Bonferroni significance in previous publications on the individual cohorts[5-7,12], including one probe annotated to *ANK1,* one probe annotated to *HOXA3,* one probe annotated to *PPT2/PRRT1* and two probes annotated to *RHBDF2*, amongst others. However, our approach has identified 207 novel Bonferroni significant DMPs (although several had been reported in previous studies at a more relaxed significance threshold, or in regional analyses). This included several additional probes residing in genes already identified (from another probe) in earlier studies, for example a further 16 probes in *HOXA3* and two probes in *PPT2/PRRT1*. Interestingly, we also identified a number of novel genes, including some which featured multiple Bonferroni significant DMPs including for example seven probes in *AGAP2* and five probes in *SLC44A2*, amongst others. One other noteworthy novel Bonferroni significant DMP in the prefrontal cortex was cg08898775 (ES = 0.019, P = 4.03 x 10^{-9}), annotated to *ADAM10*, which encodes for α-secretase which cleaves APP in the non-amyloidogenic pathway. A differentially methylated region (DMR) analysis, which allowed us to identify areas of the genome consisting of $\geq$ 2 DMPs, revealed 262 significant DMRs in the prefrontal cortex (**Supplementary Table 2**), the most significant containing 20 probes and located in *HOXA3* (chr7:27,153,212-27,155,234: Sidak-corrected p = 8.21 x 10^{-50}, **Supplementary Figure 5**), as well as several other DMRs in the *HOXA* gene cluster.

A meta-analysis of temporal gyrus EWAS datasets (N = 608 samples) identified 95 Bonferroni significant probes, of which 75 were annotated to 53 genes, with 20 unannotated probes using Illumina UCSC annotation (**Figure 1a**, **Supplementary Figure 6**, **Supplementary Table 3**). The most significant probe was cg11823178 (ES = 0.029, P = 3.97 x 10^{-16}, **Supplementary Figure 7**), which is annotated to the *ANK1* gene, with the fifth (cg05066959: ES = 0.042, P = 4.58 x 10^{-13}) and 82nd (cg16140558: ES = 0.013, P = 8.44 x 10^{-8}) most significant probes in the temporal gyrus also being annotated to nearby CpGs in that gene. This locus has been widely reported to be hypermethylated in AD from prior EWAS[5,6,8,12], as well as in other neurodegenerative diseases such as Huntington's disease and Parkinson's disease[13]. Another noteworthy gene is *RHBDF2*, where five Bonferroni significant DMPs in the temporal gyrus were annotated to (cg05810363: ES = 0.029, P = 2.25 x 10^{-11}; cg13076843: ES = 0.031, P = 2.97 x 10^{-11}; cg09123026: ES = 0.012, P = 3.46 x 10^{-9}; cg12163800: ES = 0.025, P = 5.85 x 10^{-9}; cg12309456: ES = 0.016, P = 1.33 x 10^{-8}); and which has been highlighted in previous EWAS in AD in the individual cohorts[5,6,12]. Of the 95 Bonferroni significant DMPs in the temporal gyrus, 88% (84 probes) were nominally significant in the prefrontal cortex, of which

29% (28 probes) were Bonferroni significant, whilst 54% (51 probes) were nominally significant in the entorhinal cortex, of which 6% (6 probes) were Bonferroni significant (**Figure 1b**). Given the high degree of overlapping significant loci between the temporal gyrus and other cortical regions, it was not surprising that the ES of the 95 Bonferroni significant temporal gyrus probes were highly correlated with the ES of the same loci in both the prefrontal cortex (r = 0.91, P = 5.09 x 10^{-38}) and entorhinal cortex (r = 0.77, P = 4.02 x 10^{-20}) and were enriched for the same direction of effect (sign test: prefrontal cortex P = 5.05 x 10^{-29}, entorhinal cortex = 2.30 x 10^{-25}) (**Supplementary Figure 8**). The majority of the 95 Bonferroni significant DMPs in the temporal gyrus were hypermethylated, and the mean ES was greater in the temporal gyrus than the prefrontal cortex or entorhinal cortex (**Figure 1c**). Thirty-two of the 95 Bonferroni significant DMPs in the temporal gyrus have been previously reported to be significantly differentially methylated in published EWAS, including for example three probes in *ANK1* and the five probes in *RHBDF2.* Our meta-analysis approach in the temporal gyrus has identified 63 novel DMPs (at Bonferroni significance), including some novel genes with multiple DMPs, for example four probes in *RGMA* and two probes in *CCND1*, amongst others. Finally, our regional analysis highlighted 104 DMRs (**Supplementary Table 4**); the top DMR resided in the *ANK1* gene (chr8:41,519,308-41,519,399) and contained two probes (Sidak-corrected P = 1.72 x 10^{-21}) (**Supplementary Figure 9**). The five DMPs in *RHBDF2* that we already highlighted also represented a significant DMR (Sidak-corrected P = 8.47 x 10^{-21}), with three other genomic regions containing large, significant DMRs consisting of ≥ 10 probes, such as *MCF2L* (chr13:113698408-113699016 [10 probes], Sidak-corrected P = 1.16 x 10^{-19}), *PRRT1/PPT2* (chr6:32120773-32121261 [17 probes], Sidak-corrected P = 4.90 x 10^{-15}) and *HOXA5* (chr7:27184264-27184521 [10 probes], Sidak-corrected P = 1.60 x 10^{-7}).

The final cortical region we had available was the entorhinal cortex (N = 189), where we identified ten Bonferroni significant probes in our meta-analysis, all of which were hypermethylated with higher Braak stage (**Figure 1a**, **Supplementary Figure 10, Supplementary Table 5**). These ten probes were annotated to eight genes (Illumina UCSC annotation), with two Bonferroni significant probes residing in each of the *ANK1* and *SLC15A4* genes. As with the temporal gyrus, the most significant DMP was cg11823178 (ES = 0.045, P = 5.22 x 10^{-10}, **Supplementary Figure 7**), located within the *ANK1* gene, with the fourth most significant DMP being located within 100bp of that CpG (cg05066959: ES = 0.062, P = 2.93 x 10^{-9}). In total, eight of the ten DMPs in the entorhinal cortex had been reported previously at Bonferroni significance, including the two probes in *ANK1.* Two of the Bonferroni significant

DMPs we identified in the entorhinal cortex were novel CpGs (cg11563844: *STARD13*, cg04523589: *CAMP*), having not been reported as Bonferroni significant in previous EWAS. Of the ten entorhinal cortex probes, 90% (9 probes) were nominally significant in the temporal gyrus, of which 60% (6 probes) were Bonferroni significant, whilst 70% (7 probes) were nominally significant in the prefrontal cortex, of which 30% (3 probes) were Bonferroni significant (**Figure 1b**). Of the four DMPs that were Bonferroni significant in only the entorhinal cortex, three of these were nominally significant in at least one other tissue, with just one probe unique to the entorhinal cortex, annotated to *STARD13* (cg11563844, ES = 0.027, P = 1.07 x 10^{-8}). The effect sizes of the ten Bonferroni significant DMPs in the entorhinal cortex were significantly correlated with the effect size of the same probes in the prefrontal cortex (r = 0.74, P = 0.01) and temporal gyrus (r = 0.85, P = 1.52 x 10^{-3}) and were enriched for the same direction of effect (sign test: prefrontal cortex P = 0.021, temporal gyrus P = 1.95 x 10^{-3}) (**Supplementary Figure 11**). The ten DMPs were hypermethylated in all three cortical regions, with the greatest Braak-associated ES in the entorhinal cortex (**Figure 1c**). A regional analysis identified seven DMRs (**Supplementary Table 6**); the top three DMRs (*RHBDF2:* chr17:74,475,240-74,475,402 [five probes], P = 7.68 x 10^{-14}, **Supplementary Figure 12;** *ANK1*: chr8:41519308-41519399 [two probes], P = 4.89 x 10^{-13}; *SLC15A4*: chr12:129281444-129281546 [three probes], P = 5.24 x 10^{-12}) were significant in at least one of the other cortical regions we meta-analyzed.

To date, a few independent EWAS in AD have been undertaken in the cerebellum and none of these have reported any Bonferroni significant DMPs. In our meta-analysis we identified no Bonferroni significant DMPs, nor any DMRs in the cerebellum (**Supplementary Figure 13**), despite this analysis including 533 independent samples. There was no correlation of the ES for the Bonferroni significant DMPs we had identified in the meta-analyses of the three cortical regions with the ES of the same probes in the cerebellum (prefrontal cortex: r = 0.11, P = 0.08; temporal gyrus: r = 0.14, P = 0.17; entorhinal cortex: r = 0.48, P = 0.16; **Supplementary Figure 14**).

220 CpGs are differentially methylated across the cortex in AD

We were interested in combining data from across the different cortical tissues to identify common differentially methylated loci across the cortex and also to provide more power by utilizing data from 1,408 unique individuals with cortical EWAS data available. As multiple cortical tissues were available for some cohorts, a mixed-effects model was utilized. In this analysis we controlled for age, sex and neuron/glia proportion, with surrogate variables added as appropriate to reduce inflation. Using this approach, we identified 220 Bonferroni significant probes, of which 168 were annotated to 121 genes, with 52 DMPs unannotated using Illumina UCSC annotation. **Figure 2a**, **Figure 2b**, **Table 2**, **Supplementary Table 7, Supplementary Figure 15**). All of the 220 probes were nominally significant ($P < 0.05$) in ≥ two cohorts, with ten of these probes being nominally significant in all six cohorts (**Supplementary Figure 16**), which included single probes annotated to *ANK1*, *ABR*, *SPG7* and *WDR81*, two probes in *DUSP27*, three probes in *RHBDF2* and one unannotated probe. We observed similar DNA methylation patterns across all cortical cohorts and tissues for the 220 probes with 219 of the 220 DMPs showing the same direction of effect in at least five cohorts. In total, 154 of the DMPs were hypermethylated, with 66 hypomethylated, representing an enrichment for hypermethylation ($P = 4.85 \times 10^{-10}$). This pattern of methylation was evident across all cortical tissues but was not seen in the cerebellum (**Supplementary Figure 17**). Of the 220 DMPs we identified, 46 of these have been previously reported at Bonferroni significance in published EWAS, including multiple previously identified probes in *ANK1* (cg05066959, cg11823178), *MCF2L* (cg07883124, cg09448088), *PCNT* (cg00621289, cg04147621, cg23449541) and *RHBDF2* (cg05810363, cg12163800, cg12309456, cg13076843). The most significant probe we identified in our cross-cortex analysis was cg12307200 (**Table 2**, ES = -0.015, $P = 4.48 \times 10^{-16}$), which is intergenic and found at chr3:188664632, located between the *TPRG1* and *LPP* genes and had been previously reported at Bonferroni significance by De Jager and colleagues with respect to neuritic plaque burden[6] and by Brokaw and colleagues with respect to post-mortem diagnosis[12]. Our cross-cortex meta-analysis approach has identified 174 novel DMPs (at Bonferroni significance), annotated to 102 genes. Although 11 of these genes had previously been reported at Bonferroni significance (another probe within that gene), the remaining 96 genes represent robust novel loci in AD. Many of these novel differentially methylated genes had multiple Bonferroni significant probes, for example five probes in *AGAP2*, three probes in *HOXB3* and *SLC44A2,* and two probes in *CDH9*, *CPEB4*, *DUSP27*, *GCNT2*, *MAMSTR*, *PTK6*, *RGMA*, *RHOB*, *SMURF1*, *THBS1*, *ZNF238* and *ZNF385A* (**Supplementary Table 7**). Although some of these loci may have been reported in earlier AD EWAS, none of these were at Bonferroni significance and so here represent robust novel loci.

We were interested to investigate whether specific functional pathways were differentially methylated in AD cortex and so performed a gene ontology pathway analysis of the 121 genes annotated to the 220 Bonferroni significant cross-cortex DMPs. We highlighted epigenetic dysfunction in numerous pathways (at nominal significance), interestingly including a number of developmental pathways, mainly featuring the *HOXA* and *HOXB* gene clusters (**Supplementary Table 8**). Given that we identified multiple DMPs in some genes, we were interested to investigate the correlation structure between probes in close proximity to each other to establish how many independent signals we had identified. Using a method developed to identify single nucleotide polymorphisms (SNPs) in linkage disequilibrium (LD)[14], we collapsed the 220 Bonferroni significant loci into 165 independent (non-highly correlated [r < 0.6 over 1mb]) signals (**Supplementary Table 9**). We found that the largest reduction in signals occurred in the *HOXA* and *HOXB* gene clusters, with the 18 DMPs in the *HOXA* region representing only two independent signals, whilst the seven DMPs in the *HOXB* region represented one independent signal.. Next we undertook a formal regional analysis to identify genomic regions of multiple adjacent DMPs and identified 221 DMRs, with the top DMR containing 11 probes and covering the *HOXA* region (chr7:27,153,212-27,154,305: P = 3.84 x 10^{-35}) (**Figure 2c**, **Supplementary Table 10**). The *HOXA* gene cluster further featured a number of times in our DMR analysis; four of the ten most significant DMRs fell in this genomic region, including DMRs spanning four probes (chr7:27146237-27146445: P = 4.11 x 10^{-27}), 33 probes (chr7:27183133-27184667: P = 2.22 x 10^{-20}) and ten probes (chr7:27143235-27143806: P = 1.75 x 10^{-18}).

Replication of pathology associated DMPs in the cortex

To replicate our findings and to determine the cellular origin of DNA methylomic differences we used the estimated coefficients and SEs for these 220 probes generated in a seventh independent (“Munich”) cohort, which consisted of 450K data generated in the prefrontal cortex (N = 45) and sorted neuronal and non-neuronal nuclei from the occipital cortex (N = 26) (**Table 1**). This cohort had not been used in our discovery analyses as < 50 samples were available. Notably, we identified a similar pattern of Braak-associated DNA methylation changes for the 220 Bonferroni significant cross-cortex probes in this replication cohort, with a significantly correlated effect size between the discovery dataset and the replication prefrontal cortex (r = 0.64, P = 5.24 x 10^{-27}), neuronal (r = 0.45, P = 1.56 x 10^{-12}) and non-neuronal datasets (r = 0.79, P= 1.43 x 10^{-47}) with a similar enrichment for the same direction of effect (sign test: prefrontal cortex P = 4.59 x 10^{-28}, neuronal P = 6.13 x 10^{-15}, non-neuronal P = 1.06 x 10^{-42}) (**Figure 3a**). The most significant probe from the cross-cortex meta-analysis (cg12307200) showed consistent hypomethylation in disease in all cohorts in all cortical brain regions, with this direction of effect replicated in the prefrontal cortex and non-neuronal nuclei samples, but not the neuronal nuclei samples, suggesting that this is primarily driven by non-neuronal cell types, which are likely to be glia (**Figure 3b**). We have developed an online database (www.epigenomicslab.com/ad-meta-analysis/), which can generate a forest plot showing the ES and SE across any of the discovery cohorts and the Munich sample types for any of the 403,763 probes that passed our quality control. This allows researchers to determine the consistency of effects across cohorts for a given CpG site as well as the likely cellular origin of the signature. In addition, our tool can generate mini-Manhattan plots to show DMRs utilizing the summary statistics from the cross-cortex meta-analysis.

Finally, we had access to DNA methylation data generated in an eighth independent ("Brains for Dementia Research [BDR]") cohort. This consisted of Illumina Infinium HumanMethylation EPIC BeadChip (EPIC array) data in the prefrontal cortex in 590 individuals[15]. As this is the successor to the 450K array (which had been used for the other seven cohorts), there are some differences in genome coverage, and for the 220 Bonferroni significant cross-cortex DMPs we had identified in the discovery cohorts, only 208 probes are also present on the EPIC array. For these overlapping 208 probes, we observed a significantly correlated effect size between the discovery dataset and the BDR dataset (r = 0.53, P = 4.13 x 10^{-16}) (**Figure 3c**), with all 208 probes showing the same direction of effect (sign test P = 4.86 x 10^{-63}).

Cross-cortex AD-associated DMPs are enriched in specific genomic features

To identify if the cross-cortex DMPs reside in specific genomic features, we used a Fisher's exact test to look for an enrichment of the 220 DMPs using Slieker annotations[16] (**Supplementary Table 11**, **Supplementary Figure 18**). We observed a significant over representation of Bonferroni significant DMPs in CpG islands of gene bodies (odds ratio [OR] = 3.199, P = 4.76 x 10^{-10}), and in CpG island shelves and non-CpG island areas of proximal promoters (OR = 3.571, P = 9.09 x 10^{-3} and OR = 1.641, P = 0.03, respectively). However, DMPs located in CpG islands in the proximal promoter were under-represented (OR = 0.353, P = 2.08 x 10^{-6}). There was a significant over representation of the 220 cross-cortex DMPs in the first exon (OR = 1.80, P = 0.02), with an under enrichment within 1500bp of the transcription start site (OR = 0.49, P = 3.82 x 10^{-3}) (**Supplementary Table 12**, **Supplementary Figure 19**).

DNA methylomic signatures in the cortex can explain variance in the degree of pathology

We were interested to investigate whether the Braak-associated DNA methylation patterns we had identified across the cortex could accurately predict the pathological load of a brain sample and how much variance this explained. To this end we took samples within the discovery cohorts with either low pathology (Braak 0-II “controls”: N = 407), or high pathology (Braak V-VI “AD”: N = 589) and used these as a “training” dataset. We then used elastic net regression to identify 110 probes in the 220 cross-cortex Bonferroni significant loci (**Supplementary Table 13)** that were able to explain the most variance between post-mortem low pathology “control” from high pathology “AD” status in our training dataset (N = 996) (**Supplementary Table 14**, **Figure 4**). In our training data, we achieved an Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) of 94.33% (CI = 92.88-95.64%, variance explained = 71.11%).. We then tested its performance in the Munich replication samples (N = 38) and the BDR replication samples (N = 454), where it achieved an AUC of 73.95% (CI = 55.17-88.89%, variance explained = 20.18%) and 70.36% (CI = 65.52-75.12%, variance explained = 15.87%), respectively (**Supplementary Table 14**, **Figure 4**).

DNA methylation signatures in AD cortex are largely independent of genetic effects

DNA methylomic variation can be driven by genetic variation via methylation quantitative trait loci (mQTLs). To explore whether SNPs may be driving the methylation differences we observed (in *cis*) we used the xQTL resource to identify *cis*-mQTLs associated with the 220 Bonferroni significant cross-cortex DMPs[17]. We identified 200 Bonferroni corrected mQTLs, which were associated with DNA methylation at 18 of the 220 cross-cortex DMPs (**Supplementary Table 15**). This suggests that the majority of Braak-associated DMPs are not the result of genetic variation in *cis*. None of these mQTLs overlapped with lead SNPs (or SNPs in LD) identified in the most recent genome-wide association study (GWAS) of diagnosed late-onset AD from Kunkle *et al*[18]. Next, we were interested in exploring whether DNA methylation is enriched in genes known to harbor AD-associated genomic risk variants. Using the AD variants from Kunkle *et al*[18] we examined the enrichment of Braak-associated DNA methylation in 24 LD blocks harboring risk variants. Twenty of these LD blocks contained > 1 CpG site on the 450K array and using Brown’s method we combined P values within each of these 20 genomic regions. We observed Bonferroni-adjusted significant enrichment in the cross-cortex data in the *HLADRB1* (Chr6: 32395036-32636434: adjusted P = 1.20 x 10^{-3}), *SPI1* (Chr11: 47372377- 47466790, adjusted P = 5.76 x 10^{-3}), *SORL1* (Chr11: 121433926- 121461593, adjusted P = 0.019), *ABCA7* (Chr19: 1050130- 1075979, adjusted P

= 0.022) and *ADAM10* (Chr15: 58873555- 59120077, adjusted P = 0.022) LD regions (**Supplementary Table 16**).

DISCUSSION

This study represents the first meta-analysis of AD EWAS utilizing six published independent sample cohorts with a range of cortical brain regions and cerebellum available as a discovery dataset. Two further independent cortical datasets where then used for replication, including data from sorted nuclei populations. Our data can be explored as part of an online searchable database, which can be found on our website (https://www.epigenomicslab.com/ad-meta-analysis). By performing a meta-analysis within each tissue, we have been able to identify 236, 95 and ten Bonferroni significant DMPs in the prefrontal cortex, temporal gyrus and entorhinal cortex, respectively. Although far fewer loci were identified in the entorhinal cortex compared to the other cortical regions, this is likely due to the reduced sample size in this tissue. In the cerebellum despite meta-analyzing > 500 unique samples, we identified no Braak-associated DNA methylation changes. Furthermore, there was no correlation of the effect size of Bonferroni significant DMPs identified in any of the cortical regions with the effect size of the same probes in the cerebellum. Taken together, this suggests that DNA methylomic changes in AD are cortex cell type specific. This observation is interesting as the cerebellum is said to be "spared" from AD pathology, with an absence of neurofibrillary tangles, although some diffuse amyloid-beta plaques are reported[19]. Interestingly, a recent spatial proteomics study of AD reported a large number of protein changes in the cerebellum in AD; however, the proteins identified were distinct from other regions examined and thus the authors suggested a potential protective role[20].

Although many loci showed similar patterns of Braak-associated DNA methylation across the different cortical regions, some loci did show some regional specificity. In order to identify CpG sites that showed common DNA methylation changes in disease we performed a cross-cortex meta-analysis. Using this approach we identified 220 Bonferroni significant probes associated with Braak stage of which 46 probes had been previously reported at Bonferroni significance in the individual cohort studies that we had used for our meta-analysis, for example two probes in *ANK1*, four probes in *RHBDF2* and one probe in *HOXA3*, amongst others. Interestingly, our approach did identify 174 novel CpGs, corresponding to 102 unique genes, of which 84 genes had not been previously reported at Bonferroni significance in any of the previously published AD brain EWAS, highlighting the power of our meta-analysis approach for nominating new loci. This included 15 novel genes with at least two Bonferroni significant DMPs each, including five probes in *AGAP2*, three probes in *SLC44A2* and two probes each in *CDH9*, *CPEB4, DUSP27*, *GCNT2*, *MAMSTR, PTK6*, *RGMA*, *RHOB*, *SMURF1*, *THBS1*, *ZNF238* and *ZNF385A*. These genes had not been identified previously in an AD EWAS at this significance threshold, although a number of these genes had been previously identified from DMR analyses, which have a less stringent threshold. However, we did identify one novel gene (*HOXB3*) with three Bonferroni significant DMPs, which had not been identified at this significance threshold in previous EWAS DMP or DMR analyses in AD brain. The nomination of loci in the *HOXB* gene cluster is interesting; a recent study of human Huntington's disease brain samples also highlighted significantly increased *HOXB3* gene expression in the prefrontal cortex[21], an interesting observation given that both AD and Huntington's disease are disorders that feature dementia. Furthermore, we have recently reported AD-associated hypermethylation of the *HOXB6* gene in AD blood samples[22]. Our pathway analysis highlighted differential methylation in a number of developmental pathways, mainly featuring the *HOXA* and *HOXB* gene clusters. Although it is unclear why developmental genes may be changed in a disease that primarily affects the elderly, it has been implied that genes such as these may be involved in neuroprotection after development[23]. A number of the other novel genes with multiple DMPs are also biologically relevant in the context of AD, for example *GCNT2* was recently shown to be differentially expressed in the Putamen between males and females with AD[24]. Interestingly, some of the protein products of genes we identified have also been previously linked with AD; PTK6 is a protein kinase whose activity has been shown to be altered in post-mortem AD brain[25]. Similarly, RGMA has been shown to be increased in AD brain, where it accumulated in amyloid-beta plaques[26].

Our genomic enrichment analyses identified an over representation of hypermethylated loci in AD and methylation in specific genomic features, for example CpG islands in gene bodies, and shelves and non-CpG island regions in proximal promoters. We demonstrated that the majority of DMPs we identified (N = 202) were not driven by genetic variation as only 18 of the 220 CpG sites have reported mQTLs. However, we did observe a significant enrichment of cross-cortex loci in the LD regions surrounding the AD-associated genetic variants *HLADRB1*, *SPI1*, *SORL1*, *ABCA7* and *ADAM10* after controlling for multiple testing Finally, we have developed a classifier that could accurately predict control samples with low pathology, from those with a post-mortem AD diagnosis due to high pathology using methylation values for 110 of the 220 Bonferroni significant probes, further highlighting that distinct genomic loci reproducibly show epigenetic dysfunction in AD cortex. Although the clinical utility of such a classifier is limited as it is developed in post-mortem cortical brain tissue, it does illustrate that specific robust patterns of DNA methylation differences occur as the disease progresses. These signatures require further investigation as they could represent novel therapeutic targets, particularly given the classifier had an AUC > 70% in all the training and replication datasets. However, it is worth noting that the variance explained by the 110 CpG signature was lower in the replication datasets than the discovery samples, which could be due to a low sample number (Munich) or the different Illumina array platform (BDR).

There are some limitations with our study. First, as we have largely utilized methylation data generated in bulk tissue, this will contain a mixture of different cell types. Furthermore, it is known that the proportions of the major brain cell types are altered in AD, with reduced numbers of neurons and increased glia. As such, it is possible that the identified DNA methylation changes represent a change in cell proportions. To address this, we have included neuron/glia proportions as a co-variate in our models to minimize bias and used data from sorted cell populations as part of our replication. Although this is the optimal strategy for the current study given the EWAS data had already been generated, future EWAS should be undertaken on sorted cell populations with larger sample numbers than the Munich replication cohort, or ideally at the level of the single cell. It is important to note that the data from the sorted nuclei populations in the Munich replication cohort were generated in the occipital cortex, which was not a bulk tissue used for any of the discovery cohorts. In the future it would be interesting to explore whether different disease-associated DNA methylation signatures were observed in neurons and glia isolated from different cortical brain regions. Second, our study has utilized previously generated EWAS data generated on the 450K array or EPIC array. Although the Illumina array platform has been the most widely used platform for EWAS to date, it is limited to only analyzing a relatively small proportion of the potential methylation sites in the genome (~400,000 on the 450K array) and given the falling cost of sequencing, future studies could exploit this by performing reduced representation bisulfite sequencing to substantially increase the coverage. In our study we have primarily used the UCSC annotation provided by Illumina to identify the gene relating to each DMP. However, this can lead to the annotation of overlapping genes, or no gene annotation, which can make it difficult to establish the gene of interest in the absence of functional studies. Our study has primarily focused on the results of a fixed-effects meta-analysis, as the majority of Bonferroni-significant DMPs do not display a high degree of heterogeneity. However, ~15% of the cross-cortex DMPs did have a significant heterogeneity P value and in this instance, it is worthwhile also considering the results of the random-effects meta-analysis. Although this heterogeneity could be driven by differences between cohorts, it is also plausible that it may be driven by tissue-specific effects as we used different cortical brain regions in the model. For example cg22962123 annotated to the *HOXA3* gene has a significant heterogeneity P value in the cross-cortex meta-analysis, but we had already shown this loci to be differentially methylated in the prefrontal cortex and temporal gyrus, but not the entorhinal cortex in our intra-tissue meta-analysis.

Another limitation of our study is that we have focused our analyses on Braak (neurofibrillary tangle)-associated methylation changes, as this measure was available in all cohorts. Given that amyloid-beta is another neuropathological hallmark of AD, it would also be of interest to identify neuritic plaque-associated DMPs. Unfortunately, this was not feasible in the current study as this measure was not available in all samples. In a similar vein, we did not exclude individuals with mixed pathology, or protein hallmarks of other neurodegenerative diseases, such as the presence of lewy bodies, or TDP-43 pathology. In the future, larger meta-analyses should stratify by the presence of these protein aggregates, particularly given that very few EWAS have been undertaken in other dementias. Indeed, only three DNA methylomic studies have been undertaken in cortical samples of individuals with other dementias to date[27-30], with none of these studies utilizing > 15 individuals for EWAS. Further studies exploring common and unique DNA methylation signatures and our classifier in other diseases characterized by dementia will be vital for identifying disease-specific epigenetic signatures that could represent novel therapeutic targets. Finally, one key issue for epigenetic studies in post-mortem tissue is the issue of causality, where it is not possible to determine whether disease-associated epigenetic loci are driving disease pathogenesis, or are a consequence of the disease, or even the medication used for treatment. One method that can be used to address this is Mendelian Randomization[31] however, this does require the CpG site to have a strong association with a SNP. Given that we only identified mQTLs at 18 of the 220 Bonferroni significant cross-cortex DMPs, this approach is not suitable for most of the loci we identified. At an experimental level establishing causality is difficult to address in post-mortem human studies, and therefore longitudinal studies in animal models, or modelling methylomic dysfunction through epigenetic editing *in vitro* will be useful approaches to address these issues. In addition, examining DNA methylation signatures in brain samples in pre-clinical individuals (*i.e.* during midlife) will be important for establishing the temporal pattern of epigenetic changes relative to the pathology.

In summary we present the first meta-analyses of AD EWAS, highlighting numerous Bonferroni significant DMPs in the individual cortical regions and across the cortex, but not in the cerebellum, which were replicated in two independent cohorts. A number of these loci are novel and warrant further study to explore their role in disease etiology. We highlight that the nominated epigenetic changes are largely independent of genetic effects, with only 18 of the 220 Bonferroni significant DMPs showing a mQTL. We provide the first evidence that robust epigenomic changes in the cortex can predict the level of pathology in a sample. Looking to the future it will be important to explore the relationship between DNA methylation and gene expression in AD brain.

METHODS

Cohorts

Six sample cohorts were used for "discovery" in this study as they all had DNA methylation data generated on the 450K array for > 50 donors, enabling us to take a powerful meta-analysis approach to identify DNA methylation differences in AD. As our analyses focused specifically on neuropathology (tau)-associated differential methylation, inclusion criteria for all samples used in the "discovery" or "replication" cohorts was having post-mortem neurofibrillary tangle Braak stage available. For each discovery sample cohort DNA methylation was quantified using the 450K array. The "London 1" cohort comprised of prefrontal cortex, superior temporal gyrus, entorhinal cortex, and cerebellum tissue obtained from 113 individuals archived in the MRC London Neurodegenerative Disease Brain Bank and published by Lunnon *et al.*[5]. The "London 2" cohort comprised entorhinal cortex and cerebellum samples obtained from an additional 95 individuals from the MRC London Neurodegenerative Disease Brain Bank published by Smith and colleagues[8]. The "Mount Sinai" cohort comprised of prefrontal cortex and superior temporal gyrus tissue obtained from 146 individuals archived in the Mount Sinai Alzheimer's Disease and Schizophrenia Brain Bank published by Smith and colleagues[7]. The "Arizona 1" cohort consisted of 302 middle temporal gyrus and cerebellum samples from The Sun Health Research Institute Brain Donation Program[32] published by Brokaw *et al.*[12]. The "Arizona 2" cohort consisted of an additional 88 temporal gyrus and cerebellum samples from Lardonije *et al.*[10]. The "ROSMAP" cohort consisted of 709 samples from the Rush University

Medical Center: Religious Order Study (ROS) and the Memory and Aging Project (MAP), which were previously published by De Jager and colleagues[6]. For replication purposes we used two further replication datasets. The “Munich” cohort” from Neurobiobank Munich (NBM), which had bulk prefrontal cortex 450K array data from 45 donors, and 450K array data from fluorescence-activated cell sorted neuronal and non-neuronal (glial) populations from the occipital cortex from 26 donors as described by Gasparoni *et al.*[11]. The “Brains for Dementia Research (BDR)” cohort consisted of bulk prefrontal cortex Illumina Infinium EPIC array data from 590 donors, as described by Shireby *et al*[15]. Demographic information for all eight cohorts is available in **Table 1**.

Data quality control and harmonization

All computations and statistical analyses were performed using R 3.5.2[33] and Bioconductor 3.8[34]. A *MethylumiSet* object was created from iDATs using the methylumi package[35] and *RGChannelSet* object was created using the minfi package[36]. Samples were excluded from further steps if (a) the mean background intensity of negative probes < 1,000, (b) the mean detection P values > 0.005, (c) the mean intensity of methylated or unmethylated signals were three standard deviations above or below the mean, (d) the bisulfite conversion efficiency < 80%, (e) there was a mismatch between reported and predicted sex, or (f) the 65 SNP probes on the array show a modest level of correlation (using a cut-off of 0.65) between two samples (whereby the sample with the higher Braak score was retained). Sample and probe exclusion was performed using the *pfilter* function within the wateRmelon package[37], with the following criteria used for exclusion: samples with a detection P > 0.05 in more than 5% of probes, probes with < three beadcount in 5% of samples and probes having 1% of samples with a detection P value > 0.05. Finally, probes with common (minor allele frequency > 5%) SNPs in the single base extension position or probes that are nonspecific or mis-mapped were excluded[38,39], leaving 403,763 probes for analysis. Samples numbers after quality control are those shown in **Table 1**.

Quantile normalization was applied using the *dasen* function in the wateRmelon package[37]. For the discovery cohorts, DNA methylation data was corrected by regressing out the effects of age and sex in all samples in each cohort and tissue separately, with neuron/glia proportions included as an additional covariate in cortical regions. The neuron/glia proportions were calculated using the CETS package[40], and were not included as a co-variate for the cerebellum as the neuronal nuclear protein (NeuN) that was used to generate the neuron/glia algorithm is not expressed by some cerebellar neurons[41]. These three variables (age, sex, neuron/glia proportions) were regressed out of the data as we found that they strongly correlated with either of the first two principal components of the DNA methylation data in most of the datasets. Other potential sources of technical and biological variation (post-mortem interval, ancestry, plate, chip, study and bisulfite treatment batch) did not correlate as strongly with methylation in most datasets. We opted to use surrogate variables as a consistent method to control for variation derived from these measured and other unknown variables across all datasets. Surrogate variables were calculated using the *sva* function in the SVA package[42]. Linear regression analyses were then performed with respect to Braak stage (modelled as a continuous variable) using residuals and a variable number of surrogate variables for each study until the inflation index (lambda) fell below 1.2 (see **Supplementary Table 17**). The surrogate variables included for each cohort correlated with the technical and biological variables that we had not regressed out earlier, demonstrating that this method appropriately controlled for variation not driven by Braak stage. Quantile-quantile plots for the four intra-tissue and the cross-cortex meta-analyses are shown in **Supplementary Figure 20**. Although it appears from these plots that there is P value inflation, it is worth noting that (a) lambda for all meta-analyses < 1.2 and (b) P value inflation is commonly observed in many DNA methylation studies and standard methods to control for this in GWAS are not suitable for EWAS data[43].

Intra-tissue meta-analysis

We used the estimated coefficients and SEs from the six "discovery" cohorts to undertake an inverse variance intra-tissue meta-analysis independently in each available tissue using the *metagen* function within the Meta package[44], which applies inverse variance weighting. The estimates and SEs from individual cohort Braak linear regression analyses were added to the model for each tissue. The prefrontal cortex analyses included three cohorts (N = 959: London 1, Mount Sinai, ROSMAP), the temporal gyrus analyses included four cohorts (N = 608: London 1, Mount Sinai, Arizona 1, Arizona 2) and the entorhinal cortex analyses included two cohorts (N = 189: London 1, London 2). The cerebellum analyses included data from four cohorts (N = 533: London 1, London 2, Arizona 1 and Arizona 2) although the cerebellum data for the Arizona 1 and 2 cohorts was generated in the same experiment, and so these were combined together as a single dataset. The ESs and corresponding SEs reported in this study correspond to the corrected DNA methylation (beta) difference between Braak 0 and Braak VI individuals. Bonferroni significance was defined as $P < 1.238 \times 10^{-7}$ to account for 403,763 tests. A fixed effects meta-analysis are the results primarily reported as it is the most appropriate model for our study as it can more reliably estimate the pooled effect and therefore the standard error and P value. However, in the supplementary tables we do also report the results of the random effects meta-analysis as ~10% of Bonferroni significant DMPs in the intra-tissue meta-analysis had high heterogeneity and in which case the results from the random-effects model should also be considered.

Cross-cortex meta-analysis

As multiple cortical brain regions were available for the "London 1" and "Mount Sinai" cohorts, a mixed model was performed using the *lme* function within the nlme package[45]. Estimate coefficients and SEs from each EWAS were extracted and were subjected to bacon[43] to control for bias and inflation, after which a fixed-effect inverse variance meta-analysis was performed across all discovery cohorts using the *metagen* function. A fixed effects model was selected in this instance for consistency with the intra-tissue meta-analysis, although the random effects meta-analysis results also shown in **Supplementary Table 7**.

Replication analyses

For the Munich replication cohort, we extracted the beta values for the 220 cross-cortex Bonferroni significant DMPs. This DNA methylation data was then corrected for age, sex and neuron/glia proportions (bulk tissue only) prior to performing a linear regression analysis with respect to Braak stage. For the BDR replication cohort, we were provided with beta values for the 208 cross-cortex Bonferroni significant DMPs that were present on the EPIC array. This data had been corrected for age, sex, neuron/glia proportions, batch and principal component 1, before the linear regression analysis was performed with respect to Braak stage, with Bacon used to control for inflation. Additional information on the BDR dataset can be found in Shireby *et al*[15].

Annotations, pathway and regional analyses

Probes were annotated for tables using both the Illumina (UCSC) gene annotation (which is derived from the genomic overlap of probes with RefSeq genes or up to 1500bp from the transcription start site of a gene) and "Genomic Regions Enrichment of Annotations Tool" (GREAT)[46] annotation (which annotates a DMP to genes with a transcription start site within 5kb upstream, or 1kb downstream). Pathway analyses were performed on the Illumina (UCSC) annotated genes corresponding to the 220 Bonferroni significant cross-cortex DMPs (N = 121 genes). We used the 'gometh' function within the missMethyl package (version 1.20.0)[47], which performs one-sided hypergeometric tests and adjusts the test for the uneven number of probes per gene and pathway redundancy. The identified GO terms were subjected to the online tool REViGO (available at http://revigo.irb.hr/)[48], to reduce the number of redundant functional terms based on semantic similarity between ontology terms. Resnik's measure was used to compute the similarity of terms and a medium between terms similarity of 0.7 was allowed. As methylation at neighboring CpG sites can be highly correlated we used a method developed to identify SNPs in LD to identify independent signals[14]. For the 220 Bonferroni significant cross-cortex DMPs we used a threshold of $r < 0.6$ over 1mb to identify 165 independent (non-highly correlated) methylation signals. To identify DMRs consisting of multiple DMPs we used comb-p[49] with a distance of 500bp and a seeded P value of 1.0×10^{-4}. Comb-p was selected for DMR identification over alternative methods as it uses P values as input and so was the most suitable method for calling DMRs in the cross-cortex meta-analysis where multiple brain regions were available for some of the individuals.

Genomic enrichment analyses

To test for an enrichment of DMPs in specific genomic features (*i.e.* CpG islands, shelves, shores, non-CpG island regions) in certain genomic regions (*i.e.* intergenic, distal promoter, proximal promoter, gene body, downstream) we annotated all DMPs with Slieker annotation[16] and performed a two-sided Fisher's exact test comparing to all probes analyzed (N = 403,763). We also used a Fisher's exact test to test for an enrichment of DMPs in genomic regions related to transcription based on the Illumina annotation (TSS1500, TSS200, 5' UTR, 1st exon, gene body, 3' UTR). To investigate whether any of the 220 Bonferroni significant cross-cortex DMPs were driven by genetic variation we used the xQTL resource to identify which of these DMPs are established *cis*-mQTLs[17]. To explore whether Braak-associated methylation was enriched in known AD GWAS variants we used Brown's method to combine together P values from our meta-analyses for probes residing in the LD blocks around the genome-wide significant ($P < 5.0 \times 10^{-8}$) GWAS variants identified by the stage one meta-analysis of Kunkle *et al.*[18] Of the 24 LD blocks reported by Kunkle and colleagues, 20 contained > 1 CpG site on the 450K array and the P values for each CpG in a given block were combined using Brown's method, which accounts for the correlation structure between probes, with the regional P values adjusted to correct for multiple testing.

Quantifying variance in Braak pathology explained by DNA methylation signatures

For this analysis control samples (Braak low [0-II]: N = 407) and AD cases (Braak high [V-VI]: N = 589) from the cross cortex discovery dataset were used for training a classifier. A penalized regression model was used to select the optimum (N = 110) CpG probes from the 220 cross-cortex Bonferroni significant DMPs that determined case-control status in the training dataset using the R package GLMnet[50]. Elastic net uses a combination of ridge and lasso regression, in which alpha (α) = 0 corresponds to ridge, whilst α = 1 corresponds to lasso, the elastic net α parameter used was 0.5. The lambda value was derived when using 10-fold cross validation on the training dataset. The model was then tested for AUC ROC value, confidence intervals (CI) and variance explained in the testing dataset as well as the independent replication Munich (Braak 0-II: N = 9, Braak V-VI: N = 29) and BDR (Braak 0-II: N = 196, Braak V-VI: N = 258) prefrontal cortex datasets.

Acknowledgements

This work was funded by a major project grant from the Alzheimer's Society UK (AS-PG-14-038) to KL, an Alzheimer's Association US New Investigator Research Grant (NIRG-14-320878) to KL and a project grant from the Medical Research Council (MRC) (MR/N027973/1) to KL as part of a larger collaborative project funded to KL and DLAvdH for the EPI-AD consortium through the Joint Programme—Neurodegenerative Disease Research (JPND) initiative. Data analysis was undertaken using high-performance computing supported by a Medical Research Council (MRC) Clinical Infrastructure Award (M008924) to JM. The project was also supported through PhD studentships from the Alzheimer's Society (GS), BRACE (Bristol Research into Alzheimer's and Care of the Elderly) (GW) and the MRC GW4 Doctoral Training Program (DTP) (JAYR). BDR is jointly funded by Alzheimer's Research UK and the Alzheimer's Society in association with the MRC. DNA methylation data generated in the Brains for Dementia Research cohort was supported by the Alzheimer's Society and Alzheimer's Research UK (ARUK). The project was also supported by a number of NIH grants, including P30AG19610, P30AG10161, R01AG15819, R01AG17917, R01AG36042, R01AG036039, R01AG036400 and R01AG067015. We thank the donors and families who have made this research possible.

Contributions

ARS and GW conducted laboratory experiments. RGS, EP, GS, EH, WV and MW undertook data analysis, bioinformatics and/or support with data review. RGS, EP, ARS, JAYR, DM, GG, MR, AG, AJS, LS, VH, DLAvdH, DB, PTF, AJT, SL, KM, JW, PDC, DAB, PLDJ, JM and KL provided data for the meta-analysis. LS developed the online database. KL conceived of the idea and directed the project. KL, RGS and EP drafted the manuscript. All authors read and approved the final submission.

Competing interests

The authors declare no competing interests.

Data availability

The data supporting the findings of this study are available within the article, Supplementary Information or from the authors upon request. Some of the datasets are also available on GEO including London 1 data (GSE59685), London 2 data (GSE105109), Mount Sinai data (GSE80970), Arizona 1 data (GSE134379), Arizona 2 data (GSE109627) and Munich data (GSE66351). We have developed an online database, which can present summary statistics, which is available from our website: www.epigenomicslab.com/ad-meta-analysis/. All scripts for data analyses performed in this manuscript can be found at: https://github.com/rgs212/Meta-analysis-Smith.

REFERENCES

1. Collaborators, G.B.D.D. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol* **18**, 88-106 (2019).
2. Blennow, K., de Leon, M.J. & Zetterberg, H. Alzheimer's disease. *Lancet* **368**, 387-403 (2006).
3. Sperling, R.A. *et al.* Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 280-92 (2011).
4. Jack, C.R., Jr. *et al.* Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol* **9**, 119-28 (2010).
5. Lunnon, K. *et al.* Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. *Nat Neurosci* **17**, 1164-70 (2014).
6. De Jager, P.L. *et al.* Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat Neurosci* **17**, 1156-63 (2014).
7. Smith, R.G. *et al.* Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology. *Alzheimers Dement* **14**, 1580-1588 (2018).

8. Smith, A.R. *et al.* Parallel profiling of DNA methylation and hydroxymethylation highlights neuropathology-associated epigenetic variation in Alzheimer's disease. *Clin Epigenetics* **11**, 52 (2019).
9. Watson, C.T. *et al.* Genome-wide DNA methylation profiling in the superior temporal gyrus reveals epigenetic signatures associated with Alzheimer's disease. *Genome Med* **8**, 5 (2016).
10. Lardenoije, R. *et al.* Alzheimer's disease-associated (hydroxy)methylomic changes in the brain and blood. *Clin Epigenetics* **11**, 164 (2019).
11. Gasparoni, G. *et al.* DNA methylation analysis on purified neurons and glia dissects age and Alzheimer's disease-specific changes in the human cortex. *Epigenetics Chromatin* **11**, 41 (2018).
12. Brokaw, D.L. *et al.* Cell Death and Survival Pathways in Alzheimer's Disease: An Integrative Hypothesis Testing Approach Utilizing -Omic Datasets. *Neurobiol Aging* **In Press** **https://doi.org/10.1016/j.neurobiolaging.2020.06.022**(2020).
13. Smith, A.R. *et al.* A cross-brain regions study of ANK1 DNA methylation in different neurodegenerative diseases. *Neurobiol Aging* **74**, 70-76 (2019).
14. Nyholt, D.R. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* **74**, 765-9 (2004).
15. Shireby, G.L. *et al.* Recalibrating the Epigenetic Clock: Implications for Assessing Biological Age in the Human Cortex *BioxRiv* **https://doi.org/10.1101/2020.04.27.063719**(2020).
16. Slieker, R.C. *et al.* Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics Chromatin* **6**, 26 (2013).
17. Ng, B. *et al.* An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat Neurosci* **20**, 1418-1426 (2017).
18. Kunkle, B.W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat Genet* **51**, 414-430 (2019).
19. Selkoe, D.J. The molecular pathology of Alzheimer's disease. *Neuron* **6**, 487-98 (1991).
20. Xu, J. *et al.* Regional protein expression in human Alzheimer's brain correlates with disease severity. *Commun Biol* **2**, 43 (2019).
21. Labadorf, A. *et al.* RNA Sequence Analysis of Human Huntington Disease Brain Reveals an Extensive Increase in Inflammatory and Developmental Gene Expression. *PLoS One* **10**, e0143563 (2015).
22. Roubroeks, J.A.Y. *et al.* An epigenome-wide association study of Alzheimer's disease blood highlights robust DNA hypermethylation in the HOXB6 gene. . *Neurobiol Aging* **https://doi.org/10.1016/j.neurobiolaging.2020.06.023**(2020).
23. Friedrich, J. *et al.* Hox Function Is Required for the Development and Maintenance of the Drosophila Feeding Motor Unit. *Cell Rep* **14**, 850-60 (2016).
24. Sun, L.L., Yang, S.L., Sun, H., Li, W.D. & Duan, S.R. Molecular differences in Alzheimer's disease between male and female patients determined by integrative network analysis. *J Cell Mol Med* **23**, 47-58 (2019).
25. Rosenberger, A.F. *et al.* Protein Kinase Activity Decreases with Higher Braak Stages of Alzheimer's Disease Pathology. *J Alzheimers Dis* **49**, 927-43 (2016).

26. Satoh, J., Tabunoki, H., Ishida, T., Saito, Y. & Arima, K. Accumulation of a repulsive axonal guidance molecule RGMa in amyloid plaques: a possible hallmark of regenerative failure in Alzheimer's disease brains. *Neuropathol Appl Neurobiol* **39**, 109-20 (2013).
27. Fernandez, A.F. *et al.* A DNA methylation fingerprint of 1628 human samples. *Genome Res* **22**, 407-19 (2012).
28. Masliah, E., Dumaop, W., Galasko, D. & Desplats, P. Distinctive patterns of DNA methylation associated with Parkinson disease: identification of concordant epigenetic changes in brain and peripheral blood leukocytes. *Epigenetics* **8**, 1030-8 (2013).
29. Sanchez-Mut, J.V. *et al.* Human DNA methylomes of neurodegenerative diseases show common epigenomic patterns. *Transl Psychiatry* **6**, e718 (2016).
30. MacBean, L.F., Smith, A.R. & Lunnon, K. Exploring beyond the DNA sequence: A Review of Epigenomic Studies of DNA and Histone Modifications in Dementia. *Current Genetic Medicine Reports* **In Press**(2020).
31. Relton, C.L. & Davey Smith, G. Mendelian randomization: applications and limitations in epigenetic studies. *Epigenomics* **7**, 1239-43 (2015).
32. Beach, T.G. *et al.* Arizona Study of Aging and Neurodegenerative Disorders and Brain and Body Donation Program. *Neuropathology* **35**, 354-89 (2015).
33. R Development Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria 2012* (2012).
34. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).
35. Spiers, H. *et al.* Methylomic trajectories across human fetal brain development. *Genome Res* **25**, 338-52 (2015).
36. Aryee, M.J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-9 (2014).
37. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).
38. Chen, Y.A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203-9 (2013).
39. Price, M.E. *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* **6**, 4 (2013).
40. Guintivano, J., Aryee, M.J. & Kaminsky, Z.A. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* **8**, 290-302 (2013).
41. Gusel'nikova, V.V. & Korzhevskiy, D.E. NeuN As a Neuronal Nuclear Antigen and Neuron Differentiation Marker. *Acta Naturae* **7**, 42-7 (2015).
42. Chakraborty, S., Datta, S. & Datta, S. Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies. *Bioinformatics* **28**, 799-806 (2012).
43. van Iterson, M., van Zwet, E.W., Consortium, B. & Heijmans, B.T. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol* **18**, 19 (2017).
44. Schwarzer, G. meta: A R package for meta-analysis. *R News* **7**, 40-45 (2007).

45. Pinheiro, J., Bates, D., DebRoy, S. & Sarkar, D. nlme: Linear and Nonlinear Mixed Effects Models. (2019).

46. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).

47. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32**, 286-8 (2016).

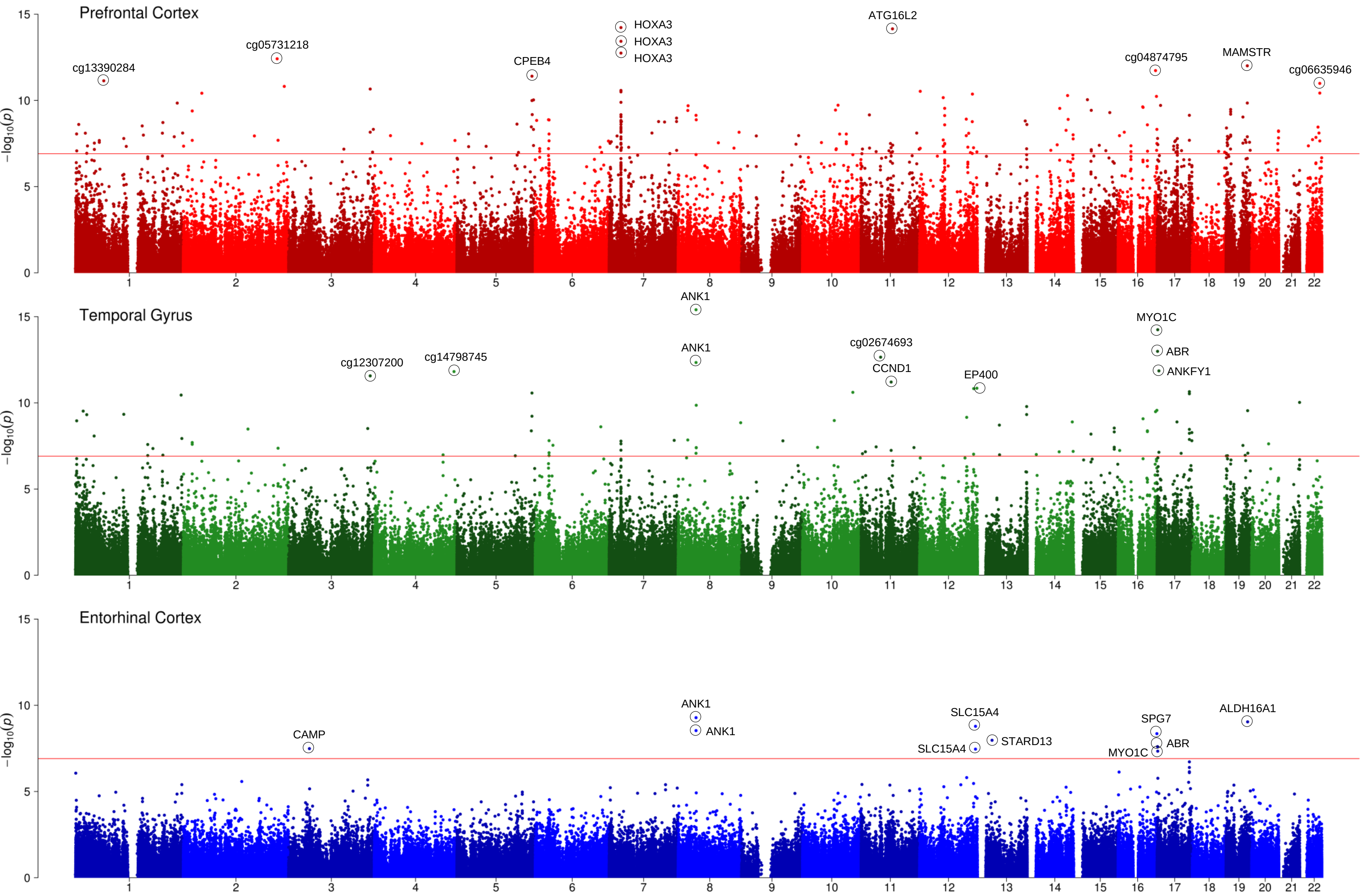
48. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).

49. Pedersen, B.S., Schwartz, D.A., Yang, I.V. & Kechris, K.J. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* **28**, 2986-8 (2012).

50. Friedman, J. *et al.* glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. (2019).

Figure 1: Intra-tissue meta-analyses of AD methylomic studies highlights Bonferroni significant differentially methylated positions (DMPs) in all cortical tissues. (a) A Manhattan plot for the prefrontal cortex (red), temporal gyrus (green) and entorhinal cortex (blue) meta-analyses, with the ten most significant DMPs circled on the plot and Illumina UCSC gene name shown if annotated, or CpG ID if unannotated. The X-axis shows chromosomes 1-22 and the Y-axis shows -log10(p), with the horizontal red line denoting Bonferroni significance ($P < 1.238 \times 10^{-7}$). (**b**) A Venn diagram highlighting overlapping DMPs at Bonferroni significance across the cortical tissues. (**c**) In each cortical brain region the Bonferroni significant DMPs identified in that region usually had a greater effect size (ES) there, than in any of the other cortical regions. The X-axis represents the methylation (beta) ES between individuals that are Braak stage 0 and VI. Data is separated on the Y-axis by tissue analysis (large text) with the corresponding data at these probes in other tissues (small text). The white dot in the centre represents the median, the dark box represents the interquartile range (IQR), whilst the whisker lines represent the "minimum" (quartile 1 – 1.5 x IQR) and the "maximum" (quartile 3 + 1.5 x IQR). The coloured violin represents all samples including outliers, meaning that the "minimum" and "maximum" may not extend to the end of the violin.

Fig 1. a



Prefrontal Cortex
15
10
5
0
−log10(p)
cg13390284
cg05731218
CPEB4
HOXA3
HOXA3
HOXA3
ATG16L2
cg04874795
MAMSTR
cg06635946
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Temporal Gyrus
ANK1
ANK1
MYO1C
ABR
ANKFY1
cg12307200
cg14798745
cg02674693
CCND1
EP400
Entorhinal Cortex
CAMP
ANK1
ANK1
SLC15A4
SLC15A4
STARD13
SPG7
ABR
MYO1C
ALDH16A1

Fig 1. b

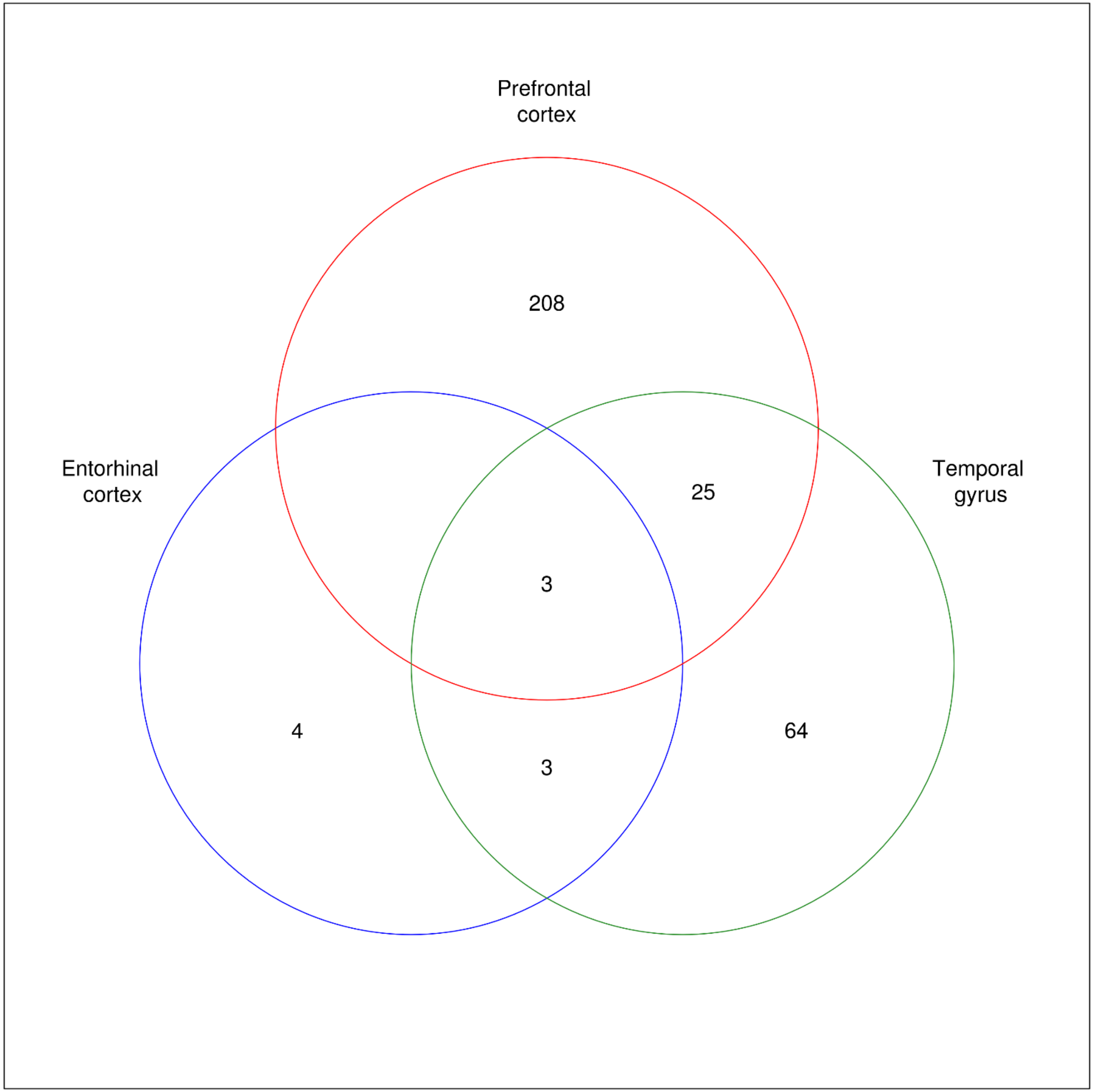
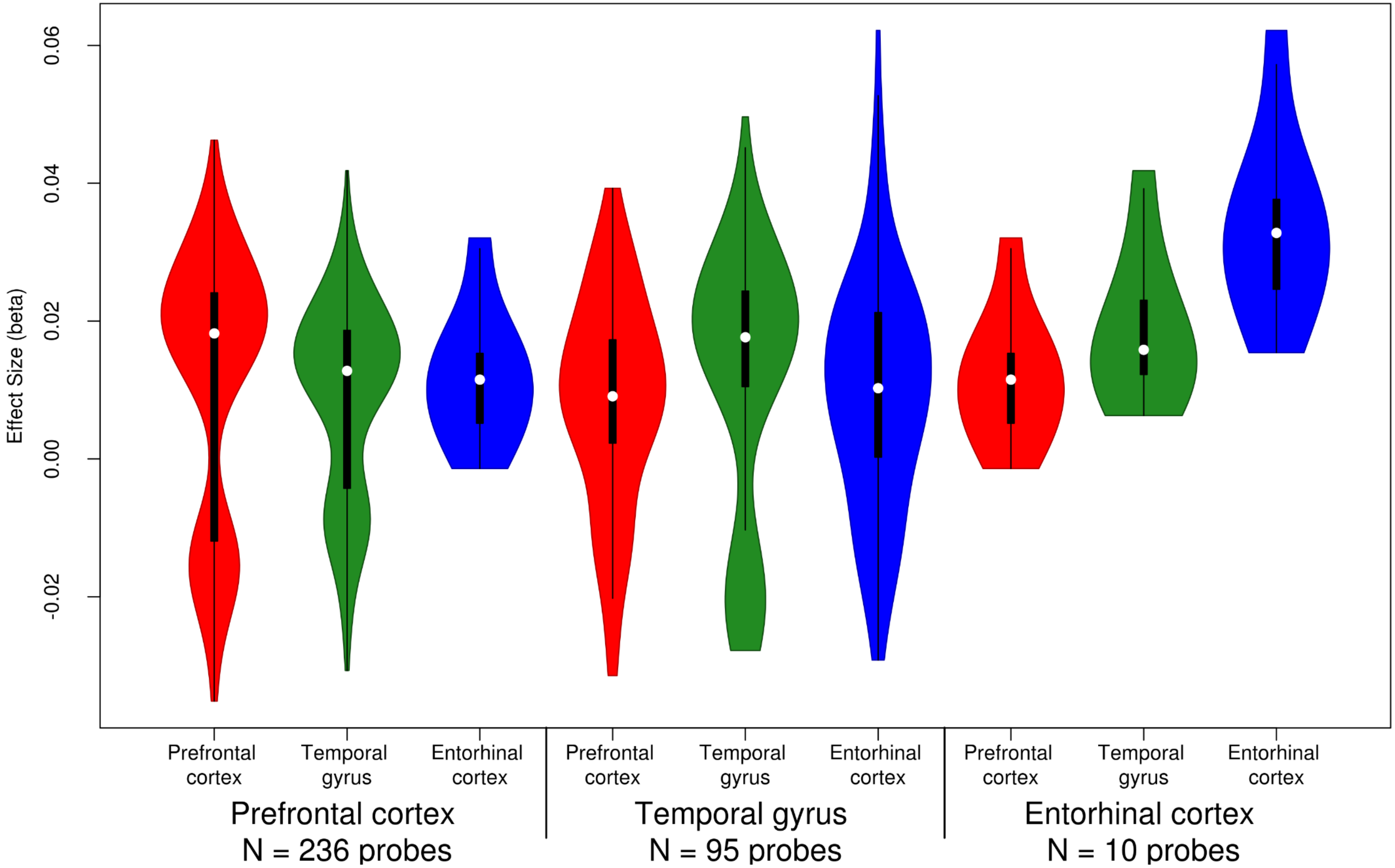

Prefrontal cortex
Entorhinal cortex
Temporal gyrus
208
25
3
4
3
64

Fig 1. c



0.06
0.04
0.02
0.00
-0.02
Effect Size (beta)
Prefrontal cortex
Temporal gyrus
Entorhinal cortex
Prefrontal cortex
Temporal gyrus
Entorhinal cortex
Prefrontal cortex
Temporal gyrus
Entorhinal cortex
Prefrontal cortex
N = 236 probes
Temporal gyrus
N = 95 probes
Entorhinal cortex
N = 10 probes

Figure 2: A cross-cortex meta-analysis identifies 220 Bonferroni significant differentially methylated positions (DMPs) associated with Braak stage. (a) A Miami plot of the cross-cortex meta-analyses. Probes shown above the X-axis indicate hypermethylation with higher Braak stage, whilst probes shown below the X-axis indicate hypomethylation with higher Braak stage. The chromosome and genomic position are shown on the X-axis. The Y-axis shows –log10(p). The red horizontal lines indicate the Bonferroni significance level of $P < 1.238 \times 10^{-7}$. Probes with a methylation (beta) effect size (ES: difference between Braak 0- Braak VI) ≥ 0.01 and $P < 1.238 \times 10^{-7}$ are shown in blue. The 20 most significant DMPs are circled on the plot and Illumina UCSC gene name is shown if annotated, or CpG ID if unannotated. (**b**) A volcano plot showing the ES (X-axis) and –log10(p) (Y-axis) for the cross-cortical meta-analysis results. Gray probes indicate an ES between ≥ 0.01, whilst blue probes indicate an ES ≥ 0.01 and $P < 1.238 \times 10^{-7}$. (**c**) The most significant cross-cortex differentially methylated region (DMR) (chr7:27153212-27154305) contained 11 probes and resided in the *HOXA* region. The horizontal red line denotes the Bonferroni significance level of $P < 1.238 \times 10^{-7}$. Red probes represent a positive ES ≥ 0.01, blue probes represent a negative ES ≥ 0.01. Underneath the gene tracks are shown in black with CpG islands in green.

Fig 2. a

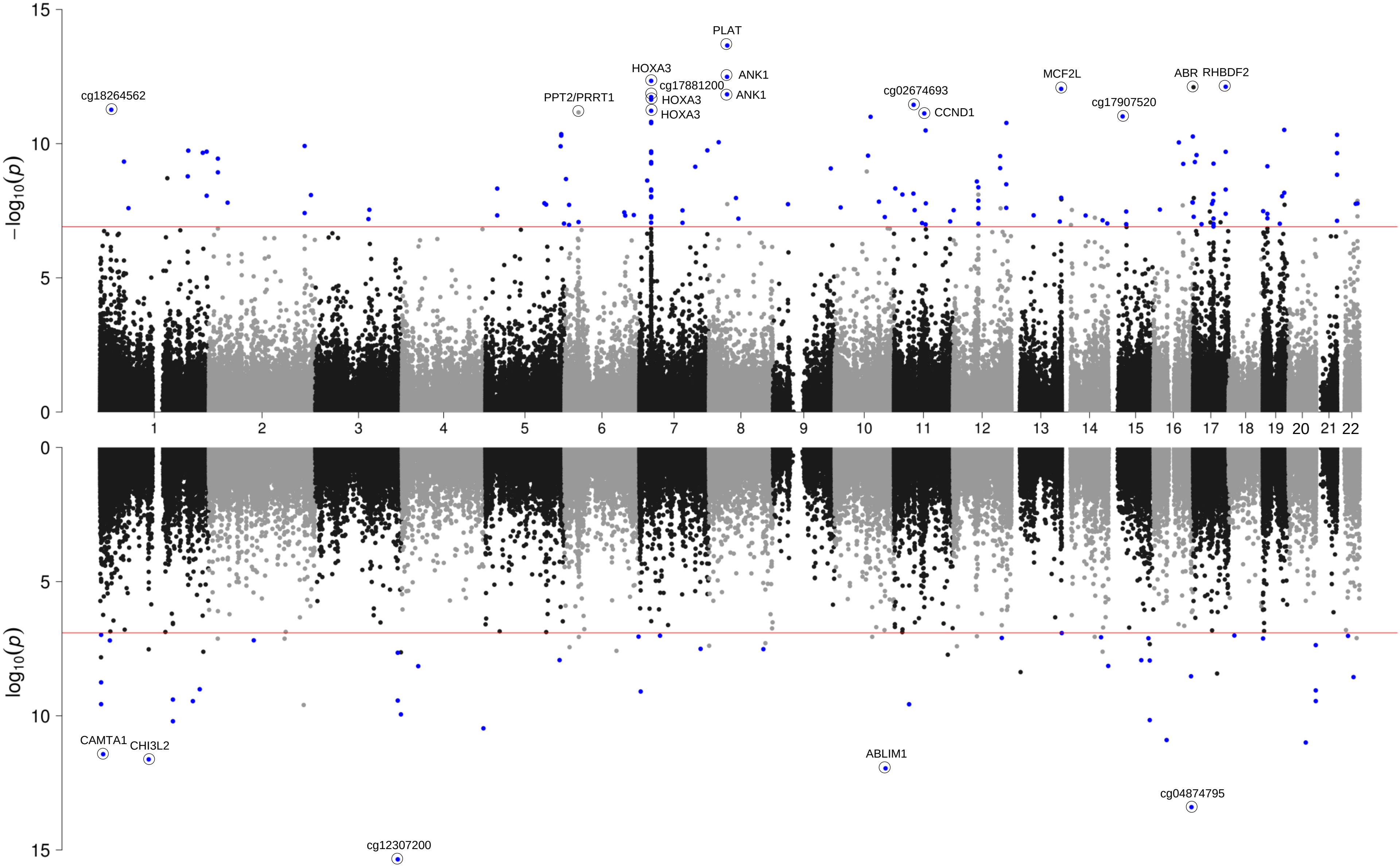

15
10
5
0
$-\log_{10}(p)$
PLAT
HOXA3
ANK1
cg17881200
HOXA3
ANK1
HOXA3
cg18264562
PPT2/PRRT1
cg02674693
CCND1
MCF2L
cg17907520
ABR
RHBDF2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
0
5
10
15
$\log_{10}(p)$
CAMTA1
CHI3L2
ABLIM1
cg04874795
cg12307200

Fig 2. b

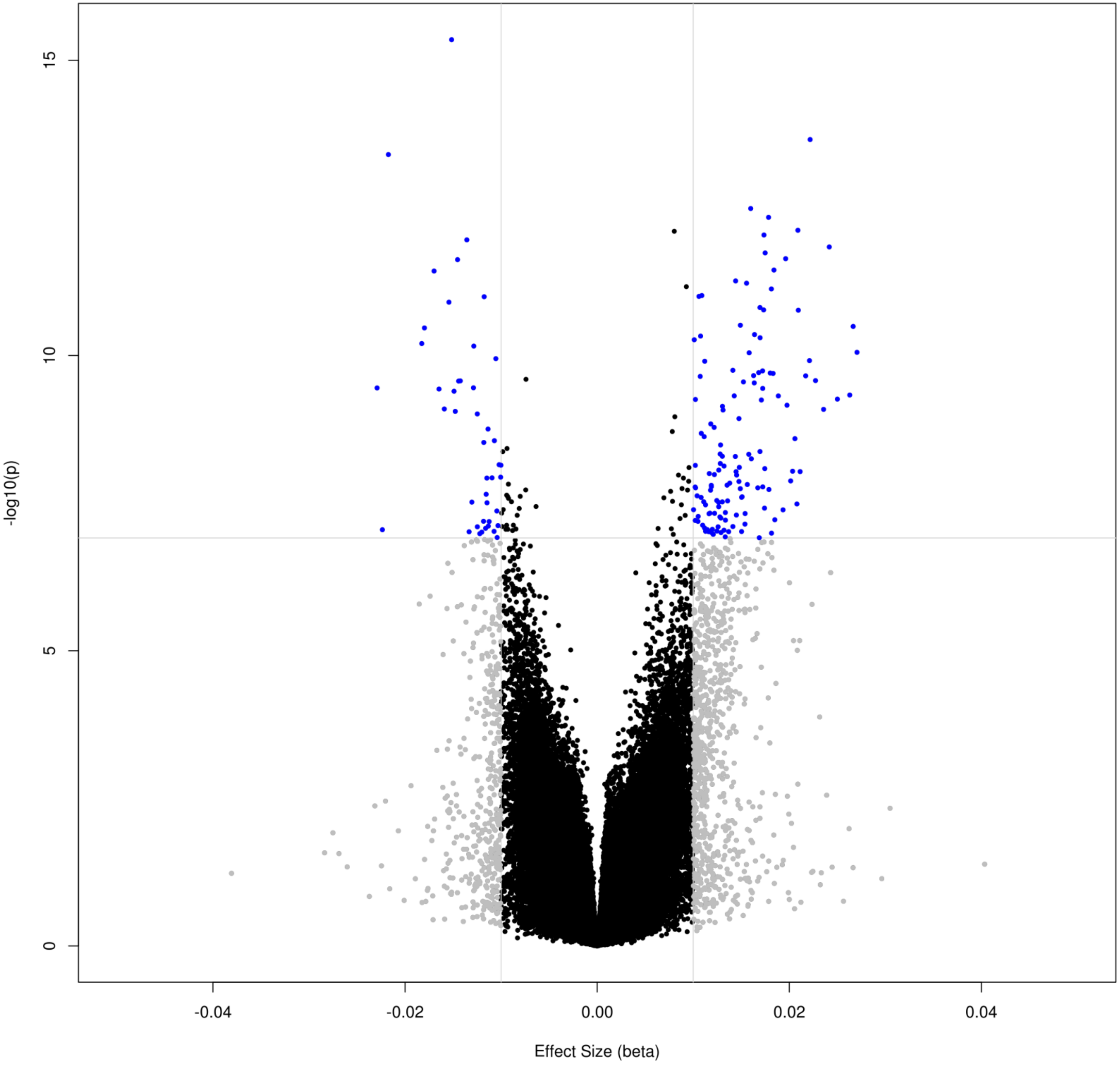


Fig 2. c

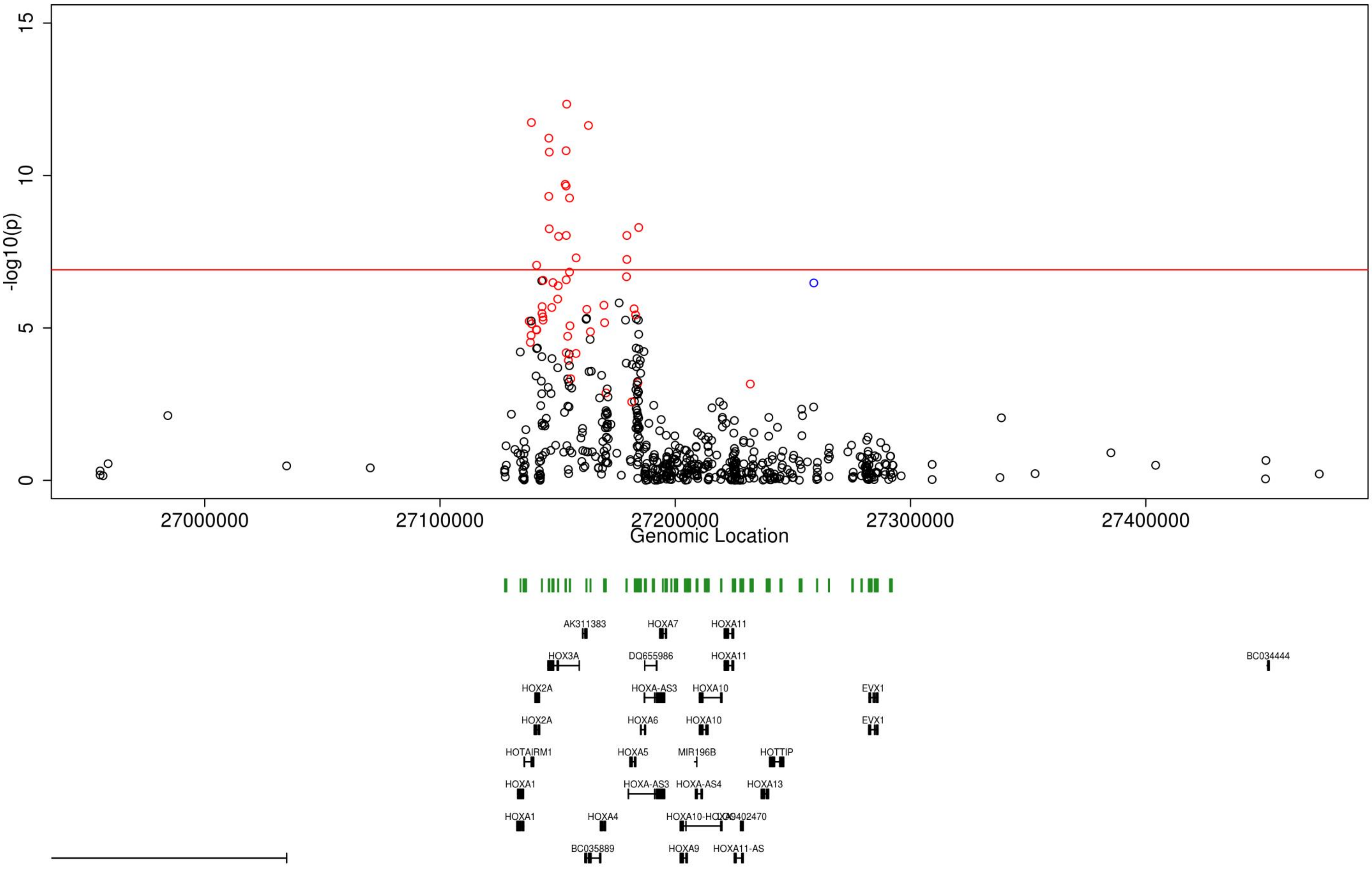

15
10
5
0
-log10(p)
27000000
27100000
27200000
27300000
27400000
Genomic Location
AK311383
HOXA7
HOXA11
HOX3A
DQ655986
HOXA11
BC034444
HOX2A
HOXA-AS3
HOXA10
EVX1
HOX2A
HOXA6
HOXA10
EVX1
HOTAIRM1
HOXA5
MIR196B
HOTTIP
HOXA1
HOXA-AS3
HOXA-AS4
HOXA13
HOXA1
HOXA4
HOXA10-HOXA9
LOC402470
BC035889
HOXA9
HOXA11-AS

Figure 3: Independent replication of the Bonferroni significant cross-cortex differentially methylated loci. (**a**) The methylation (beta) effect size (ES) of the 220 cross-cortex differentially methylated positions (DMPs) identified in the discovery cohorts (X-axis) were significantly correlated with the ES in the Munich replication cohort in the prefrontal cortex (red, r = 0.64, P = 5.24 x 10^{-27}), sorted neuronal cells (light blue, r = 0.45, P = 1.56 x 10^{-12}) and non-neuronal cells (purple, r = 0.79, P = 1.43 x 10^{-47}) (Y-axis). (**b**) A forest plot of the most significant cross-cortex DMP (cg12307200, chr3:188664632, P = 4.48 x 10^{-16}). The effect size is shown in the prefrontal cortex (red), temporal gyrus (green) and entorhinal cortex (blue) for the different discovery cohorts. The X-axis shows the beta ES, with dots representing ES and arms indicating standard error (SE). ES from the intra-tissue meta-analysis using all available individual cohorts are represented by polygons in the corresponding tissue color. The black polygon represents the cross-cortex data. Shown in purple on the plot is the ES in the Munich replication cohort in the prefrontal cortex and sorted neuronal cells and non-neuronal cells, with the direction of effect suggesting the hypomethylation seen in the discovery cohorts is driven by changes in non-neuronal cells. (**c**) In the BDR replication cohort DNA methylation data was available in the prefrontal cortex for 208 of the 220 Bonferroni significant cross-cortex DMPs. The ES of these 208 cross-cortex DMPs in the discovery cohorts (X-axis) were significantly correlated with the ES in the BDR replication cohort (r = 0.53, P = 4.13 x 10^{-16}) (Y-axis).

Fig 3. a

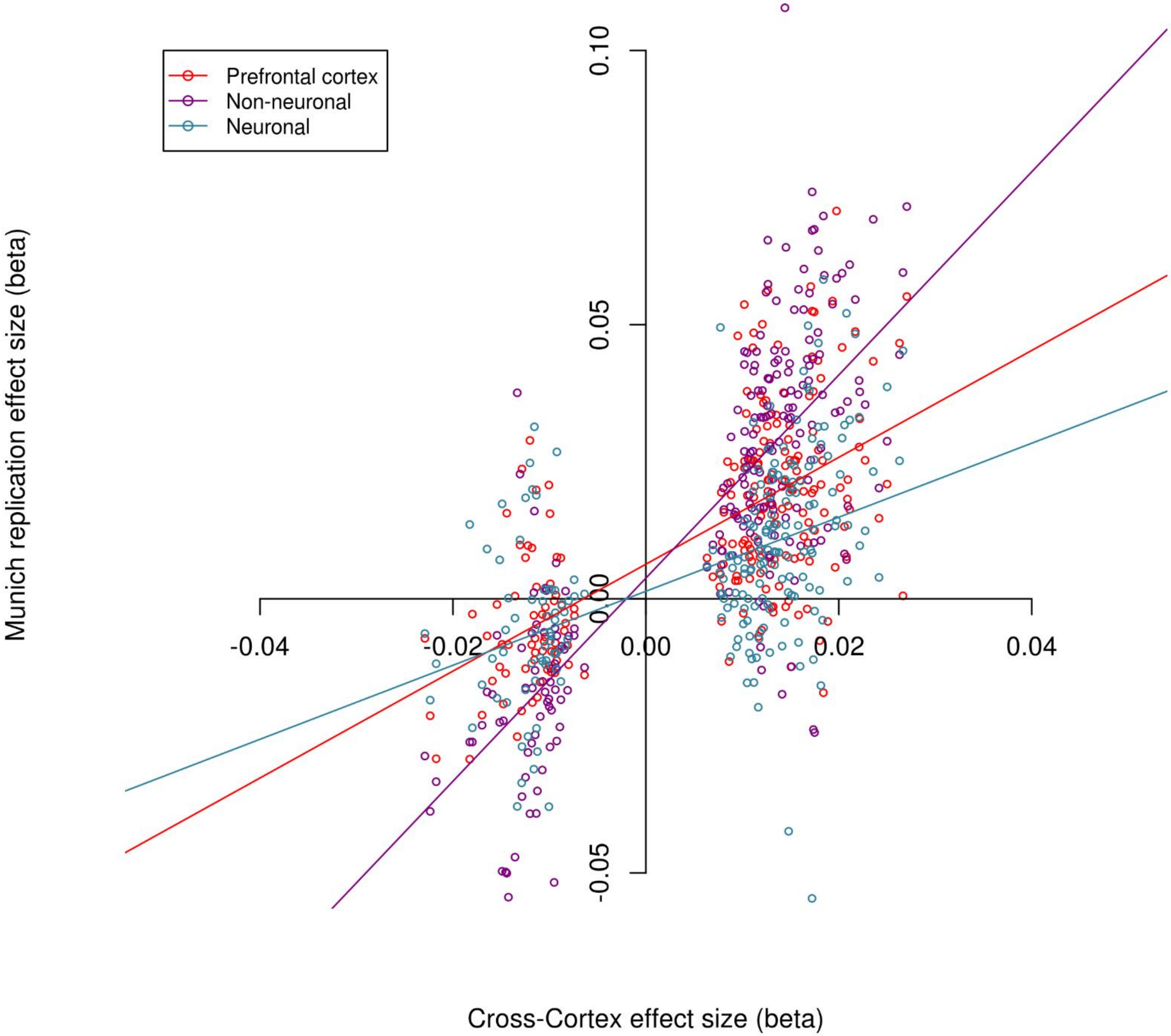

Prefrontal cortex
Non-neuronal
Neuronal
Munich replication effect size (beta)
0.10
0.05
0.00
-0.05
-0.04
-0.02
0.00
0.02
0.04
Cross-Cortex effect size (beta)

Fig 3. b

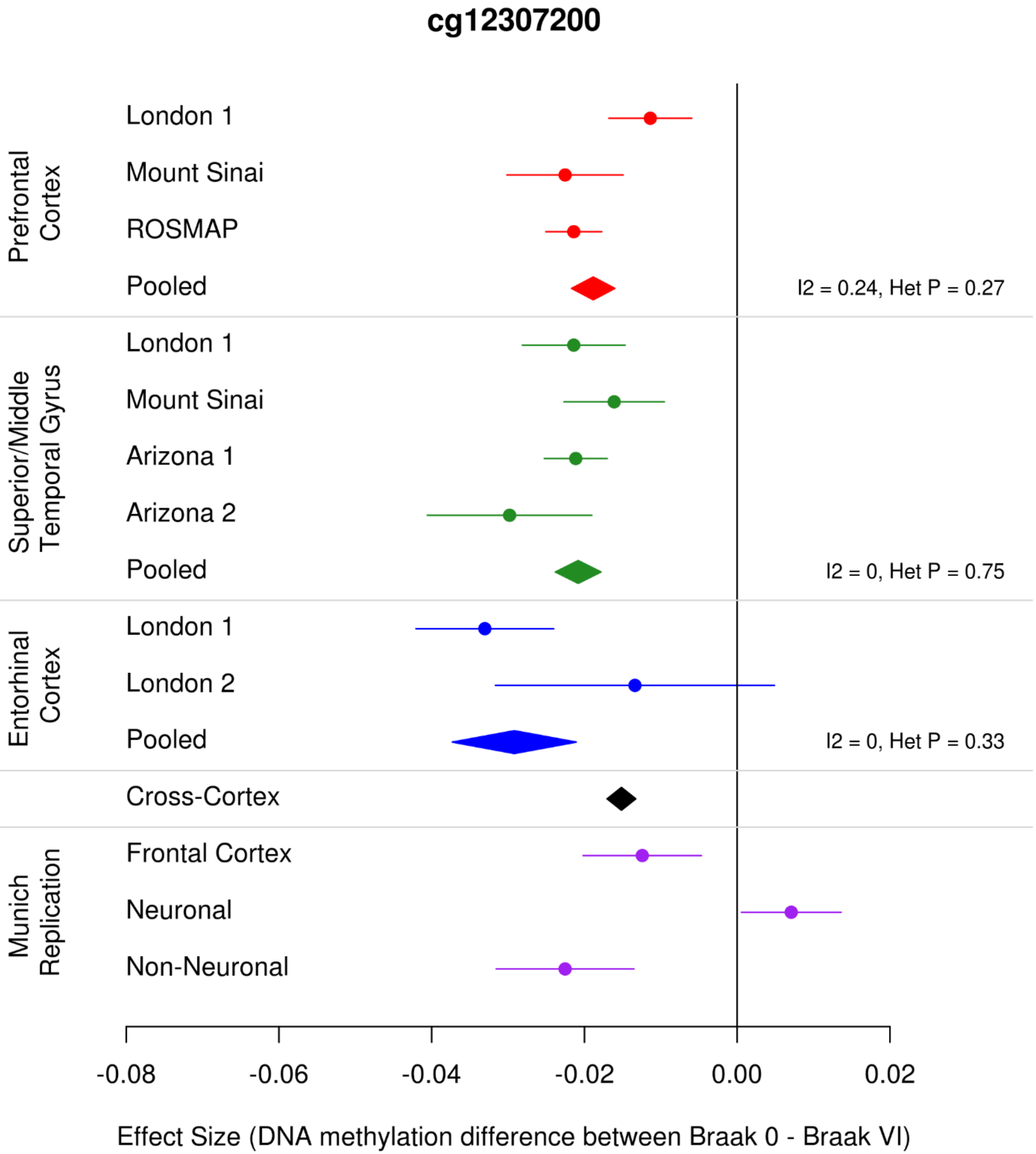

cg12307200
Prefrontal Cortex
London 1
Mount Sinai
ROSMAP
Pooled
I2 = 0.24, Het P = 0.27
Superior/Middle Temporal Gyrus
London 1
Mount Sinai
Arizona 1
Arizona 2
Pooled
I2 = 0, Het P = 0.75
Entorhinal Cortex
London 1
London 2
Pooled
I2 = 0, Het P = 0.33
Cross-Cortex
Munich Replication
Frontal Cortex
Neuronal
Non-Neuronal
-0.08
-0.06
-0.04
-0.02
0.00
0.02
Effect Size (DNA methylation difference between Braak 0 - Braak VI)

Fig 3. c

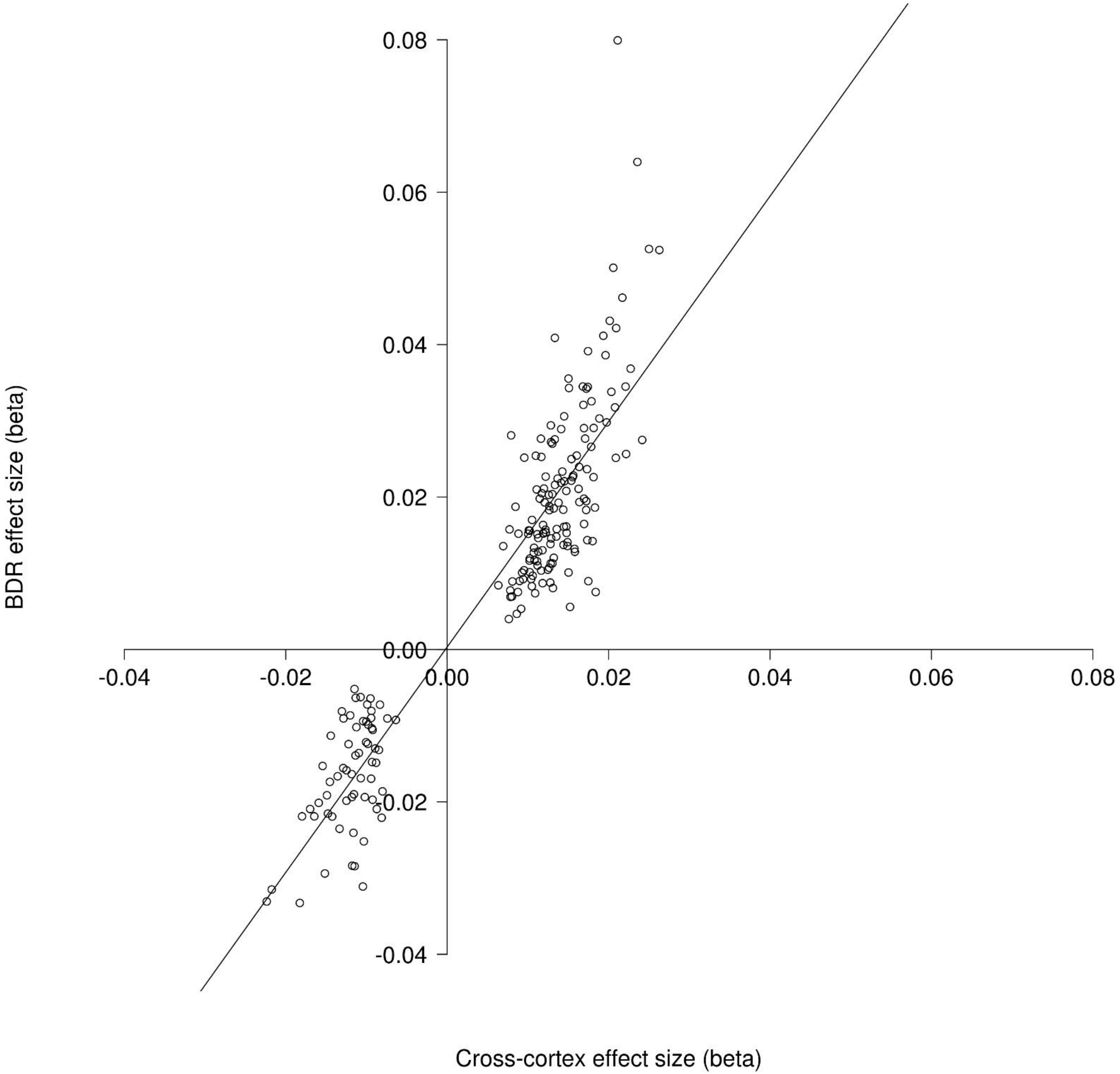

BDR effect size (beta)
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.04
-0.02
0.00
0.02
0.04
0.06
0.08
Cross-cortex effect size (beta)

Figure 4: Receiver Operating Characteristic (ROC) graphs highlighting the Area Under the Curve (AUC) for the 110 cross-cortex probes that can best explain the variance in Braak pathology. An elastic net penalized regression model was used to identify a subset of 110 of the Bonferroni significant cross-cortex probes that could best predict whether a sample has low pathology (Braak 0-II: “control”) compared to high pathology (Braak V-VI: “AD”) in a training dataset comprised of 996 discovery samples (Braak 0-II: N = 407, Braak V-VI: N = 589). This model had an Area Under the ROC Curve (AUC) of 94.33% (confidence interval [CI] = 92.88-95.64%) and explained 71.11% of the pathological variance (black line). The 110 probe signature was then tested in two independent replication cohorts. In the Munich prefrontal cortex samples (Braak 0-II: N = 9, Braak V-VI: N = 29) the model had an AUC of 73.95% (CI = 55.17-88.89%), explaining 20.18% of the variance (blue line). In the BDR prefrontal cortex samples (Braak 0-II: N = 196, Braak V-VI: N = 258) the model had an AUC = 70.36% (CI = 65.52-75.12%), explaining 15.87% of the variance (green line). A list of the 110 probes and their performance characteristics can be found in Supplementary Tables 13 and 14, respectively.

Fig 4.

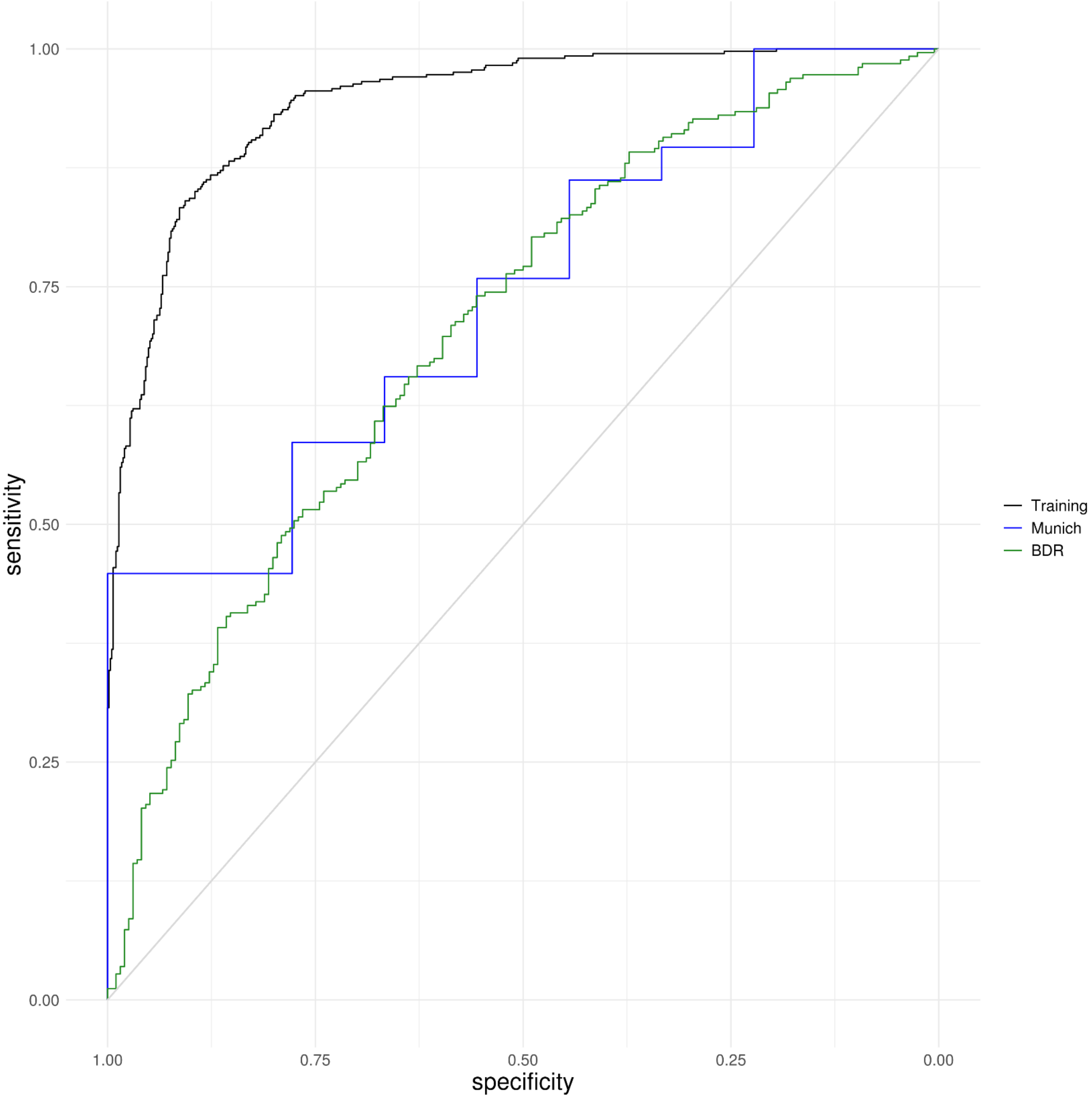

1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
Training
Munich
BDR

Table 1: Demographic information for cohorts included in the meta-analyses. Sample numbers, split of males (M)/females (F) and mean age at death in years (± standard deviation [SD]) are shown for individuals with low pathology (Braak 0-II), mid-stage pathology (Braak III-IV) and severe pathology (Braak V-VI) in each cohort. Shown are the bulk tissues available from each cohort, which included the cerebellum, entorhinal cortex, middle temporal gyrus, prefrontal cortex and superior temporal gyrus. In the discovery meta-analyses, we used data from six EWAS using the 450K array, which all had > 50 unique donors. For replication we used two cohorts. The Munich cohort had 450K data from bulk prefrontal cortex tissue, as well as data available from sorted neuronal and non-neuronal cell populations from the occipital cortex. The BDR cohort had EPIC array data available from bulk prefrontal cortex samples. For the meta-analyses, superior temporal gyrus and middle temporal gyrus samples were both classed as temporal gyrus samples. Shown are final numbers for all cohorts after data quality control. Ancestry is reported for the discovery cohorts and is the number of unique individuals that had the following inferred ethnicities from the 1000 genomes reference panel: European (Eu), African (Af), American (Am), East Asian (As).

Stage	**Cohort**	**Unique individuals**	**Ancestry (Eu/Af/Am/As)**	**Braak**	**Number**	**Sex (M/F)**	**Age at death in (± SD)**	**Tissues analysed**
DISCOVERY	London 1	113	112/0/1/0	0-II	29	13/16	77.6 (12.8)	Prefrontal cortex, entorhinal cortex, superior temporal gyrus, cerebellum (Bulk)
				III-IV	18	7/11	88.5 (5.2)	
				V-VI	66	26/40	85.4 (8.1)	
	London 2	95	92/1/2/0	0-II	23	12/11	76.1 (10.0)	Entorhinal cortex, cerebellum (Bulk)
				III-IV	16	3/13	87.6 (6.4)	
				V-VI	56	26/30	81.5 (8.6)	
	Mount Sinai	146	113/20/11/2	0-II	60	32/28	82 (7.6)	Prefrontal cortex, superior temporal gyrus (Bulk)
				III-IV	42	12/30	88.8 (6.6)	
				V-VI	44	12/32	88.0 (7.5)	
	Arizona 1	302	302/0/0/0	0-II	61	40/21	80.3 (8.2)	Middle temporal gyrus, cerebellum (Bulk)
				III-IV	97	50/47	86.9 (6.9)	
				V-VI	144	63/81	82.3 (8.5)	
	Arizona 2	88	88/0/0/0	0-II	16	10/6	82.5 (5.0)	Middle temporal gyrus, cerebellum (Bulk)
				III-IV	45	21/24	86.7 (5.1)	
				V-VI	27	12/15	84.6 (7.1)	
	ROS/MAP	709	709/0/0/0	0-II	143	70/73	83.2 (6.0)	Prefrontal cortex (Bulk)
				III-IV	409	144/266	86.9 (4.1)	
				V-VI	157	45/113	87.8 (3.5)	
REPLICATION	Munich	45	-	0-II	9	5/4	76.7 (10.9)	Prefrontal cortex (Bulk)
				III-IV	7	1/6	82.1 (5.2)	
				V-VI	29	12/17	79.2 (8.5)	
		26	-	0-II	11	7/4	75.9 (8.5)	Occipital cortex (Sorted cells)
				III-IV	5	1/4	85.0 (6.5)	
				V-VI	10	4/6	77.9 (6.6)	
	BDR	590	-	0-II	196	100/96	83.6 (10.6)	Prefrontal cortex (Bulk)
				III-IV	136	91/65	85.1 (7.45)	
				V-VI	258	128/130	82.5 (8.5)	

Table 2: The 25 most significant differentially methylated positions (DMPs) associated with Braak stage from the cross-cortex meta-analysis. Probe information is provided corresponding to chromosomal location (hg19/GRCh37 genomic annotation), Illumina gene annotation, closest genes with a transcription start site upstream or downstream (from GREAT annotation). Shown for each DMP is the methylation (beta) effect size (ES), standard error (SE) and corresponding unadjusted P value from the inverse variance fixed effects meta-analysis model in the cross-cortex data. All ES and SE have been multiplied by six to demonstrate the difference between Braak stage 0 and Braak stage VI samples. A more comprehensive table is provided in Supplementary Table 7.

Probe	Position	Illumina Gene Annotation	GREAT annotation - closest genes with transcription start site upstream (distance to site)	GREAT annotation - closest genes with transcription start site downstream (distance to site)	ES	SE	P
cg12307200	chr3:188664632		*TPRG1 (-225131)*	*LPP (+733912)*	-0.015	0.002	4.48E-16
cg01419713	chr8:42038135	*PLAT*		*PLAT (+27107), AP3M2 (+27672)*	0.022	0.003	2.20E-14
cg04874795	chr16:86477638		*FOXF1 (-66495)*	*IRF8 (+545230)*	-0.022	0.003	3.95E-14
cg11823178	chr8:41519399	*ANK1;MIR486*	*NKX6-3 (-14522)*	*ANK1 (+234881)*	0.016	0.002	3.24E-13
cg07061298	chr7:27153847	*HOXA3*	*HOXA2 (-11418)*	*HOXA3 (+5367)*	0.018	0.002	4.57E-13
cg13076843	chr17:74475294	*RHBDF2*		*RHBDF2 (+22195), AANAT (+25862)*	0.021	0.003	7.57E-13
cg25018458	chr17:980014	*ABR*		*TIMM22 (+79658), ABR (+103154)*	0.008	0.001	7.87E-13
cg07883124	chr13:113634042	*MCF2L*	*F7 (-126079)*	*MCF2L (+10508)*	0.017	0.002	9.10E-13
cg03223072	chr10:116398913	*ABLIM1*	*AFAP1L2 (-234670)*	*ABLIM1 (+19144)*	-0.014	0.002	1.10E-12
cg05066959	chr8:41519308	*ANK1;MIR486*	*NKX6-3 (-14431)*	*ANK1 (+234972)*	0.024	0.003	1.45E-12
cg17881200	chr7:27138850		*HOXA1 (-3258)*		0.017	0.002	1.83E-12
cg19240213	chr7:27163095	*HOXA3*	*HOXA3 (-3882)*		0.020	0.003	2.29E-12
cg10045881	chr1:111770291	*CHI3L2*	*CHIA (-63247)*	*CHI3L2 (+26899)*	-0.015	0.002	2.38E-12
cg02674693	chr11:45109122		*TP53I11 (-137412), PRDM11 (-59772)*		0.018	0.003	3.57E-12
cg06800235	chr1:7692367	*CAMTA1*	*VAMP3 (-138962)*	*CAMTA1 (+846984)*	-0.017	0.002	3.71E-12
cg18264562	chr1:26253412		*STMN1 (-20456)*	*PAFAH2 (+71236)*	0.014	0.002	5.46E-12
cg01964852	chr7:27146262	*HOXA3*	*HOXA2 (-3833)*		0.016	0.002	5.96E-12
cg01111041	chr6:32121055	*PPT2;PRRT1*	*PRRT1 (-1327), PPT2-EGFL8 (-944), PPT2 (-245)*		0.009	0.001	6.83E-12
cg15974867	chr11:69464012	*CCND1*		*CCND1 (+8158), ORAOV1 (+26103)*	0.018	0.003	7.46E-12
cg17907520	chr15:31680189			*KLF13 (+61132), OTUD7A (+267353)*	0.011	0.002	9.65E-12
cg16988611	chr10:82224946	*TSPAN14*		*TSPAN14 (+11025)*	0.011	0.002	9.98E-12
cg13579486	chr20:39314091			*MAFB (+3789)*	-0.012	0.002	1.01E-11
cg01681367	chr16:29676071	*SPN*	*QPRT (-14287)*	*SPN (+1492)*	-0.015	0.002	1.25E-11
cg01301319	chr7:27153580	*HOXA3*	*HOXA2 (-11151)*	*HOXA3 (+5634)*	0.017	0.003	1.54E-11
cg02317313	chr12:122235206	*LOC338799*	*RHOF (-3039)*		0.017	0.003	1.69E-11

Bibliography

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. doi:10.1038/nature15393

Aguilar, B. J., Zhu, Y., & Lu, Q. (2017). Rho GTPases as therapeutic targets in Alzheimer’s disease. *Alzheimer’s research & therapy*, *9*(1), 97. doi:10.1186/s13195-017-0320-4

Alafuzoff, I., Arzberger, T., Al-Sarraj, S., Bodi, I., Bogdanovic, N., Braak, H., … Kretzschmar, H. (2008). Staging of neurofibrillary pathology in Alzheimer’s disease: a study of the BrainNet Europe Consortium. *Brain Pathology*, *18*(4), 484–496. doi:10.1111/j.1750-3639.2008.00147.x

Altuna, M., Urdánoz-Casado, A., Sánchez-Ruiz de Gordoa, J., Zelaya, M. V., Labarga, A., Lepesant, J. M. J., … Mendioroz, M. (2019). DNA methylation signature of human hippocampus in Alzheimer’s disease is linked to neurogenesis. *Clinical epigenetics*, *11*(1), 91. doi:10.1186/s13148-019-0672-7

Alzheimer’s Association. (2019). 2019 Alzheimer’s disease facts and figures. *Alzheimer’s & Dementia*, *15*(3), 321–387. doi:10.1016/j.jalz.2019.01.010

Alzheimer’s Association Calcium Hypothesis Workgroup. (2017). Calcium Hypothesis of Alzheimer’s disease and brain aging: A framework for integrating new evidence into a comprehensive theory of pathogenesis. *Alzheimer’s & Dementia*, *13*(2), 178–182.e17. doi:10.1016/j.jalz.2016.12.006

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30. doi:10.1186/s13059-020-1935-5

Amlie-Wolf, A., Tang, M., Way, J., Dombroski, B., Jiang, M., Vrettos, N., … Schellenberg, G. D. (2018). Inferring the molecular mechanisms of noncoding Alzheimer’s disease-associated genetic variants. *BioRxiv*. doi:10.1101/401471

Andrés-Benito, P., Gelpi, E., Povedano, M., Ausín, K., Fernández-Irigoyen, J., Santamaría, E., & Ferrer, I. (2019). Combined transcriptomics and proteomics in frontal cortex area 8 in frontotemporal lobar degeneration linked to C9ORF72 expansion. *Journal of Alzheimer’s Disease*, *68*(3), 1287–1307. doi:10.3233/JAD-181123

Antequera, F., & Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(24), 11995–11999. doi:10.1073/pnas.90.24.11995

Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Molecular autism*, *8*, 21. doi:10.1186/s13229-017-0137-9

Axelrud, L. K., Santoro, M. L., Pine, D. S., Talarico, F., Gadelha, A., Manfro, G. G., … Salum, G. A. (2018). Polygenic risk score for alzheimer’s disease: implications for memory performance and hippocampal volumes in early life. *The American Journal of Psychiatry*, *175*(6), 555–563. doi:10.1176/appi.ajp.2017.17050529

Aziz, A.-L., Giusiano, B., Joubert, S., Duprat, L., Didic, M., Gueriot, C., … Ceccaldi, M. (2017). Difference in imaging biomarkers of neurodegeneration between early and late-onset amnestic Alzheimer’s disease. *Neurobiology of Aging*, *54*, 22–30. doi:10.1016/j.neurobiolaging.2017.02.010

Baker, D. J., Wijshake, T., Tchkonia, T., LeBrasseur, N. K., Childs, B. G., van de Sluis, B., … van Deursen, J. M. (2011). Clearance of p16Ink4a-positive senescent cells delays ageing-associated disorders. *Nature*, *479*(7372), 232–236. doi:10.1038/nature10600

Bakulski, K. M., Dolinoy, D. C., Sartor, M. A., Paulson, H. L., Konen, J. R., Lieberman, A. P., … Rozek, L. S. (2012). Genome-wide DNA methylation differences between late-onset Alzheimer's disease and cognitively normal controls in human frontal cortex. *Journal of Alzheimer's Disease*, *29*(3), 571–588. doi:10.3233/JAD-2012-111223

Bandaru, V. V. R., Troncoso, J., Wheeler, D., Pletnikova, O., Wang, J., Conant, K., & Haughey, N. J. (2009). ApoE4 disrupts sterol and sphingolipid metabolism in Alzheimer's but not normal brain. *Neurobiology of Aging*, *30*(4), 591–599. doi:10.1016/j.neurobiolaging.2007.07.024

Bannwarth, S., Ait-El-Mkadem, S., Chaussenot, A., Genin, E. C., Lacas-Gervais, S., Fragaki, K., … Paquis-Flucklinger, V. (2014). A mitochondrial origin for frontotemporal dementia and amyotrophic lateral sclerosis through CHCHD10 involvement. *Brain: A Journal of Neurology*, *137*(Pt 8), 2329–2345. doi:10.1093/brain/awu138

Bartzokis, G. (2004). Age-related myelin breakdown: a developmental model of cognitive decline and Alzheimer's disease. *Neurobiology of Aging*, *25*(1), 5–18; author reply 49. doi:10.1016/j.neurobiolaging.2003.03.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

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), 213–227. doi:10.1177/0891988710383571

Bell, J. E., Alafuzoff, I., Al-Sarraj, S., Arzberger, T., Bogdanovic, N., Budka, H., … Kretzschmar, H. (2008). Management of a twenty-first century brain bank: experience in the BrainNet Europe consortium. *Acta Neuropathologica*, *115*(5), 497–507. doi:10.1007/s00401-008-0360-8

Bellenguez, Céline, Charbonnier, C., Grenier-Boley, B., Quenez, O., Le Guennec, K., Nicolas, G., … CNR MAJ collaborators. (2017). Contribution to Alzheimer's disease risk of rare variants in *TREM2*, *SORL1*, and *ABCA7* in 1779 cases and 1273 controls. *Neurobiology of Aging*, *59*, 220.e1–220.e9. doi:10.1016/j.neurobiolaging.2017.07.001

Bellenguez, Celine, Kucukali, F., Jansen, I., Andrade, V., Morenau-Grau, S., Amin, N., … Garcia, P. (2020). Large meta-analysis of genome-wide association studies expands knowledge of the genetic etiology of Alzheimer disease and highlights potential translational opportunities. *medRxiv*. doi:10.1101/2020.10.01.20200659

Bertram, L., & Tanzi, R. E. (2020). Genomic mechanisms in Alzheimer's disease. *Brain Pathology*, *30*(5), 966–977. doi:10.1111/bpa.12882

Bettencourt, C., Foti, S. C., Miki, Y., Botia, J., Chatterjee, A., Warner, T. T., … Holton, J. L. (2020). White matter DNA methylation profiling reveals deregulation of *HIP1*, LMAN2, MOBP, and other loci in multiple system atrophy. *Acta Neuropathologica*, *139*(1), 135–156. doi:10.1007/s00401-019-02074-0

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., … Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, *98*(4), 288–295. doi:10.1016/j.ygeno.2011.07.007

Biffi, A., Shulman, J. M., Jagiella, J. M., Cortellini, L., Ayres, A. M., Schwab, K., … Rosand, J. (2012). Genetic variation at *CR1* increases risk of cerebral amyloid angiopathy. *Neurology*, *78*(5), 334–341. doi:10.1212/WNL.0b013e3182452b40

Biolabs, N. E. (2016). Bisulfite Conversion | NEB. Retrieved February 2, 2021, from https://international.neb.com/applications/epigenetics/dna-methylation-analysis/bisulfite-conversion

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature*, *321*(6067), 209–213. doi:10.1038/321209a0

Blauwendraat, C., Faghri, F., Pihlstrom, L., Geiger, J. T., Elbaz, A., Lesage, S., … Scholz, S. W. (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. doi:10.1016/j.neurobiolaging.2017.05.009

Bonder, M. J., Luijk, R., Zhernakova, D. V., Moed, M., Deelen, P., Vermaat, M., … Bot, J. (2017). Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics*, *49*(1), 131–138. doi:10.1038/ng.3721

Borson, S., Scanlan, J. M., Watanabe, J., Tu, S.-P., & Lessig, M. (2005). Simplifying detection of cognitive impairment: comparison of the Mini-Cog and Mini-Mental State Examination in a multiethnic sample. *Journal of the American Geriatrics Society*, *53*(5), 871–874. doi:10.1111/j.1532-5415.2005.53269.x

Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H., & Del Tredici, K. (2006). Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathologica*, *112*(4), 389–404. doi:10.1007/s00401-006-0127-z

Braak, H., Bohl, J. R., Müller, C. M., Rüb, U., de Vos, R. A. I., & Del Tredici, K. (2006). Stanley Fahn Lecture 2005: The staging procedure for the inclusion body pathology associated with sporadic Parkinson’s disease reconsidered. *Movement Disorders*, *21*(12), 2042–2051. doi:10.1002/mds.21065

Bradley, R. M., Mardian, E. B., Bloemberg, D., Aristizabal Henao, J. J., Mitchell, A. S., Marvyn, P. M., … Duncan, R. E. (2017). Mice Deficient in lysophosphatidic acid acyltransferase delta (Lpaatδ)/acylglycerophosphate acyltransferase 4 (*AGPAT4*) Have Impaired Learning and Memory. *Molecular and Cellular Biology*, *37*(22). doi:10.1128/MCB.00245-17

Bradley, R. M., Marvyn, P. M., Aristizabal Henao, J. J., Mardian, E. B., George, S., Aucoin, M. G., … Duncan, R. E. (2015). Acylglycerophosphate acyltransferase 4 (*AGPAT4*) is a mitochondrial lysophosphatidic acid acyltransferase that regulates brain phosphatidylcholine, phosphatidylethanolamine, and phosphatidylinositol levels. *Biochimica et Biophysica Acta*, *1851*(12), 1566–1576. doi:10.1016/j.bbalip.2015.09.005

Breton, C. V., & Marutani, A. N. (2014). Air pollution and epigenetics: recent findings. *Current environmental health reports*, *1*(1), 35–45. doi:10.1007/s40572-013-0001-9

Broce, I. J., Tan, C. H., Fan, C. C., Jansen, I., Savage, J. E., Witoelar, A., … Desikan, R. S. (2019). Dissecting the genetic relationship between cardiovascular risk factors and Alzheimer’s disease. *Acta Neuropathologica*, *137*(2), 209–226. doi:10.1007/s00401-018-1928-6

Brookes, K. J., McConnell, G., Williams, K., Chaudhury, S., Madhan, G., Patel, T., … 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), 355–362. doi:10.3233/JAD-180191

Buck, N., & McFall, S. (2011). Understanding Society: design overview. *Longitudinal and Life Course Studies*.

Bullido, M. J., Martínez-García, A., Artiga, M. J., Aldudo, J., Sastre, I., Gil, P., … Valdivieso, F. (2007). A TAP2 genotype associated with Alzheimer’s disease in APOE4 carriers. *Neurobiology of Aging*, *28*(4), 519–523. doi:10.1016/j.neurobiolaging.2006.02.011

Burgess, S., Dudbridge, F., & Thompson, S. G. (2016). Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in Medicine*, *35*(11), 1880–1906. doi:10.1002/sim.6835

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, *8*(12), e1002822. doi:10.1371/journal.pcbi.1002822

Cacabelos, R., Cacabelos, P., & Torrellas, C. (2014). Personalized medicine of alzheimer's disease. In *Handbook of pharmacogenomics and stratified medicine* (pp. 563–615). Elsevier. doi:10.1016/B978-0-12-386882-4.00027-X

Cao, W., & Zheng, H. (2018). Peripheral immune system in aging and Alzheimer's disease. *Molecular Neurodegeneration*, *13*(1), 51. doi:10.1186/s13024-018-0284-2

Cervantes, S., Samaranch, L., Vidal-Taboada, J. M., Lamet, I., Bullido, M. J., Frank-García, A., … Pastor, P. (2011). Genetic variation in *APOE* cluster region and Alzheimer's disease risk. *Neurobiology of Aging*, *32*(11), 2107.e7–17. doi:10.1016/j.neurobiolaging.2011.05.023

Chai, B., Gao, F., Wu, R., Dong, T., Gu, C., Lin, Q., & Zhang, Y. (2019). Vitamin D deficiency as a risk factor for dementia and Alzheimer's disease: an updated meta-analysis. *BMC Neurology*, *19*(1), 284. doi:10.1186/s12883-019-1500-6

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7. doi:10.1186/s13742-015-0047-8

Chang, D., Nalls, M. A., Hallgrímsdóttir, I. B., Hunkapiller, J., van der Brug, M., Cai, F., … Graham. (2017). A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nature Genetics*, *49*(10), 1511–1516. doi:10.1038/ng.3955

Chen, P.-H., Wang, R.-L., Liou, D.-J., & Shaw, J.-S. (2013). Gait disorders in parkinson's disease: assessment and management. *International journal of gerontology*, *7*(4), 189–193. doi:10.1016/j.ijge.2013.03.005

Chen, Y., Lemire, M., Choufani, S., Butcher, D. T., Grafodatskaya, D., Zanke, B. W., … Weksberg, R. (2013). Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, *8*(2), 203–209. doi:10.4161/epi.23470

Choi, B. W., Kim, S., Kang, S., Won, K. S., Yi, H.-A., & Kim, H. W. (2020). Relationship between thyroid hormone levels and the pathology of alzheimer's disease in euthyroid subjects. *Thyroid*. doi:10.1089/thy.2019.0727

Choi, S. W., & O'Reilly, P. F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*, *8*(7). doi:10.1093/gigascience/giz082

Chou, C.-C., Zhang, Y., Umoh, M. E., Vaughan, S. W., Lorenzini, I., Liu, F., … Rossoll, W. (2018). TDP-43 pathology disrupts nuclear pore complexes and nucleocytoplasmic transport in ALS/FTD. *Nature Neuroscience*, *21*(2), 228–239. doi:10.1038/s41593-017-0047-3

Chouliaras, L., Pishva, E., Haapakoski, R., Zsoldos, E., Mahmood, A., Filippini, N., … Ebmeier, K. P. (2018). Peripheral DNA methylation, cognitive decline and brain aging: pilot findings from the Whitehall II imaging study. *Epigenomics*, *10*(5), 585–595. doi:10.2217/epi-2017-0132

Chuang, Y.-H., Paul, K. C., Bronstein, J. M., Bordelon, Y., Horvath, S., & Ritz, B. (2017). Parkinson's disease is associated with DNA methylation levels in human blood and saliva. *Genome Medicine*, *9*(1), 76. doi:10.1186/s13073-017-0466-5

Coneys, R., & Wood, I. C. (2020). Alzheimer's disease: the potential of epigenetic treatments and current clinical candidates. *Neurodegenerative disease management*, *10*(3), 543–558. doi:10.2217/nmt-2019-0034

Coutelier, M., Blesneac, I., Monteil, A., Monin, M.-L., Ando, K., Mundwiller, E., … Stevanin, G. (2015). A Recurrent Mutation in CACNA1G Alters Cav3.1 T-Type Calcium-Channel Conduction and Causes Autosomal-Dominant Cerebellar Ataxia. *American Journal of Human Genetics*, *97*(5), 726–737. doi:10.1016/j.ajhg.2015.09.007

Cruchaga, C., Kauwe, J. S. K., Harari, O., Jin, S. C., Cai, Y., Karch, C. M., … Goate, A. M. (2013). GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease. *Neuron*, *78*(2), 256–268. doi:10.1016/j.neuron.2013.02.026

Cummings, J. (2019). The National Institute on Aging-Alzheimer's Association Framework on Alzheimer's disease: Application to clinical trials. *Alzheimer's & Dementia*, *15*(1), 172–178. doi:10.1016/j.jalz.2018.05.006

Czinn, S. J., & Blanchard, S. S. (2011). Developmental anatomy and physiology of the stomach. In *Pediatric gastrointestinal and liver disease* (pp. 262–268.e1). Elsevier. doi:10.1016/B978-1-4377-0774-8.10025-9

Danysz, W., & Parsons, C. G. (2012). Alzheimer's disease, β-amyloid, glutamate, NMDA receptors and memantine--searching for the connections. *British Journal of Pharmacology*, *167*(2), 324–352. doi:10.1111/j.1476-5381.2012.02057.x

Das, P. M., & Singal, R. (2004). DNA methylation and cancer. *Journal of Clinical Oncology*, *22*(22), 4632–4642. doi:10.1200/JCO.2004.07.151

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. doi:10.1038/ng.3656

Datta, S. R., McQuillin, A., Rizig, M., Blaveri, E., Thirumalai, S., Kalsi, G., … Gurling, H. M. D. (2010). A threonine to isoleucine missense mutation in the pericentriolar material 1 gene is strongly associated with schizophrenia. *Molecular Psychiatry*, *15*(6), 615–628. doi:10.1038/mp.2008.128

Deane, R., Sagare, A., Hamm, K., Parisi, M., Lane, S., Finn, M. B., ... & Zlokovic, B. V. (2008). apoE isoform–specific disruption of amyloid β peptide clearance from mouse brain. *The Journal of clinical investigation*, *118*(12), 4002-4013.

De Jager, P. L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L. C., Yu, L., … Bennett, D. A. (2014). Alzheimer's disease: early alterations in brain DNA methylation at *ANK1*, *BIN1*, *RHBDF2* and other loci. *Nature Neuroscience*, *17*(9), 1156–1163. doi:10.1038/nn.3786

de Rojas, I., Moreno-Grau, S., Tesi, N., Grenier-Boley, B., Andrade, V., Jansen, I., … Hernández, I. (2020). Common variants in Alzheimer's disease: Novel association of six genetic variants with AD and risk stratification by polygenic risk scores. *medRxiv*. doi:10.1101/19012021

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., & Fuks, F. (2011). Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, *3*(6), 771–784. doi:10.2217/epi.11.105

Deming, Y., Filipello, F., Cignarella, F., Cantoni, C., Hsu, S., Mikesell, R., … Cruchaga, C. (2019). The MS4A gene cluster is a key modulator of soluble *TREM2* and Alzheimer's disease risk. *Science Translational Medicine*, *11*(505). doi:10.1126/scitranslmed.aau2291

Desai, M. K., Mastrangelo, M. A., Ryan, D. A., Sudol, K. L., Narrow, W. C., & Bowers, W. J. (2010). Early oligodendrocyte/myelin pathology in Alzheimer's disease mice constitutes a novel therapeutic target. *The American Journal of Pathology*, *177*(3), 1422–1435. doi:10.2353/ajpath.2010.100087

Desikan, R S, Schork, A. J., Wang, Y., Witoelar, A., Sharma, M., McEvoy, L. K., … ADNI, ADGC, GERAD, CHARGE and IPDGC Investigators. (2015). Genetic overlap between Alzheimer's disease and Parkinson's disease at the MAPT locus. *Molecular Psychiatry*, *20*(12), 1588–1595. doi:10.1038/mp.2015.6

Desikan, Rahul S, Fan, C. C., Wang, Y., Schork, A. J., Cabral, H. J., Cupples, L. A., … 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. doi:10.1371/journal.pmed.1002258

DeTure, M. A., & Dickson, D. W. (2019). The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration*, *14*(1), 32. doi:10.1186/s13024-019-0333-5

Di Paolo, G., & Kim, T.-W. (2011). Linking lipids to Alzheimer's disease: cholesterol and beyond. *Nature Reviews. Neuroscience*, *12*(5), 284–296. doi:10.1038/nrn3012

Dickerson, B. C., Brickhouse, M., McGinnis, S., & Wolk, D. A. (2017). Alzheimer's disease: The influence of age on clinical heterogeneity through the human brain connectome. *Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring*, *6*, 122–135. doi:10.1016/j.dadm.2016.12.007

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, *11*, 587. doi:10.1186/1471-2105-11-587

Dubois, B., Feldman, H. H., Jacova, C., Hampel, H., Molinuevo, J. L., Blennow, K., … Cummings, J. L. (2014). Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *Lancet Neurology*, *13*(6), 614–629. doi:10.1016/S1474-4422(14)70090-0

Eggert, S., Thomas, C., Kins, S., & Hermey, G. (2018). Trafficking in alzheimer's disease: modulation of *APP* transport and processing by the transmembrane proteins LRP1, sorla, sorcs1c, sortilin, and calsyntenin. *Molecular Neurobiology*, *55*(7), 5809–5829. doi:10.1007/s12035-017-0806-x

El Khoury, L. Y., Gorrie-Stone, T., Smart, M., Hughes, A., Bao, Y., Andrayas, A., … Schalkwyk, L. C. (2019). Systematic underestimation of the epigenetic clock and age acceleration in older subjects. *Genome Biology*, *20*(1), 283. doi:10.1186/s13059-019-1810-4

Elliott, H. R., Tillin, T., McArdle, W. L., Ho, K., Duggirala, A., Frayling, T. M., … Relton, C. L. (2014). Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clinical epigenetics*, *6*(1), 4. doi:10.1186/1868-7083-6-4

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., … AIBL Research Group. (2009). The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, *21*(4), 672–687. doi:10.1017/S1041610209009405

Ellison, E. M., Bradley-Whitman, M. A., & Lovell, M. A. (2017). Single-Base Resolution Mapping of 5-Hydroxymethylcytosine Modifications in Hippocampus of Alzheimer's Disease Subjects. *Journal of Molecular Neuroscience*, *63*(2), 185–197. doi:10.1007/s12031-017-0969-y

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. doi:10.14283/jpad.2018.46

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

Escott-Price, Valentina, 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.e7–214.e11. doi:10.1016/j.neurobiolaging.2016.07.018

Escott-Price, Valentina, Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., … Williams, J. (2015). Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain: A Journal of Neurology*, *138*(Pt 12), 3673–3684. doi:10.1093/brain/awv268

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic Risk Score software. *Bioinformatics*, *31*(9), 1466–1468. doi:10.1093/bioinformatics/btu848

Farrer, L. A. (1997). Effects of Age, Sex, and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease. *The Journal of the American Medical Association*, *278*(16), 1349. doi:10.1001/jama.1997.03550160069041

Felsky, D., Patrick, E., Schneider, J. A., Mostafavi, S., Gaiteri, C., Patsopoulos, N., … De Jager, P. L. (2018). Polygenic analysis of inflammatory disease variants and effects on microglia in the aging brain. *Molecular Neurodegeneration*, *13*(1), 38. doi:10.1186/s13024-018-0272-6

Ferguson-Smith, A. C. (2011). Genomic imprinting: the emergence of an epigenetic paradigm. *Nature Reviews. Genetics*, *12*(8), 565–575. doi:10.1038/nrg3032

Fernandez, A. F., Assenov, Y., Martin-Subero, J. I., Balint, B., Siebert, R., Taniguchi, H., … Esteller, M. (2012). A DNA methylation fingerprint of 1628 human samples. *Genome Research*, *22*(2), 407–419. doi:10.1101/gr.119867.110

Ferrari, R., Hernandez, D. G., Nalls, M. A., Rohrer, J. D., Ramasamy, A., Kwok, J. B. J., … Halliday, G. M. (2014). Frontotemporal dementia and its subtypes: a genome-wide association study. *Lancet Neurology*, *13*(7), 686–699. doi:10.1016/S1474-4422(14)70065-1

Ferrari, R., Wang, Y., Vandrovcova, J., Guelfi, S., Witeolar, A., Karch, C. M., … Desikan, R. S. (2017). Genetic architecture of sporadic frontotemporal dementia and overlap with Alzheimer’s and Parkinson's diseases. *Journal of Neurology, Neurosurgery, and Psychiatry*, *88*(2), 152–164. doi:10.1136/jnnp-2016-314411

Filzmoser, P., Hron, K., & Reimann, C. (2012). Interpretation of multivariate outliers for compositional data. *Computers & geosciences*, *39*, 77–85. doi:10.1016/j.cageo.2011.06.014

Finch, N. A., Wang, X., Baker, M. C., Heckman, M. G., Gendron, T. F., Bieniek, K. F., … van Blitterswijk, M. (2017). Abnormal expression of homeobox genes and transthyretin in C9ORF72 expansion carriers. *Neurology. Genetics*, *3*(4), e161. doi:10.1212/NXG.0000000000000161

Fitzpatrick, A. L., Kuller, L. H., Lopez, O. L., Diehr, P., O’Meara, E. S., Longstreth, W. T., & Luchsinger, J. A. (2009). Midlife and late-life obesity and the risk of dementia: cardiovascular health study. *Archives of Neurology*, *66*(3), 336–342. doi:10.1001/archneurol.2008.582

Flutre, T., Wen, X., Pritchard, J., & Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics*, *9*(5), e1003486. doi:10.1371/journal.pgen.1003486

Folstein, M. F., Robins, L. N., & Helzer, J. E. (1983). The Mini-Mental State Examination. *Archives of General Psychiatry*, *40*(7), 812. doi:10.1001/archpsyc.1983.01790060110016

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. doi:10.3233/JAD-180699

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, *33*(1), 1–22. doi:10.18637/jss.v033.i01

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., … Shah, H. R. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience*, *19*(11), 1442–1453. doi:10.1038/nn.4399

Funayama, M., Ohe, K., Amo, T., Furuya, N., Yamaguchi, J., Saiki, S., … Hattori, N. (2015). CHCHD2 mutations in autosomal dominant late-onset Parkinson’s disease: a genome-wide linkage and sequencing study. *Lancet Neurology*, *14*(3), 274–282. doi:10.1016/S1474-4422(14)70266-2

Galvin, J. E., Roe, C. M., Powlishta, K. K., Coats, M. A., Muich, S. J., Grant, E., … Morris, J. C. (2005). The AD8: a brief informant interview to detect dementia. *Neurology*, *65*(4), 559–564. doi:10.1212/01.wnl.0000172958.95282.2a

Gao, L., Fang, Z., Zhang, K., Zhi, D., & Cui, X. (2011). Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics*, *27*(5), 662–669. doi:10.1093/bioinformatics/btr005

Gaskin, F., Finley, J., Fang, Q., Xu, S., & Fu, S. M. (1993). Human antibodies reactive with beta-amyloid protein in Alzheimer's disease. *The Journal of Experimental Medicine*, *177*(4), 1181–1186. doi:10.1084/jem.177.4.1181

Gasparoni, G., Bultmann, S., Lutsik, P., Kraus, T. F. J., Sordon, S., Vlcek, J., … Walter, J. (2018). DNA methylation analysis on purified neurons and glia dissects age and Alzheimer's disease-specific changes in the human cortex. *Epigenetics & Chromatin*, *11*(1), 41. doi:10.1186/s13072-018-0211-3

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

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., … Relton, C. L. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, *17*, 61. doi:10.1186/s13059-016-0926-z

Gauthreaux, K., Bonnett, T. A., Besser, L. M., Brenowitz, W. D., Teylan, M., Mock, C., … Kukull, W. A. (2020). Concordance of clinical alzheimer diagnosis and neuropathological features at autopsy. *Journal of Neuropathology and Experimental Neurology*, *79*(5), 465–473. doi:10.1093/jnen/nlaa014

Ge, T., Sabuncu, M. R., Smoller, J. W., Sperling, R. A., Mormino, E. C., & Alzheimer's Disease Neuroimaging Initiative. (2018). Dissociable influences of *APOE* ε4 and polygenic risk of AD dementia on amyloid and cognition. *Neurology*, *90*(18), e1605–e1612. doi:10.1212/WNL.0000000000005415

Geeleher, P., Hartnett, L., Egan, L. J., Golden, A., Raja Ali, R. A., & Seoighe, C. (2013). Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics*, *29*(15), 1851–1857. doi:10.1093/bioinformatics/btt311

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, *10*(5), e1004383. doi:10.1371/journal.pgen.1004383

Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., … Roussos, P. (2018). A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, *34*(15), 2538–2545. doi:10.1093/bioinformatics/bty147

Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H., & Kellis, M. (2015). Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature*, *518*(7539), 365–369. doi:10.1038/nature14252

Gnjec, A., D'Costa, K. J., Laws, S. M., Hedley, R., Balakrishnan, K., Taddei, K., … Martins, R. N. (2008). Association of alleles carried at TNFA -850 and BAT1 -22 with Alzheimer's disease. *Journal of Neuroinflammation*, *5*, 36. doi:10.1186/1742-2094-5-36

Gómez-Isla, T., Hollister, R., West, H., Mui, S., Growdon, J. H., Petersen, R. C., … Hyman, B. T. (1997). Neuronal loss correlates with but exceeds neurofibrillary tangles in Alzheimer's disease. *Annals of Neurology*, *41*(1), 17–24. doi:10.1002/ana.410410106

Goodman, R. A., Lochner, K. A., Thambisetty, M., Wingo, T. S., Posner, S. F., & Ling, S. M. (2017). Prevalence of dementia subtypes in United States Medicare fee-for-service beneficiaries, 2011-2013. *Alzheimer's & Dementia*, *13*(1), 28–37. doi:10.1016/j.jalz.2016.04.002

Gopalan, S., Gaige, J., & Henn, B. M. (2019). DNA methylation-based forensic age estimation in human bone. *BioRxiv*. doi:10.1101/801647

Gorrie-Stone, T. J., Smart, M. C., Saffari, A., Malki, K., Hannon, E., Burrage, J., … Schalkwyk, L. C. (2019). Bigmelon: tools for analysing large DNA methylation datasets. *Bioinformatics*, *35*(6), 981–986. doi:10.1093/bioinformatics/bty713

Greco, E., Aita, A., Galozzi, P., Gava, A., Sfriso, P., Negm, O. H., … Punzi, L. (2015). The novel S59P mutation in the *TNFRSF1A* gene identified in an adult onset TNF receptor associated periodic syndrome (TRAPS) constitutively activates NF-κB pathway. *Arthritis Research & Therapy*, *17*, 93. doi:10.1186/s13075-015-0604-7

Griciuc, A., Serrano-Pozo, A., Parrado, A. R., Lesinski, A. N., Asselin, C. N., Mullin, K., … Tanzi, R. E. (2013). Alzheimer's disease risk gene *CD33* inhibits microglial uptake of amyloid beta. *Neuron*, *78*(4), 631–643. doi:10.1016/j.neuron.2013.04.014

GTEx Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. doi:10.1126/science.1262110

Guerreiro, R., & Bras, J. (2015). The age factor in Alzheimer's disease. *Genome Medicine*, *7*, 106. doi:10.1186/s13073-015-0232-5

Guerreiro, R., Escott-Price, V., Darwent, L., Parkkinen, L., Ansorge, O., Hernandez, D. G., … Bras, J. (2016). Genome-wide analysis of genetic correlation in dementia with Lewy bodies, Parkinson's and Alzheimer's diseases. *Neurobiology of Aging*, *38*, 214.e7–214.e10. doi:10.1016/j.neurobiolaging.2015.10.028

Guerreiro, R., Ross, O. A., Kun-Rodrigues, C., Hernandez, D. G., Orme, T., Eicher, J. D., … Heckman, M. G. (2018). Investigating the genetic architecture of dementia with Lewy bodies: a two-stage genome-wide association study. *Lancet Neurology*, *17*(1), 64–74. doi:10.1016/S1474-4422(17)30400-3

Guintivano, J., Aryee, M. J., & Kaminsky, Z. A. (2013). A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*, *8*(3), 290–302. doi:10.4161/epi.23924

Guo, H., Fortune, M. D., Burren, O. S., Schofield, E., Todd, J. A., & Wallace, C. (2015). Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Human Molecular Genetics*, *24*(12), 3305–3313. doi:10.1093/hmg/ddv077

Gurnot, C., Martin-Subero, I., Mah, S. M., Weikum, W., Goodman, S. J., Brain, U., … Hensch, T. K. (2015). Prenatal antidepressant exposure associated with CYP2E1 DNA methylation change in neonates. *Epigenetics*, *10*(5), 361–372. doi:10.1080/15592294.2015.1026031

Hamilton, R. L. (2006). Lewy Bodies in Alzheimer's Disease: A Neuropathological Review of 145 Cases Using α-Synuclein Immunohistochemistry. *Brain Pathology*, *10*(3), 378–384. doi:10.1111/j.1750-3639.2000.tb00269.x

Hannon, E., Bray, N., Weedon, M., Gorrie-Stone, T., Smart, M., Kumari, M., … Mill, J. (2019). Pleiotropic effects of genetic variation associated with psychiatric disorders on dna methylation. *European Neuropsychopharmacology*, *29*, S984–S985. doi:10.1016/j.euroneuro.2017.08.362

Hannon, E., Dempster, E., Viana, J., Burrage, J., Smith, A. R., Macdonald, R., … Mill, J. (2016). An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biology*, *17*(1), 176. doi:10.1186/s13059-016-1041-x

Hannon, E., Gorrie-Stone, T. J., Smart, M. C., Burrage, J., Hughes, A., Bao, Y., … Mill, J. (2018). Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits. *American Journal of Human Genetics*, *103*(5), 654–665. doi:10.1016/j.ajhg.2018.09.007

Hannon, E., Knox, O., Sugden, K., Burrage, J., Wong, C. C. Y., Belsky, D. W., … Mill, J. (2018). Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genetics*, *14*(8), e1007544. doi:10.1371/journal.pgen.1007544

Hannon, E., Lunnon, K., Schalkwyk, L., & Mill, J. (2015). Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics*, *10*(11), 1024–1032. doi:10.1080/15592294.2015.1100786

Hannon, E., Schendel, D., Ladd-Acosta, C., Grove, J., iPSYCH-Broad ASD Group, Hansen, C. S., … Mill, J. (2018). Elevated polygenic burden for autism is associated with differential DNA methylation at birth. *Genome Medicine*, *10*(1), 19. doi:10.1186/s13073-018-0527-4

Hannon, E., Shireby, G. L., Brookes, K., Attems, J., Sims, R., Cairns, N. J., … Mill, J. (2020). Genetic risk for Alzheimer's disease influences neuropathology via multiple biological pathways. *Brain Communications*, *2*(2), fcaa167. doi:10.1093/braincomms/fcaa167

Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., … Mill, J. (2016). Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nature Neuroscience*, *19*(1), 48–54. doi:10.1038/nn.4182

Hannon, E., Weedon, M., Bray, N., O'Donovan, M., & Mill, J. (2017). Pleiotropic Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining GWAS Loci. *American Journal of Human Genetics*, *100*(6), 954–959. doi:10.1016/j.ajhg.2017.04.013

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., … Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, *49*(2), 359–367. doi:10.1016/j.molcel.2012.10.016

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. doi:10.1016/0165-6147(91)90609-V

Hardy, John, & Selkoe, D. J. (2002). The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science*, *297*(5580), 353–356. doi:10.1126/science.1072994

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

Harper, S. (2014). Economic and social implications of aging societies. *Science*, *346*(6209), 587–591. doi:10.1126/science.1254405

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–1774. doi:10.1101/gr.135350.111

He, M., Rutledge, S. L., Kelly, D. R., Palmer, C. A., Murdoch, G., Majumder, N., … Vockley, J. (2007). A new genetic disorder in mitochondrial fatty acid beta-oxidation: ACAD9 deficiency. *American Journal of Human Genetics*, *81*(1), 87–103. doi:10.1086/519219

He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X., & Li, H. (2013). Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *American Journal of Human Genetics*, *92*(5), 667–680. doi:10.1016/j.ajhg.2013.03.022

Heard, E., Clerc, P., & Avner, P. (1997). X-chromosome inactivation in mammals. *Annual Review of Genetics*, *31*, 571–610. doi:10.1146/annurev.genet.31.1.571

Henikoff, S., & Matzke, M. A. (1997). Exploring and explaining epigenetic effects. *Trends in Genetics*, *13*(8), 293–295.

Heppner, F. L., Ransohoff, R. M., & Becher, B. (2015). Immune attack: the role of inflammation in Alzheimer disease. *Nature Reviews. Neuroscience*, *16*(6), 358–372. doi:10.1038/nrn3880

Huichalaf, C. H., Al-Ramahi, I., Park, K. W., Grunke, S. D., Lu, N., de Haro, M., ... & Jankowsky, J. L. (2019). Cross-species genetic screens to identify kinase targets for *APP* reduction in Alzheimer's disease. *Human molecular genetics*, *28*(12), 2014-2029.

Hogan, D. B., Jetté, N., Fiest, K. M., Roberts, J. I., Pearson, D., Smith, E. E., … Maxwell, C. J. (2016). The prevalence and incidence of frontotemporal dementia: a systematic review. *The Canadian Journal of Neurological Sciences. Le Journal Canadien des Sciences Neurologiques*, *43 Suppl 1*, S96–S109. doi:10.1017/cjn.2016.25

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

Holtzman, D. M., Herz, J., & Bu, G. (2012). Apolipoprotein E and apolipoprotein E receptors: normal biology and roles in Alzheimer disease. *Cold Spring Harbor perspectives in medicine*, *2*(3), a006312. doi:10.1101/cshperspect.a006312

Honea, R. A., Vidoni, E. D., Swerdlow, R. H., Burns, J. M., & Alzheimer’s Disease Neuroimaging Initiative. (2012). Maternal family history is associated with Alzheimer’s disease biomarkers. *Journal of Alzheimer’s Disease*, *31*(3), 659–668. doi:10.3233/JAD-2012-120676

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*(10), R115. doi:10.1186/gb-2013-14-10-r115

Horvath, S., & Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews. Genetics*, *19*(6), 371–384. doi:10.1038/s41576-018-0004-3

Horvath, S., & Ritz, B. R. (2015). Increased epigenetic age and granulocyte counts in the blood of Parkinson’s disease patients. *Aging*, *7*(12), 1130–1142. doi:10.18632/aging.100859

Hoss, A. G., Kartha, V. K., Dong, X., Latourelle, J. C., Dumitriu, A., Hadzi, T. C., … Myers, R. H. (2014). MicroRNAs located in the Hox gene clusters are implicated in huntington’s disease pathogenesis. *PLoS Genetics*, *10*(2), e1004188. doi:10.1371/journal.pgen.1004188

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., … Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, *13*, 86. doi:10.1186/1471-2105-13-86

Hsu, P. D., Lander, E. S., & Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, *157*(6), 1262–1278. doi:10.1016/j.cell.2014.05.010

Hu, H., Eggers, K., Chen, W., Garshasbi, M., Motazacker, M. M., Wrogemann, K., … Kuss, A. W. (2011). ST3GAL3 mutations impair the development of higher cognitive functions. *American Journal of Human Genetics*, *89*(3), 407–414. doi:10.1016/j.ajhg.2011.08.008

Hu, Y., Lu, Q., Liu, W., Zhang, Y., Li, M., & Zhao, H. (2017). Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS Genetics*, *13*(6), e1006836. doi:10.1371/journal.pgen.1006836

Hyman, B. T., Phelps, C. H., Beach, T. G., Bigio, E. H., Cairns, N. J., Carrillo, M. C., … Montine, T. J. (2012). National Institute on Aging-Alzheimer’s Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimer’s & Dementia*, *8*(1), 1–13. doi:10.1016/j.jalz.2011.10.007

Ibanez, L., Dube, U., Davis, A. A., Fernandez, M. V., Budde, J., Cooper, B., … Benitez, B. A. (2018). Pleiotropic effects of variants in dementia genes in parkinson disease. *Frontiers in Neuroscience*, *12*, 230. doi:10.3389/fnins.2018.00230

Imai, F. L., Uzawa, K., Nimura, Y., Moriya, T., Imai, M. A., Shiiba, M., … Tanzawa, H. (2005). Chromosome 1 open reading frame 10 (C1orf10) gene is frequently down-regulated and inhibits cell proliferation in oral squamous cell carcinoma. *The International Journal of Biochemistry & Cell Biology*, *37*(8), 1641–1655. doi:10.1016/j.biocel.2005.02.005

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. doi:10.1038/nature04226

International Schizophrenia Consortium. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, *455*(7210), 237–241. doi:10.1038/nature07239

Iotchkova, V., Ritchie, G. R. S., Geihs, M., Morganella, S., Min, J. L., Walter, K., … Soranzo, N. (2016). GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. *BioRxiv*. doi:10.1101/085738

Iotchkova, V., Ritchie, G. R. S., Geihs, M., Morganella, S., Min, J. L., Walter, K., … Soranzo, N. (2019). GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nature Genetics*, *51*(2), 343–353. doi:10.1038/s41588-018-0322-6

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., … Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurology*, *9*(1), 119–128. doi:10.1016/S1474-4422(09)70299-6

Jaffe, A. E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T. M., Weinberger, D. R., & Kleinman, J. E. (2016). Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nature Neuroscience*, *19*(1), 40–47. doi:10.1038/nn.4181

Jamshidi, Y., Snieder, H., Ge, D., Spector, T. D., & O'Dell, S. D. (2007). The SH2B gene is associated with serum leptin and body fat in normal female twins. *Obesity*, *15*(1), 5–9. doi:10.1038/oby.2007.637

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., … Athanasiu, L. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics*, *51*(3), 404–413. doi:10.1038/s41588-018-0311-9

Jia, Y.-L., Jing, L.-J., Li, J.-Y., Lu, J.-J., Han, R., Wang, S.-Y., … Jia, Y.-J. (2011). Expression and significance of DSCAM in the cerebral cortex of *APP* transgenic mice. *Neuroscience Letters*, *491*(2), 153–157. doi:10.1016/j.neulet.2011.01.028

Jiang, S., Avraham, H. K., Park, S.-Y., Kim, T.-A., Bu, X., Seng, S., & Avraham, S. (2005). Process elongation of oligodendrocytes is promoted by the Kelch-related actin-binding protein Mayven. *Journal of Neurochemistry*, *92*(5), 1191–1203. doi:10.1111/j.1471-4159.2004.02946.x

Jones, L., Holmans, P. A., Hamshere, M. L., Harold, D., Moskvina, V., Ivanov, D., … Sims, R. (2010). Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer's disease. *Plos One*, *5*(11), e13950. doi:10.1371/journal.pone.0013950

Jongsma, H. E., Gayer-Anderson, C., Lasalvia, A., Quattrone, D., Mulè, A., Szöke, A., … European Network of National Schizophrenia Networks Studying Gene-Environment Interactions Work Package 2 (EU-GEI WP2) Group. (2018). Treated Incidence of Psychotic Disorders in the Multinational EU-GEI Study. *JAMA psychiatry*, *75*(1), 36–46. doi:10.1001/jamapsychiatry.2017.3554

Jorm, A. F., & Jacomb, P. A. (1989). The Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): Socio-demographic correlates, reliability, validity and some norms. *Psychological Medicine*, *19*(4), 1015–1022. doi:10.1017/S0033291700005742

Josephs, K. A., Whitwell, J. L., Weigand, S. D., Murray, M. E., Tosakulwong, N., Liesinger, A. M., … Dickson, D. W. (2014). TDP-43 is a key player in the clinical features associated with Alzheimer's disease. *Acta Neuropathologica*, *127*(6), 811–824. doi:10.1007/s00401-014-1269-z

Jostins, L., Morley, K. I., & Barrett, J. C. (2011). Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *European Journal of Human Genetics*, *19*(6), 662–666. doi:10.1038/ejhg.2011.10

Jouanne, M., Rault, S., & Voisin-Chiret, A.-S. (2017). Tau protein aggregation in Alzheimer's disease: An attractive target for the development of novel therapeutic agents. *European Journal of Medicinal Chemistry*, *139*, 153–167. doi:10.1016/j.ejmech.2017.07.070

Joubert, S., Gour, N., Guedj, E., Didic, M., Guériot, C., Koric, L., … Ceccaldi, M. (2016). Early-onset and late-onset Alzheimer's disease are associated with distinct patterns of memory impairment. *Cortex*, *74*, 217–232. doi:10.1016/j.cortex.2015.10.014

Jylhävä, J., Jiang, M., Foebel, A. D., Pedersen, N. L., & Hägg, S. (2019). Can markers of biological age predict dependency in old age? *Biogerontology*, *20*(3), 321–329. doi:10.1007/s10522-019-09795-5

Jylhävä, J., Pedersen, N. L., & Hägg, S. (2017). Biological Age Predictors. *EBioMedicine*, *21*, 29–36. doi:10.1016/j.ebiom.2017.03.046

Kametani, F., & Hasegawa, M. (2018). Reconsideration of amyloid hypothesis and tau hypothesis in alzheimer's disease. *Frontiers in Neuroscience*, *12*, 25. doi:10.3389/fnins.2018.00025

Kane, J. P. M., Surendranathan, A., Bentley, A., Barker, S. A. H., Taylor, J.-P., Thomas, A. J., … O'Brien, J. T. (2018). Clinical prevalence of Lewy body dementia. *Alzheimer's research & therapy*, *10*(1), 19. doi:10.1186/s13195-018-0350-6

Kapasi, A., DeCarli, C., & Schneider, J. A. (2017). Impact of multiple pathologies on the threshold for clinically overt dementia. *Acta Neuropathologica*, *134*(2), 171–186. doi:10.1007/s00401-017-1717-7

Karch, C. M., Ezerskiy, L. A., Bertelsen, S., Alzheimer's Disease Genetics Consortium (ADGC), & Goate, A. M. (2016). Alzheimer's disease risk polymorphisms regulate gene expression in the *ZCWPW1* and the *CELF1* loci. *Plos One*, *11*(2), e0148717. doi:10.1371/journal.pone.0148717

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., … University of California Santa Cruz. (2003). The UCSC genome browser database. *Nucleic Acids Research*, *31*(1), 51–54. doi:10.1093/nar/gkg129

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), 250. doi:10.1038/s41398-020-00934-y

Kent, W. J., 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. doi:10.1101/gr.229102

Kibinge, N. K., Relton, C. L., Gaunt, T. R., & Richardson, T. G. (2020). Characterizing the Causal Pathway for Genetic Variants Associated with Neurological Phenotypes Using Human Brain-Derived Proteome Data. *American Journal of Human Genetics*, *106*(6), 885–892. doi:10.1016/j.ajhg.2020.04.007

Kikuchi, M., Hara, N., Hasegawa, M., Miyashita, A., Kuwano, R., Ikeuchi, T., & Nakaya, A. (2019). Enhancer variants associated with Alzheimer's disease affect gene expression via chromatin looping. *BMC Medical Genomics*, *12*(1), 128. doi:10.1186/s12920-019-0574-8

Kim, Y., Sato, K., Asagiri, M., Morita, I., Soma, K., & Takayanagi, H. (2005). Contribution of nuclear factor of activated T cells c1 to the transcriptional control of immunoreceptor osteoclast-associated receptor but not triggering receptor expressed by myeloid cells-2 during osteoclastogenesis. *The Journal of Biological Chemistry*, *280*(38), 32905–32913. doi:10.1074/jbc.M505820200

Kobayashi, N., Shinagawa, S., Nagata, T., Shimada, K., Shibata, N., Ohnuma, T., … Kondo, K. (2016). Development of biomarkers based on DNA methylation in the NCAPH2/LMF2 promoter region for diagnosis of alzheimer's disease and amnesic mild cognitive impairment. *Plos One*, *11*(1), e0146449. doi:10.1371/journal.pone.0146449

Kocherhans, S., Madhusudan, A., Doehner, J., Breu, K. S., Nitsch, R. M., Fritschy, J.-M., & Knuesel, I. (2010). Reduced Reelin expression accelerates amyloid-beta plaque formation and tau pathology in transgenic Alzheimer's disease mice. *The Journal of Neuroscience*, *30*(27), 9228–9240. doi:10.1523/JNEUROSCI.0418-10.2010

Koedam, E. L. G. E., Lauffer, V., van der Vlies, A. E., van der Flier, W. M., Scheltens, P., & Pijnenburg, Y. A. L. (2010). Early-versus late-onset Alzheimer's disease: more than age alone. *Journal of Alzheimer's Disease*, *19*(4), 1401–1408. doi:10.3233/JAD-2010-1337

Konki, M., Malonzo, M., Karlsson, I. K., Lindgren, N., Ghimire, B., Smolander, J., … Lund, R. J. (2019). Peripheral blood DNA methylation differences in twin pairs discordant for Alzheimer's disease. *Clinical epigenetics*, *11*(1), 130. doi:10.1186/s13148-019-0729-7

Kosik, K. S., Joachim, C. L., & Selkoe, D. J. (1986). Microtubule-associated protein tau (tau) is a major antigenic component of paired helical filaments in Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America*, *83*(11), 4044–4048. doi:10.1073/pnas.83.11.4044

Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., … Amlie-Wolf, A. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nature Genetics*, *51*(3), 414–430. doi:10.1038/s41588-019-0358-2

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of statistical software*, *82*(13). doi:10.18637/jss.v082.i13

Labadorf, A., Hoss, A. G., Lagomarsino, V., Latourelle, J. C., Hadzi, T. C., Bregu, J., … Myers, R. H. (2015). RNA sequence analysis of human huntington disease brain reveals an extensive increase in inflammatory and developmental gene expression. *Plos One*, *10*(12), e0143563. doi:10.1371/journal.pone.0143563

Lachén-Montes, M., González-Morales, A., de Morentin, X. M., Pérez-Valderrama, E., Ausín, K., Zelaya, M. V., … Santamaría, E. (2016). An early dysregulation of FAK and MEK/ERK signaling pathways precedes the β-amyloid deposition in the olfactory bulb of *APP*/PS1 mouse model of Alzheimer's disease. *Journal of Proteomics*, *148*, 149–158. doi:10.1016/j.jprot.2016.07.032

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, *37*(13), 4181–4193. doi:10.1093/nar/gkp552

Laird, P. W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nature Reviews. Genetics*, *11*(3), 191–203. doi:10.1038/nrg2732

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., … Grenier-Boley, B. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, *45*(12), 1452–1458. doi:10.1038/ng.2802

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

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., … Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862–8. doi:10.1093/nar/gkv1222

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559. doi:10.1186/1471-2105-9-559

Lardenoije, R., Roubroeks, J. A. Y., Pishva, E., Leber, M., Wagner, H., Iatrou, A., … van den Hove, D. L. A. (2019). Alzheimer's disease-associated (hydroxy)methylomic changes in the brain and blood. *Clinical epigenetics*, *11*(1), 164. doi:10.1186/s13148-019-0755-5

Larsen, F., Gundersen, G., Lopez, R., & Prydz, H. (1992). CpG islands as gene markers in the human genome. *Genomics*, *13*(4), 1095–1107. doi:10.1016/0888-7543(92)90024-M

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., … GENEVA Investigators. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, *34*(6), 591–602. doi:10.1002/gepi.20516

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882–883. doi:10.1093/bioinformatics/bts034

Leonenko, G., Shoai, M., Bellou, E., Sims, R., Williams, J., Hardy, J., … Alzheimer's Disease Neuroimaging Initiative. (2019). Genetic risk for alzheimer disease is distinct from genetic risk for amyloid deposition. *Annals of Neurology*, *86*(3), 427–435. doi:10.1002/ana.25530

Lescai, F., Als, T. D., Li, Q., Nyegaard, M., Andorsdottir, G., Biskopstø, M., … Børglum, A. D. (2017). Whole-exome sequencing of individuals from an isolated population implicates rare risk variants in bipolar disorder. *Translational psychiatry*, *7*(2), e1034. doi:10.1038/tp.2017.3

Levine, M. E., Lu, A. T., Bennett, D. A., & Horvath, S. (2015). Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. *Aging*, *7*(12), 1198–1211. doi:10.18632/aging.100864

Levine, M. E., Lu, A. T., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S., … Horvath, S. (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging*, *10*(4), 573–591. doi:10.18632/aging.101414

Lewis, A. C. F., & Green, R. C. (2021). Polygenic risk scores in the clinic: new perspectives needed on familiar ethical issues. *Genome Medicine*, *13*(1), 14. doi:10.1186/s13073-021-00829-7

Li, Q. S., Sun, Y., & Wang, T. (2020). Epigenome-wide association study of Alzheimer's disease replicates 22 differentially methylated positions and 30 differentially methylated regions. *Clinical epigenetics*, *12*(1), 149. doi:10.1186/s13148-020-00944-z

Li, S., He, T., Pawlikowska, I., & Lin, T. (2017). Correcting length-bias in gene set analysis for DNA methylation data. *Statistics and its interface*, *10*(2), 279–289. doi:10.4310/SII.2017.v10.n2.a11

Li, W., Liu, H., Yu, M., Zhang, X., Zhang, Y., Liu, H., … Huang, G. (2016). Folic Acid Alters Methylation Profile of JAK-STAT and Long-Term Depression Signaling Pathways in Alzheimer's Disease Models. *Molecular Neurobiology*, *53*(9), 6548–6556. doi:10.1007/s12035-015-9556-9

Li, X., Song, D., & Leng, S. X. (2015). Link between type 2 diabetes and Alzheimer's disease: from epidemiology to mechanism and treatment. *Clinical Interventions in Aging*, *10*, 549–560. doi:10.2147/CIA.S74042

Li, Yonghong, Grupe, A., Rowland, C., Holmans, P., Segurado, R., Abraham, R., … Williams, J. (2008). Evidence that common variation in NEDD9 is associated with susceptibility to late-onset Alzheimer’s and Parkinson's disease. *Human Molecular Genetics*, *17*(5), 759–767. doi:10.1093/hmg/ddm348

Li, Yuanyuan, & Tollefsbol, T. O. (2011). DNA methylation detection: bisulfite genomic sequencing analysis. *Methods in Molecular Biology*, *791*, 11–21. doi:10.1007/978-1-61779-316-5_2

Li, Yun, Chen, J. A., Sears, R. L., Gao, F., Klein, E. D., Karydas, A., … Coppola, G. (2014). An epigenetic signature in peripheral blood associated with the haplotype on 17q21.31, a risk factor for neurodegenerative tauopathy. *PLoS Genetics*, *10*(3), e1004211. doi:10.1371/journal.pgen.1004211

Liao, H.-K., Hatanaka, F., Araoka, T., Reddy, P., Wu, M.-Z., Sui, Y., … Izpisua Belmonte, J. C. (2017). In Vivo Target Gene Activation via CRISPR/Cas9-Mediated Trans-epigenetic Modulation. *Cell*, *171*(7), 1495–1507.e15. doi:10.1016/j.cell.2017.10.025

Lim, A., Tsuang, D., Kukull, W., Nochlin, D., Leverenz, J., McCormick, W., … Larson, E. B. (1999). Clinico-neuropathological correlation of Alzheimer’s disease in a community-based case series. *Journal of the American Geriatrics Society*, *47*(5), 564–569. doi:10.1111/j.1532-5415.1999.tb02571.x

Lin, A.-P., Abbas, S., Kim, S.-W., Ortega, M., Bouamar, H., Escobedo, Y., … Aguiar, R. C. T. (2015). *D2HGDH* regulates alpha-ketoglutarate levels and dioxygenase function by modulating IDH2. *Nature Communications*, *6*, 7768. doi:10.1038/ncomms8768

Linnevers, C., Smeekens, S. P., & Brömme, D. (1997). Human cathepsin W, a putative cysteine protease predominantly expressed in CD8$^+$ T-lymphocytes. *FEBS Letters*, *405*(3), 253–259. doi:10.1016/S0014-5793(97)00118-X

Liu, D., Wang, Y., Jing, H., Meng, Q., & Yang, J. (2021). Mendelian randomization integrating GWAS and mQTL data identified novel pleiotropic DNA methylation loci for neuropathology of Alzheimer’s disease. *Neurobiology of Aging*, *97*, 18–27. doi:10.1016/j.neurobiolaging.2020.09.019

Liu, G., Zhang, Y., Wang, L., Xu, J., Chen, X., Bao, Y., … Jiang, Q. (2018). Alzheimer’s Disease rs11767557 Variant Regulates *EPHA1* Gene Expression Specifically in Human Whole Blood. *Journal of Alzheimer’s Disease*, *61*(3), 1077–1088. doi:10.3233/JAD-170468

Liu, Yiyuan, Wang, M., Marcora, E. M., Zhang, B., & Goate, A. M. (2019). Promoter DNA hypermethylation - Implications for Alzheimer’s disease. *Neuroscience Letters*, *711*, 134403. doi:10.1016/j.neulet.2019.134403

Liu, Yun, Li, X., Aryee, M. J., Ekström, T. J., Padyukov, L., Klareskog, L., … Feinberg, A. P. (2014). GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. *American Journal of Human Genetics*, *94*(4), 485–495. doi:10.1016/j.ajhg.2014.02.011

Lizen, B., Hutlet, B., Bissen, D., Sauvegarde, D., Hermant, M., Ahn, M.-T., & Gofflot, F. (2017). *HOXA*5 localization in postnatal and adult mouse brain is suggestive of regulatory roles in postmitotic neurons. *The Journal of Comparative Neurology*, *525*(5), 1155–1175. doi:10.1002/cne.24123

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., … Visscher, P. M. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature Communications*, *10*(1), 5086. doi:10.1038/s41467-019-12653-0

Lobo, A., Launer, L. J., Fratiglioni, L., Andersen, K., Di Carlo, A., Breteler, M. M., … Hofman, A. (2000). Prevalence of dementia and major subtypes in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. *Neurology*, *54*(11 Suppl 5), S4–9.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., … L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, *48*(11), 1443–1448. doi:10.1038/ng.3679

Love, S. (2005). Neuropathological investigation of dementia: a guide for neurologists. *Journal of Neurology, Neurosurgery, and Psychiatry*, *76 Suppl 5*, v8–14. doi:10.1136/jnnp.2005.080754

Lovestone, S., Francis, P., Kloszewska, I., Mecocci, P., Simmons, A., Soininen, H., … AddNeuroMed Consortium. (2009). AddNeuroMed--the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Annals of the New York Academy of Sciences*, *1180*, 36–46. doi:10.1111/j.1749-6632.2009.05064.x

Lunnon, K., Sattlecker, M., Furney, S. J., Coppola, G., Simmons, A., Proitsi, P., … dNeuroMed Consortium. (2013). A blood gene expression marker of early Alzheimer's disease. *Journal of Alzheimer's Disease*, *33*(3), 737–753. doi:10.3233/JAD-2012-121363

Lunnon, K., Smith, R., Hannon, E., De Jager, P. L., Srivastava, G., Volta, M., … Mill, J. (2014). Methylomic profiling implicates cortical deregulation of *ANK1* in Alzheimer's disease. *Nature Neuroscience*, *17*(9), 1164–1170. doi:10.1038/nn.3782

Ma, Y., Vardarajan, B. N., CHARGE ADSP FUS, Bennett, D. A., Fornage, M., Seshadri, S., … De Jager, P. L. (2020). Alzheimer's disease GWAS weighted by multi-omics and endophenotypes identifies novel risk loci. *Alzheimer's & Dementia*, *16*(S2). doi:10.1002/alz.043977

MacBean, L. F., Smith, A. R., & Lunnon, K. (2020). Exploring beyond the DNA sequence: A review of epigenomic studies of DNA and histone modifications in dementia. *Current genetic medicine reports*. doi:10.1007/s40142-020-00190-y

Mackenzie, I. R., Rademakers, R., & Neumann, M. (2010). TDP-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia. *Lancet Neurology*, *9*(10), 995–1007. doi:10.1016/S1474-4422(10)70195-2

Madrid, A., Hogan, K. J., Papale, L. A., Clark, L. R., Asthana, S., Johnson, S. C., & Alisch, R. S. (2018). DNA hypomethylation in blood links B3GALT4 and ZADH2 to alzheimer's disease. *Journal of Alzheimer's Disease*, *66*(3), 927–934. doi:10.3233/JAD-180592

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. doi:10.1038/nature08494

Mansell, G., Gorrie-Stone, T. J., Bao, Y., Kumari, M., Schalkwyk, L. S., Mill, J., & Hannon, E. (2019). Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. *BMC Genomics*, *20*(1), 366. doi:10.1186/s12864-019-5761-7

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, *27*(2), e1608. doi:10.1002/mpr.1608

Marioni, R. E., Campbell, A., Hagenaars, S. P., Nagy, R., Amador, C., Hayward, C., … Deary, I. J. (2017). Genetic stratification to identify risk groups for alzheimer's disease. *Journal of Alzheimer's Disease*, *57*(1), 275–283. doi:10.3233/JAD-161070

Marioni, R. E., Harris, S. E., Zhang, Q., McRae, A. F., Hagenaars, S. P., Hill, W. D., … Visscher, P. M. (2018). GWAS on family history of Alzheimer's disease. *Translational psychiatry*, *8*(1), 99. doi:10.1038/s41398-018-0150-6

Marioni, R. E., Shah, S., McRae, A. F., Ritchie, S. J., Muniz-Terrera, G., Harris, S. E., … Deary, I. J. (2015). The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *International Journal of Epidemiology*, *44*(4), 1388–1396. doi:10.1093/ije/dyu277

Martiskainen, H., Helisalmi, S., Viswanathan, J., Kurki, M., Hall, A., Herukka, S.-K., … Hiltunen, M. (2015). Effects of Alzheimer's disease-associated risk loci on cerebrospinal fluid biomarkers and disease progression: a polygenic risk score approach. *Journal of Alzheimer's Disease*, *43*(2), 565–573. doi:10.3233/JAD-140777

Marzi, S. J., Leung, S. K., Ribarska, T., Hannon, E., Smith, A. R., Pishva, E., … Mill, J. (2018). A histone acetylome-wide association study of Alzheimer's disease identifies disease-associated H3K27ac differences in the entorhinal cortex. *Nature Neuroscience*, *21*(11), 1618–1627. doi:10.1038/s41593-018-0253-7

Masliah, E., Dumaop, W., Galasko, D., & Desplats, P. (2013). Distinctive patterns of DNA methylation associated with Parkinson disease: identification of concordant epigenetic changes in brain and peripheral blood leukocytes. *Epigenetics*, *8*(10), 1030–1038. doi:10.4161/epi.25865

Matosin, N., Green, M. J., Newell, K. A., & Fernandez-Enright, F. (2017). Effects of GRASP variation on memory in psychiatrically healthy individuals and cognitive dysfunction in schizophrenics. *Gene Reports*, *6*, 121–127. doi:10.1016/j.genrep.2016.12.010

Mawuenyega, K. G., Sigurdson, W., Ovod, V., Munsell, L., Kasten, T., Morris, J. C., ... & Bateman, R. J. (2010). Decreased clearance of CNS β-amyloid in Alzheimer's disease. *Science*, *330*(6012), 1774-1774.

Maze, I., Covington, H. E., Dietz, D. M., LaPlant, Q., Renthal, W., Russo, S. J., … Nestler, E. J. (2010). Essential role of the histone methyltransferase G9a in cocaine-induced plasticity. *Science*, *327*(5962), 213–216. doi:10.1126/science.1179438

McCartney, D. L., Stevenson, A. J., Walker, R. M., Gibson, J., Morris, S. W., Campbell, A., … Marioni, R. E. (2018). Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease. *Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring*, *10*, 429–437. doi:10.1016/j.dadm.2018.05.006

McCartney, D. L., Walker, R. M., Morris, S. W., McIntosh, A. M., Porteous, D. J., & Evans, K. L. (2016). Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genomics data*, *9*, 22–24. doi:10.1016/j.gdata.2016.05.012

McEwen, L. M., O'Donnell, K. J., McGill, M. G., Edgar, R. D., Jones, M. J., MacIsaac, J. L., … Kobor, M. S. (2019). The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.1820843116

McKinney, B. C., Lin, H., Ding, Y., Lewis, D. A., & Sweet, R. A. (2018). DNA methylation age is not accelerated in brain or blood of subjects with schizophrenia. *Schizophrenia Research*, *196*, 39–44. doi:10.1016/j.schres.2017.09.025

McRae, A. F., Marioni, R. E., Shah, S., Yang, J., Powell, J. E., Harris, S. E., … Montgomery, G. W. (2018). Identification of 55,000 replicated DNA methylation QTL. *Scientific Reports*, *8*(1), 17605. doi:10.1038/s41598-018-35871-w

McRae, A. F., Powell, J. E., Henders, A. K., Bowdler, L., Hemani, G., Shah, S., … Montgomery, G. W. (2014). Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biology*, *15*(5), R73. doi:10.1186/gb-2014-15-5-r73

Mendizabal, I., Berto, S., Usui, N., Toriumi, K., Chatterjee, P., Douglas, C., … Yi, S. V. (2019). Cell type-specific epigenetic links to schizophrenia risk in the brain. *Genome Biology*, *20*(1), 135. doi:10.1186/s13059-019-1747-7

Mendizabal, I., & Yi, S. V. (2016). Whole-genome bisulfite sequencing maps from multiple human tissues reveal novel CpG islands associated with tissue-specific regulation. *Human Molecular Genetics*, *25*(1), 69–82. doi:10.1093/hmg/ddv449

Mi, G., Di, Y., Emerson, S., Cumbie, J. S., & Chang, J. H. (2012). Length bias correction in gene ontology enrichment analysis using logistic regression. *Plos One*, *7*(10), e46128. doi:10.1371/journal.pone.0046128

Mill, J., & Heijmans, B. T. (2013). From promises to practical strategies in epigenetic epidemiology. *Nature Reviews. Genetics*, *14*(8), 585–594. doi:10.1038/nrg3405

Min, J. L., Hemani, G., Hannon, E., Dekkers, K. F., Castillo-Fernandez, J., Luijk, R., … Suderman, M. (2020). Genomic and phenomic insights from an atlas of genetic effects on DNA methylation. *medRxiv*. doi:10.1101/2020.09.01.20180406

Mirra, S. S., Heyman, A., McKeel, D., Sumi, S. M., Crain, B. J., Brownlee, L. M., … Berg, L. (1991). The Consortium to Establish a Registry for Alzheimer’s Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology*, *41*(4), 479–486. doi:10.1212/wnl.41.4.479

Montine, T. J., Phelps, C. H., Beach, T. G., Bigio, E. H., Cairns, N. J., Dickson, D. W., … Alzheimer’s Association. (2012). National Institute on Aging-Alzheimer’s Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. *Acta Neuropathologica*, *123*(1), 1–11. doi:10.1007/s00401-011-0910-3

Moran, S., Arribas, C., & Esteller, M. (2016). Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, *8*(3), 389–399. doi:10.2217/epi.15.114

Morgan, A. R., Touchard, S., O’Hagan, C., Sims, R., Majounie, E., Escott-Price, V., … Morgan, B. P. (2017). The Correlation between Inflammatory Biomarkers and Polygenic Risk Score in Alzheimer’s Disease. *Journal of Alzheimer’s Disease*, *56*(1), 25–36. doi:10.3233/JAD-160889

Mormino, E. C., Sperling, R. A., Holmes, A. J., Buckner, R. L., De Jager, P. L., Smoller, J. W., … Alzheimer’s Disease Neuroimaging Initiative. (2016). Polygenic risk of Alzheimer disease is associated with early- and late-life processes. *Neurology*, *87*(5), 481–488. doi:10.1212/WNL.0000000000002922

Morris, D. W., Pearson, R. D., Cormican, P., Kenny, E. M., O’Dushlaine, C. T., Perreault, L.-P. L., … Wormley, B. (2014). An inherited duplication at the gene p21 Protein-Activated Kinase 7 (PAK7) is a risk factor for psychosis. *Human Molecular Genetics*, *23*(12), 3316–3326. doi:10.1093/hmg/ddu025

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*, *43*(11), 2412–2414. doi:10.1212/wnl.43.11.2412-a

Morris, J. C. (1997). Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *International Psychogeriatrics*, *9 Suppl 1*, 173–6; discussion 177. doi:10.1017/S1041610297004870

Morshed, N., Lee, M., Rodriguez, F. H., Lauffenburger, D. A., Mastroeni, D., & White, F. (2020). Quantitative phosphoproteomics uncovers dysregulated kinase networks in Alzheimer’s disease. *BioRxiv*. doi:10.1101/2020.08.18.255778

Mortimer, J. A., van Duijn, C. M., Chandra, V., Fratiglioni, L., Graves, A. B., Heyman, A., … Rocca, W. A. (1991). Head trauma as a risk factor for Alzheimer’s disease: a collaborative re-analysis of case-control studies. EURODEM Risk Factors Research Group. *International Journal of Epidemiology*, *20 Suppl 2*, S28–35. doi:10.1093/ije/20.supplement_2.s28

Nabais, M., Laws, S., Lin, T., Vallerga, C., Armstrong, N., Wray, N., & McRae, A. (2021). Shared associations across neurodegenerative disorders. *Genome Biology*, *In press*.

Nagy, Z., Esiri, M. M., Hindley, N. J., Joachim, C., Morris, J. H., King, E. M., … Smith, A. D. (1998). Accuracy of clinical operational diagnostic criteria for Alzheimer’s disease in relation to different pathological diagnostic protocols. *Dementia and Geriatric Cognitive Disorders*, *9*(4), 219–226. doi:10.1159/000017050

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

Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., … Xue, A. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurology*, *18*(12), 1091–1102. doi:10.1016/S1474-4422(19)30320-5

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., … Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. doi:10.1111/j.1532-5415.2005.53221.x

Nevin, C., & Carroll, M. (2015). Sperm DNA methylation, infertility and transgenerational epigenetics. *Journal of Human Genetics and Clinical Embryology*, *1*(004), 9-10.

Ng, H.-H., & Adrian, B. (1999). DNA methylation and chromatin modification. *Current Opinion in Genetics & Development*, *9*(2), 158–163. doi:10.1016/S0959-437X(99)80024-0

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics*, *6*(4), e1000895. doi:10.1371/journal.pgen.1000895

Nordengen, K., Kirsebom, B.-E., Henjum, K., Selnes, P., Gísladóttir, B., Wettergreen, M., … Fladby, T. (2019). Glial activation and inflammation along the Alzheimer's disease continuum. *Journal of Neuroinflammation*, *16*(1), 46. doi:10.1186/s12974-019-1399-2

Nussbaum, R. L., & Ellis, C. E. (2003). Alzheimer's disease and Parkinson's disease. *The New England Journal of Medicine*, *348*(14), 1356–1364. doi:10.1056/NEJM2003ra020003

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Ako-Adjei, D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–45. doi:10.1093/nar/gkv1189

Oksuzyan, A., Juel, K., Vaupel, J. W., & Christensen, K. (2008). Men: good health and high mortality. Sex differences in health and aging. *Aging Clinical and Experimental Research*, *20*(2), 91–102. doi:10.1007/BF03324754

Outeiro, T. F., Koss, D. J., Erskine, D., Walker, L., Kurzawa-Akanbi, M., Burn, D., … McKeith, I. (2019). Dementia with Lewy bodies: an update and outlook. *Molecular Neurodegeneration*, *14*(1), 5. doi:10.1186/s13024-019-0306-8

Owens, I. P. F. (2002). Ecology and evolution. Sex differences in mortality rate. *Science*, *297*(5589), 2008–2009. doi:10.1126/science.1076813

Pandey, G., & Ramakrishnan, V. (2020). Invasive and non-invasive therapies for Alzheimer's disease and other amyloidosis. *Biophysical reviews*, *12*(5), 1175–1186. doi:10.1007/s12551-020-00752-y

Panni, T., Mehta, A. J., Schwartz, J. D., Baccarelli, A. A., Just, A. C., Wolf, K., … Peters, A. (2016). Genome-Wide Analysis of DNA Methylation and Fine Particulate Matter Air Pollution in Three Study Populations: KORA F3, KORA F4, and the Normative Aging Study. *Environmental Health Perspectives*, *124*(7), 983–990. doi:10.1289/ehp.1509966

Parisiadou, L., Bethani, I., Michaki, V., Krousti, K., Rapti, G., & Efthimiopoulos, S. (2008). Homer2 and Homer3 interact with amyloid precursor protein and inhibit Abeta production. *Neurobiology of Disease*, *30*(3), 353–364. doi:10.1016/j.nbd.2008.02.004

Patel, T., Brookes, K. J., Turton, J., Chaudhury, S., Guetta-Baranes, T., Guerreiro, R., … 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), 506–521. doi:10.1111/nan.12452

Penke, B., Paragi, G., Gera, J., Berkecz, R., Kovács, Z., Crul, T., & VÍgh, L. (2018). The role of lipids and membranes in the pathogenesis of alzheimer’s disease: A comprehensive view. *Current Alzheimer research*, *15*(13), 1191–1212. doi:10.2174/1567205015666180911151716

Pereira, C. D., Martins, F., Wiltfang, J., da Cruz E Silva, O. A. B., & Rebelo, S. (2018). ABC transporters are key players in alzheimer’s disease. *Journal of Alzheimer’s Disease*, *61*(2), 463–485. doi:10.3233/JAD-170639

Petrovitch, H., White, L. R., Ross, G. W., Steinhorn, S. C., Li, C. Y., Masaki, K. H., … Markesbery, W. R. (2001). Accuracy of clinical criteria for AD in the Honolulu-Asia Aging Study, a population-based study. *Neurology*, *57*(2), 226–234. doi:10.1212/wnl.57.2.226

Philippidou, P., & Dasen, J. S. (2013). Hox genes: choreographers in neural development, architects of circuit organization. *Neuron*, *80*(1), 12–34. doi:10.1016/j.neuron.2013.09.020

Picard, C., Julien, C., Frappier, J., Miron, J., Théroux, L., Dea, D., … Poirier. (2018). Alterations in cholesterol metabolism-related genes in sporadic Alzheimer’s disease. *Neurobiology of Aging*, *66*, 180.e1–180.e9. doi:10.1016/j.neurobiolaging.2018.01.018

Pidsley, R., Viana, J., Hannon, E., Spiers, H., Troakes, C., Al-Saraj, S., … Mill, J. (2014). Methylomic profiling of human brain tissue' ' supports a neurodevelopmental origin for schizophrenia. *Genome Biology*, *15*(10), 483. doi:10.1186/s13059-014-0483-2

Pidsley, R., Y Wong, C. C., Volta, M., Lunnon, K., Mill, J., & Schalkwyk, L. C. (2013). A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, *14*, 293. doi:10.1186/1471-2164-14-293

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., … Clark, S. J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, *17*(1), 208. doi:10.1186/s13059-016-1066-1

Piras, I. S., Mills, G., Llaci, L., Naymik, M., Ramsey, K., Belnap, N., … Schrauwen, I. (2017). Exploring genome-wide DNA methylation patterns in Aicardi syndrome. *Epigenomics*, *9*(11), 1373–1386. doi:10.2217/epi-2017-0060

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, *47*(7), 702–709. doi:10.1038/ng.3285

Pombo, A., & Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. *Nature Reviews. Molecular Cell Biology*, *16*(4), 245–257. doi:10.1038/nrm3965

Popp, J., Meichsner, S., Kölsch, H., Lewczuk, P., Maier, W., Kornhuber, J., … Lütjohann, D. (2013). Cerebral and extracerebral cholesterol metabolism and CSF markers of Alzheimer’s disease. *Biochemical Pharmacology*, *86*(1), 37–42. doi:10.1016/j.bcp.2012.12.007

Port, S. A., Mendes, A., Valkova, C., Spillner, C., Fahrenkrog, B., Kaether, C., & Kehlenbach, R. H. (2016). The Oncogenic Fusion Proteins SET-*NUP214* and Sequestosome-1 (SQSTM1)-*NUP214* Form Dynamic Nuclear Bodies and Differentially Affect Nuclear Protein and Poly(A)+ RNA Export. *The Journal of Biological Chemistry*, *291*(44), 23068–23083. doi:10.1074/jbc.M116.735340

Prendecki, M., Florczak-Wyspianska, J., Kowalska, M., Ilkowski, J., Grzelak, T., Bialas, K., … Dorszewska, J. (2018). Biothiols and oxidative stress markers and polymorphisms of *TOMM40* and *APOC1* genes in Alzheimer's disease patients. *Oncotarget*, *9*(81), 35207–35225. doi:10.18632/oncotarget.26184

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews. Genetics*, *11*(7), 459–463. doi:10.1038/nrg2813

Price, M. E., Cotton, A. M., Lam, L. L., Farré, P., Emberly, E., Brown, C. J., … Kobor, M. S. (2013). Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics & Chromatin*, *6*(1), 4. doi:10.1186/1756-8935-6-4

Prins, B. P., Kuchenbaecker, K. B., Bao, Y., Smart, M., Zabaneh, D., Fatemifar, G., … Zeggini, E. (2017). Genome-wide analysis of health-related biomarkers in the UK Household Longitudinal Study reveals novel associations. *Scientific Reports*, *7*(1), 11008. doi:10.1038/s41598-017-10812-1

Project MinE ALS Sequencing Consortium. (2018). Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *European Journal of Human Genetics*, *26*(10), 1537–1546. doi:10.1038/s41431-018-0177-4

Qazi, T. J., Quan, Z., Mir, A., & Qing, H. (2018). Epigenetics in alzheimer's disease: perspective of DNA methylation. *Molecular Neurobiology*, *55*(2), 1026–1044. doi:10.1007/s12035-016-0357-6

Qi, T., Wu, Y., Zeng, J., Zhang, F., Xue, A., Jiang, L., … Yang, J. (2018). Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nature Communications*, *9*(1), 2282. doi:10.1038/s41467-018-04558-1

Qing-Qing, T., Yu-Chao, C., & Zhi-Ying, W. (2018). The role of *CD2AP* in the Pathogenesis of Alzheimer's Disease. *Aging and disease*. doi:10.14336/AD.2018.1025

Quach, A., Levine, M. E., Tanaka, T., Lu, A. T., Chen, B. H., Ferrucci, L., … Horvath, S. (2017). Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging*, *9*(2), 419–446. doi:10.18632/aging.101168

Rahman, M. F., Wang, J., Patterson, T. A., Saini, U. T., Robinson, B. L., Newport, G. D., … Ali, S. F. (2009). Expression of genes related to oxidative stress in the mouse brain after exposure to silver-25 nanoparticles. *Toxicology Letters*, *187*(1), 15–21. doi:10.1016/j.toxlet.2009.01.020

Rajan, K. B., Wilson, R. S., Weuve, J., Barnes, L. L., & Evans, D. A. (2015). Cognitive impairment 18 years before clinical diagnosis of Alzheimer disease dementia. *Neurology*, *85*(10), 898–904. doi:10.1212/WNL.0000000000001774

Ramos de Matos, M., Ferreira, C., Herukka, S.-K., Soininen, H., Janeiro, A., Santana, I., … Martins, M. (2018). Quantitative genetics validates previous genetic variants and identifies novel genetic players influencing alzheimer's disease cerebrospinal fluid biomarkers. *Journal of Alzheimer's Disease*, *66*(2), 639–652. doi:10.3233/JAD-180512

Ren, X., & Kuan, P. F. (2019). methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. *Bioinformatics*, *35*(11), 1958–1959. doi:10.1093/bioinformatics/bty892

Rice, R. A., Berchtold, N. C., Cotman, C. W., & Green, K. N. (2014). Age-related downregulation of the CaV3.1 T-type calcium channel as a mediator of amyloid beta production. *Neurobiology of Aging*, *35*(5), 1002–1011. doi:10.1016/j.neurobiolaging.2013.10.090

Ridge, P. G., Hoyt, K. B., Boehme, K., Mukherjee, S., Crane, P. K., Haines, J. L., … Alzheimer's Disease Genetics Consortium (ADGC). (2016). Assessment of the genetic variance of late-onset

Alzheimer's disease. *Neurobiology of Aging*, *41*, 200.e13–200.e20. doi:10.1016/j.neurobiolaging.2016.02.024

Ridge, P. G., Mukherjee, S., Crane, P. K., Kauwe, J. S. K., & Alzheimer's Disease Genetics Consortium. (2013). Alzheimer's disease: analyzing the missing heritability. *Plos One*, *8*(11), e79771. doi:10.1371/journal.pone.0079771

Risacher, S. L., Anderson, W. H., Charil, A., Castelluccio, P. F., Shcherbinin, S., Saykin, A. J., … Alzheimer's Disease Neuroimaging Initiative. (2017). Alzheimer disease brain atrophy subtypes are associated with cognition and rate of decline. *Neurology*, *89*(21), 2176–2186. doi:10.1212/WNL.0000000000004670

Roesler, R., Henriques, J. A. P., & Schwartsmann, G. (2006). Gastrin-releasing peptide receptor as a molecular target for psychiatric and neurological disorders. *CNS & Neurological Disorders Drug Targets*, *5*(2), 197–204. doi:10.2174/187152706776359673

Rongve, A., Witoelar, A., Ruiz, A., Athanasiu, L., Abdelnour, C., Clarimon, J., … Andreassen, O. A. (2019). GBA and *APOE* ε4 associate with sporadic dementia with Lewy bodies in European genome wide association study. *Scientific Reports*, *9*(1), 7013. doi:10.1038/s41598-019-43458-2

Rosenthal, S. L., & Kamboh, M. I. (2014). Late-Onset Alzheimer's Disease Genes and the Potentially Implicated Pathways. *Current genetic medicine reports*, *2*, 85–101. doi:10.1007/s40142-014-0034-x

Roses, A. D., Lutz, M. W., Amrine-Madsen, H., Saunders, A. M., Crenshaw, D. G., Sundseth, S. S., … Reiman, E. M. (2010). A *TOMM40* variable-length polymorphism predicts the age of late-onset Alzheimer's disease. *The Pharmacogenomics Journal*, *10*(5), 375–384. doi:10.1038/tpj.2009.69

Ross, C. A., & Poirier, M. A. (2004). Protein aggregation and neurodegenerative disease. *Nature Medicine*, *10 Suppl*, S10–7. doi:10.1038/nm1066

Roubroeks, J. A. Y., Smith, A. R., Smith, R. G., Pishva, E., Ibrahim, Z., Sattlecker, M., … Lunnon, K. (2020). An epigenome-wide association study of Alzheimer's disease blood highlights robust DNA hypermethylation in the HOXB6 gene. *Neurobiology of Aging*, *95*, 26–45. doi:10.1016/j.neurobiolaging.2020.06.023

Rusznák, Z., Henskens, W., Schofield, E., Kim, W. S., & Fu, Y. (2016). Adult neurogenesis and gliogenesis: possible mechanisms for neurorestoration. *Experimental neurobiology*, *25*(3), 103–112. doi:10.5607/en.2016.25.3.103

Ryder, J., Su, Y., & Ni, B. (2004). Akt/GSK3beta serine/threonine kinases: evidence for a signalling pathway mediated by familial Alzheimer's disease mutations. *Cellular Signalling*, *16*(2), 187–200. doi:10.1016/j.cellsig.2003.07.004

Sabbagh, M. N., Malek-Ahmadi, M., Kataria, R., Belden, C. M., Connor, D. J., Pearson, C., … Singh, U. (2010). The Alzheimer's questionnaire: a proof of concept study for a new informant-based dementia assessment. *Journal of Alzheimer's Disease*, *22*(3), 1015–1021. doi:10.3233/JAD-2010-101185

Sabuncu, M. R., Buckner, R. L., Smoller, J. W., Lee, P. H., Fischl, B., Sperling, R. A., & Alzheimer's Disease Neuroimaging Initiative. (2012). The association between a polygenic Alzheimer score and cortical thickness in clinically normal subjects. *Cerebral Cortex*, *22*(11), 2653–2661. doi:10.1093/cercor/bhr348

Sadigh-Eteghad, S., Sabermarouf, B., Majdi, A., Talebi, M., Farhoudi, M., & Mahmoudi, J. (2015). Amyloid-beta: a crucial factor in Alzheimer's disease. *Medical principles and practice : international journal of the Kuwait University, Health Science Centre*, *24*(1), 1–10. doi:10.1159/000369101

Salazar, S. V., Cox, T. O., Lee, S., Brody, A. H., Chyung, A. S., Haas, L. T., & Strittmatter, S. M. (2019). Alzheimer's Disease Risk Factor Pyk2 Mediates Amyloid-β-Induced Synaptic Dysfunction and Loss. *The Journal of Neuroscience*, *39*(4), 758–772. doi:10.1523/JNEUROSCI.1873-18.2018

Samarasekera, N., Al-Shahi Salman, R., Huitinga, I., Klioueva, N., McLean, C. A., Kretzschmar, H., … Ironside, J. W. (2013). Brain banking for neurological disorders. *Lancet Neurology*, *12*(11), 1096–1105. doi:10.1016/S1474-4422(13)70202-3

Sanchez-Mut, J. V., Heyn, H., Vidal, E., Moran, S., Sayols, S., Delgado-Morales, R., … Esteller, M. (2016). Human DNA methylomes of neurodegenerative diseases show common epigenomic patterns. *Translational psychiatry*, *6*, e718. doi:10.1038/tp.2015.214

Sanders, J. L., & Newman, A. B. (2013). Telomere length in epidemiology: a biomarker of aging, age-related disease, both, or neither? *Epidemiologic Reviews*, *35*, 112–131. doi:10.1093/epirev/mxs008

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*(7510), 421–427. doi:10.1038/nature13595

Schübeler, D. (2015). Function and information content of DNA methylation. *Nature*, *517*(7534), 321–326. doi:10.1038/nature14192

Schwartzentruber, J., Cooper, S., Liu, J. Z., Barrio-Hernandez, I., Bello, E., Kumasaka, N., … Bassett, A. (2021). Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nature Genetics*.

Selkoe, D. J., & Hardy, J. (2016). The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Molecular Medicine*, *8*(6), 595–608. doi:10.15252/emmm.201606210

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, *28*(10), 1353–1358. doi:10.1093/bioinformatics/bts163

Shang, Z., Lv, H., Zhang, M., Duan, L., Wang, S., Li, J., … Jiang, Y. (2015). Genome-wide haplotype association study identify *TNFRSF1A*, CASP7, LRP1B, CDH1 and TG genes associated with Alzheimer's disease in Caribbean Hispanic individuals. *Oncotarget*, *6*(40), 42504–42514. doi:10.18632/oncotarget.6391

Shao, Y., Shaw, M., Todd, K., Khrestian, M., D'Aleo, G., Barnard, P. J., … Bekris, L. M. (2018). DNA methylation of *TOMM40-APOE*-APOC2 in Alzheimer's disease. *Journal of Human Genetics*, *63*(4), 459–471. doi:10.1038/s10038-017-0393-8

Shaw, C. A., Li, Y., Wiszniewska, J., Chasse, S., Zaidi, S. N. Y., Jin, W., … Szigeti, K. (2011). Olfactory copy number association with age at onset of Alzheimer disease. *Neurology*, *76*(15), 1302–1309. doi:10.1212/WNL.0b013e3182166df5

Sheffield, L. G., Miskiewicz, H. B., Tannenbaum, L. B., & Mirra, S. S. (2006). Nuclear pore complex proteins in Alzheimer disease. *Journal of Neuropathology and Experimental Neurology*, *65*(1), 45–54. doi:10.1097/01.jnen.0000195939.40410.08

Shen, Y., Xia, Y., Meng, S., Lim, N. K. H., Wang, W., & Huang, F. (2017). SH2B1 is Involved in the Accumulation of Amyloid-β42 in Alzheimer's Disease. *Journal of Alzheimer's Disease*, *55*(2), 835–847. doi:10.3233/JAD-160233

Shireby, G. L., Davies, J. P., Francis, P. T., Burrage, J., Walker, E. M., Neilson, G. W. A., … Mill, J. (2020). Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex. *Brain: A Journal of Neurology*, *143*(12), 3763–3775. doi:10.1093/brain/awaa334

Shoemaker, R., Deng, J., Wang, W., & Zhang, K. (2010). Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Research*, *20*(7), 883–889. doi:10.1101/gr.104695.109

Shulman, J. M., Chen, K., Keenan, B. T., Chibnik, L. B., Fleisher, A., Thiyyagura, P., … Bennett, D. A. (2013). Genetic susceptibility for Alzheimer disease neuritic plaque pathology. *JAMA neurology*, *70*(9), 1150–1157. doi:10.1001/jamaneurol.2013.2815

Sierksma, A., Lu, A., Mancuso, R., Fattorelli, N., Thrupp, N., Salta, E., … Fiers, M. (2020). Novel Alzheimer risk genes determine the microglia response to amyloid-β but not to TAU pathology. *EMBO Molecular Medicine*, *12*(3), e10606. doi:10.15252/emmm.201910606

Sierra, F. (2019). Geroscience and the challenges of aging societies. *Aging Medicine*, *2*(3), 132–134. doi:10.1002/agm2.12082

Silva, M. V. F., Loures, C. de M. G., Alves, L. C. V., de Souza, L. C., Borges, K. B. G., & Carvalho, M. das G. (2019). Alzheimer's disease: risk factors and potentially protective measures. *Journal of Biomedical Science*, *26*(1), 33. doi:10.1186/s12929-019-0524-y

Sims, R., Hill, M., & Williams, J. (2020). The multiplex model of the genetics of Alzheimer's disease. *Nature Neuroscience*, *23*(3), 311–322. doi:10.1038/s41593-020-0599-5

Sims, R., van der Lee, S. J., Naj, A. C., Bellenguez, C., Badarinarayan, N., Jakobsdottir, J., … Bis, J. C. (2017). Rare coding variants in PLCG2, *ABI3*, and *TREM2* implicate microglial-mediated innate immunity in Alzheimer's disease. *Nature Genetics*, *49*(9), 1373–1384. doi:10.1038/ng.3916

Skoog, I., Lernfelt, B., Landahl, S., Palmertz, B., Andreasson, L. A., Nilsson, L., … Svanborg, A. (1996). 15-year longitudinal study of blood pressure and dementia. *The Lancet*, *347*(9009), 1141–1145. doi:10.1016/s0140-6736(96)90608-x

Smith, A. K., Kilaru, V., Kocak, M., Almli, L. M., Mercer, K. B., Ressler, K. J., … Conneely, K. N. (2014). Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics*, *15*, 145. doi:10.1186/1471-2164-15-145

Smith, A. R., Smith, R. G., Burrage, J., Troakes, C., Al-Sarraj, S., Kalaria, R. N., … Lunnon, K. (2019). A cross-brain regions study of *ANK1* DNA methylation in different neurodegenerative diseases. *Neurobiology of Aging*, *74*, 70–76. doi:10.1016/j.neurobiolaging.2018.09.024

Smith, A. R., Smith, R. G., Condliffe, D., Hannon, E., Schalkwyk, L., Mill, J., & Lunnon, K. (2016). Increased DNA methylation near *TREM2* is consistently seen in the superior temporal gyrus in Alzheimer's disease brain. *Neurobiology of Aging*, *47*, 35–40. doi:10.1016/j.neurobiolaging.2016.07.008

Smith, A. R., Smith, R. G., Pishva, E., Hannon, E., Roubroeks, J. A. Y., Burrage, J., … Lunnon, K. (2019). Parallel profiling of DNA methylation and hydroxymethylation highlights neuropathology-associated epigenetic variation in Alzheimer's disease. *Clinical epigenetics*, *11*(1), 52. doi:10.1186/s13148-019-0636-y

Smith, R. G., Hannon, E., De Jager, P. L., Chibnik, L., Lott, S. J., Condliffe, D., … Lunnon, K. (2018). Elevated DNA methylation across a 48-kb region spanning the *HOXA* gene cluster is associated with Alzheimer's disease neuropathology. *Alzheimer's & Dementia*, *14*(12), 1580–1588. doi:10.1016/j.jalz.2018.01.017

Smith, R. G., Pishva, E., Shireby, G., Smith, A. R., Roubroeks, J. A. Y., Hannon, E., … Lunnon, K. (2021). A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex. *Nature Communications*, *12*(1), 3517. https://doi.org/10.1038/s41467-021-23243-4

Smolders, J., Remmerswaal, E. B. M., Schuurman, K. G., Melief, J., van Eden, C. G., van Lier, R. A. W., … Hamann, J. (2013). Characteristics of differentiated CD8(+) and CD4 (+) T cells present in the human brain. *Acta Neuropathologica*, *126*(4), 525–535. doi:10.1007/s00401-013-1155-0

Soltysik-Espanola, M., Rogers, R. A., Jiang, S., Kim, T. A., Gaedigk, R., White, R. A., … Avraham, S. (1999). Characterization of Mayven, a novel actin-binding protein predominantly expressed in brain. *Molecular Biology of the Cell*, *10*(7), 2361–2375. doi:10.1091/mbc.10.7.2361

Sosnoff, J. J., & Newell, K. M. (2006). Are age-related increases in force variability due to decrements in strength? *Experimental Brain Research*, *174*(1), 86–94. doi:10.1007/s00221-006-0422-x

Staessen, J. A., Richart, T., & Birkenhäger, W. H. (2007). Less atherosclerosis and lower blood pressure for a meaningful life perspective with more brain. *Hypertension*, *49*(3), 389–400. doi:10.1161/01.HYP.0000258151.00728.d8

Steeland, S., Gorlé, N., Vandendriessche, C., Balusu, S., Brkic, M., Van Cauwenberghe, C., … Vandenbroucke, R. E. (2018). Counteracting the effects of TNF receptor-1 has therapeutic potential in Alzheimer’s disease. *EMBO Molecular Medicine*, *10*(4). doi:10.15252/emmm.201708300

Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P. H., Ingason, A., Steinberg, S., … Buizer-Voskamp, J. E. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, *455*(7210), 232–236. doi:10.1038/nature07229

Stirzaker, C., Taberlay, P. C., Statham, A. L., & Clark, S. J. (2014). Mining cancer methylomes: prospects and challenges. *Trends in Genetics*, *30*(2), 75–84. doi:10.1016/j.tig.2013.11.004

Strang, K. H., Golde, T. E., & Giasson, B. I. (2019). MAPT mutations, tauopathy, and mechanisms of neurodegeneration. *Laboratory Investigation*, *99*(7), 912–928. doi:10.1038/s41374-019-0197-x

Strichman-Almashanu, L. Z., Lee, R. S., Onyango, P. O., Perlman, E., Flam, F., Frieman, M. B., & Feinberg, A. P. (2002). A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome Research*, *12*(4), 543–554. doi:10.1101/gr.224102

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., & Roses, A. D. (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(5), 1977–1981. doi:10.1073/pnas.90.5.1977

Suderman, M., Staley, J. R., French, R., Arathimos, R., Simpkin, A., & Tilling, K. (2018). dmrff: identifying differentially methylated regions efficiently with power and control. *BioRxiv*. doi:10.1101/508556

Sugden, K., Hannon, E. J., Arseneault, L., Belsky, D. W., Broadbent, J. M., Corcoran, D. L., … Caspi, A. (2019). Establishing a generalized polyepigenetic biomarker for tobacco smoking. *Translational psychiatry*, *9*(1), 92. doi:10.1038/s41398-019-0430-9

Sullivan, P. F., Daly, M. J., & O’Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews. Genetics*, *13*(8), 537–551. doi:10.1038/nrg3240

Sun, L.-L., Yang, S.-L., Sun, H., Li, W.-D., & Duan, S.-R. (2019). Molecular differences in Alzheimer’s disease between male and female patients determined by integrative network analysis. *Journal of Cellular and Molecular Medicine*, *23*(1), 47–58. doi:10.1111/jcmm.13852

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*, 699. doi:10.3389/fnins.2018.00699

Tate, P. H., & Bird, A. P. (1993). Effects of DNA methylation on DNA-binding proteins and gene expression. *Current Opinion in Genetics & Development*, *3*(2), 226–231. doi:10.1016/0959-437X(93)90027-M

Thal, D. R., Rüb, U., Orantes, M., & Braak, H. (2002). Phases of A beta-deposition in the human brain and its relevance for the development of AD. *Neurology*, *58*(12), 1791–1800. doi:10.1212/wnl.58.12.1791

Thambisetty, M., Beason-Held, L. L., An, Y., Kraut, M., Nalls, M., Hernandez, D. G., … Resnick, S. M. (2013). Alzheimer risk variant *CLU* and brain function during aging. *Biological Psychiatry*, *73*(5), 399–405. doi:10.1016/j.biopsych.2012.05.026

Thomas, D. X., Bajaj, S., McRae-McKee, K., Hadjichrysanthou, C., Anderson, R. M., & Collinge, J. (2020). Association of TDP-43 proteinopathy, cerebral amyloid angiopathy, and Lewy bodies with cognitive impairment in individuals with or without Alzheimer's disease neuropathology. *Scientific Reports*, *10*(1), 14579. doi:10.1038/s41598-020-71305-2

Tikhmyanova, N., Little, J. L., & Golemis, E. A. (2010). CAS proteins in normal and pathological cell growth control. *Cellular and Molecular Life Sciences*, *67*(7), 1025–1048. doi:10.1007/s00018-009-0213-1

Unternaehrer, E., Luers, P., Mill, J., Dempster, E., Meyer, A. H., Staehli, S., … Meinlschmidt, G. (2012). Dynamic changes in DNA methylation of stress-associated genes (OXTR, *BDNF*) after acute psychosocial stress. *Translational psychiatry*, *2*, e150. doi:10.1038/tp.2012.77

Vallerga, C. L., Zhang, F., Fowdar, J., McRae, A. F., Qi, T., Nabais, M. F., … Gratten, J. (2020). Analysis of DNA methylation associates the cystine-glutamate antiporter SLC7A11 with risk of Parkinson's disease. *Nature Communications*, *11*(1), 1238. doi:10.1038/s41467-020-15065-7

Van Deveire, K. N., Scranton, S. K., Kostek, M. A., Angelopoulos, T. J., Clarkson, P. M., Gordon, P. M., … Pescatello, L. S. (2012). Variants of the ankyrin repeat domain 6 gene (ANKRD6) and muscle and physical activity phenotypes among European-derived American adults. *Journal of Strength and Conditioning Research*, *26*(7), 1740–1748. doi:10.1519/JSC.0b013e31825c2bef

van der Lee, S. J., Wolters, F. J., Ikram, M. K., Hofman, A., Ikram, M. A., Amin, N., & van Duijn, C. M. (2018). The effect of *APOE* and other common genetic variants on the onset of Alzheimer's disease and dementia: a community-based cohort study. *Lancet Neurology*, *17*(5), 434–444. doi:10.1016/S1474-4422(18)30053-X

van Dongen, J., Nivard, M. G., Willemsen, G., Hottenga, J.-J., Helmer, Q., Dolan, C. V., … Boomsma, D. I. (2016). Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature Communications*, *7*, 11115. doi:10.1038/ncomms11115

van Iterson, M., van Zwet, E. W., BIOS Consortium, & Heijmans, B. T. (2017). Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology*, *18*(1), 19. doi:10.1186/s13059-016-1131-9

van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., … Robinson, M. R. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics*, *48*(9), 1043–1048. doi:10.1038/ng.3622

van Rheenen, W., van der Spek, R. A. A., Bakker, M. K., van Vugt, J. J. F. A., Hop, P. J., Zwamborn, R. A. J., … Veldink, J. H. (2021). 1 Common and rare variant association analyses in Amyotrophic Lateral Sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *medRxiv*, *in preparation*.

Vasanthakumar, A., Davis, J. W., Idler, K., Waring, J. F., Asque, E., Riley-Gillis, B., … Alzheimer's Disease Neuroimaging Initiative (ADNI). (2020). Harnessing peripheral DNA methylation differences in the Alzheimer's Disease Neuroimaging Initiative (ADNI) to reveal novel biomarkers of disease. *Clinical epigenetics*, *12*(1), 84. doi:10.1186/s13148-020-00864-y

Veerappan, C. S., Sleiman, S., & Coppola, G. (2013). Epigenetics of Alzheimer's disease and frontotemporal dementia. *Neurotherapeutics*, *10*(4), 709–721. doi:10.1007/s13311-013-0219-0

Viana, J., Hannon, E., Dempster, E., Pidsley, R., Macdonald, R., Knox, O., … Mill, J. (2017). Schizophrenia-associated methylomic variation: molecular signatures of disease and polygenic risk burden across multiple brain regions. *Human Molecular Genetics*, *26*(1), 210–225. doi:10.1093/hmg/ddw373

Villegas-Llerena, C., Phillips, A., Garcia-Reitboeck, P., Hardy, J., & Pocock, J. M. (2016). Microglial genes regulating neuroinflammation in the progression of Alzheimer's disease. *Current Opinion in Neurobiology*, *36*, 74–81. doi:10.1016/j.conb.2015.10.004

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *American Journal of Human Genetics*, *101*(1), 5–22. doi:10.1016/j.ajhg.2017.06.005

Viuff, A.-C. F., Pedersen, L. H., Kyng, K., Staunstrup, N. H., Børglum, A., & Henriksen, T. B. (2016). Antidepressant medication during pregnancy and epigenetic changes in umbilical cord blood: a systematic review. *Clinical epigenetics*, *8*(1), 94. doi:10.1186/s13148-016-0262-x

Voisin, S., Harvey, N. R., Haupt, L. M., Griffiths, L. R., Ashton, K. J., Coffey, V. G., … Eynon, N. (2020). An epigenetic clock for human skeletal muscle. *Journal of cachexia, sarcopenia and muscle*. doi:10.1002/jcsm.12556

Waddington, C. H. (1957). *The strategy of the genes*. Routledge. doi:10.4324/9781315765471

Wagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., & Blanchette, M. (2014). The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology*, *15*(2), R37. doi:10.1186/gb-2014-15-2-r37

Wald, A. (1940). The Fitting of Straight Lines if Both Variables are Subject to Error. In *The Annals of Mathematical Statistics* (Vol. vol. 11, pp. 284–300).

Wallace, C. (2013). Statistical testing of shared genetic control for potentially related traits. *Genetic Epidemiology*, *37*(8), 802–813. doi:10.1002/gepi.21765

Wang, H. X., Wahlin, A., Basun, H., Fastbom, J., Winblad, B., & Fratiglioni, L. (2001). Vitamin B(12) and folate in relation to the development of Alzheimer's disease. *Neurology*, *56*(9), 1188–1194. doi:10.1212/wnl.56.9.1188

Wang, W. Y. S., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews. Genetics*, *6*(2), 109–118. doi:10.1038/nrg1522

Wang, Y., & Wang, Z. (2020). Identification of dysregulated genes and pathways of different brain regions in Alzheimer's disease. *The International journal of neuroscience*, *130*(11), 1082–1094. doi:10.1080/00207454.2020.1720677

Wanker, E. E. (2002). Hip1 and Hippi participate in a novel cell death-signaling pathway. *Developmental Cell*, *2*(2), 126–128. doi:10.1016/s1534-5807(02)00121-1

Watson, C. T., Roussos, P., Garg, P., Ho, D. J., Azam, N., Katsel, P. L., … Sharp, A. J. (2016). Genome-wide DNA methylation profiling in the superior temporal gyrus reveals epigenetic signatures associated with Alzheimer's disease. *Genome Medicine*, *8*(1), 5. doi:10.1186/s13073-015-0258-8

Weingarten, C. P., Sundman, M. H., Hickey, P., & Chen, N. (2015). Neuroimaging of Parkinson's disease: Expanding views. *Neuroscience and Biobehavioral Reviews*, *59*, 16–52. doi:10.1016/j.neubiorev.2015.09.007

Weinhold, L., Wahl, S., Pechlivanis, S., Hoffmann, P., & Schmid, M. (2016). A statistical model for the analysis of beta values in DNA methylation studies. *BMC Bioinformatics*, *17*(1), 480. doi:10.1186/s12859-016-1347-4

Westra, H.-J., Arends, D., Esko, T., Peters, M. J., Schurmann, C., Schramm, K., … Andiappan, A. K. (2015). Cell Specific eQTL Analysis without Sorting Cells. *PLoS Genetics*, *11*(5), e1005223. doi:10.1371/journal.pgen.1005223

Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics*, *76*(5), 887–893. doi:10.1086/429864

Wightman, D. P., Jansen, I. E., Savage, J. E., Shadrin, A. A., Bahrami, S., Rongve, A., … Martinsen, A. E. (2020). Largest GWAS (N=1,126,563) of alzheimer's disease implicates microglia and immune cells. *medRxiv*. doi:10.1101/2020.11.20.20235275

Wilson, A. C., Dugger, B. N., Dickson, D. W., & Wang, D.-S. (2011). TDP-43 in aging and Alzheimer's disease - a review. *International journal of clinical and experimental pathology*, *4*(2), 147–155.

Wimo, A., Guerchet, M., Ali, G.-C., Wu, Y.-T., Prina, A. M., Winblad, B., … Prince, M. (2017). The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's & Dementia*, *13*(1), 1–7. doi:10.1016/j.jalz.2016.07.150

Wingo, T. S., Lah, J. J., Levey, A. I., & Cutler, D. J. (2012). Autosomal recessive causes likely in early-onset Alzheimer disease. *Archives of Neurology*, *69*(1), 59–64. doi:10.1001/archneurol.2011.221

Wong, C. C. Y., Smith, R. G., Hannon, E., Ramaswami, G., Parikshak, N. N., Assary, E., … Mill, J. (2019). Genome-wide DNA methylation profiling identifies convergent molecular signatures associated with idiopathic and syndromic autism in post-mortem human brain tissue. *Human Molecular Genetics*, *28*(13), 2201–2211. doi:10.1093/hmg/ddz052

World Health Organization. (2020). *International Statistical Classification of Diseases and Related Health Problems (ICD-11)* (11th ed.). WHO.

Wu, X., Zhao, J., Ruan, Y., Sun, L., Xu, C., & Jiang, H. (2018). Sialyltransferase ST3GAL1 promotes cell migration, invasion, and TGF-β1-induced EMT and confers paclitaxel resistance in ovarian cancer. *Cell death & disease*, *9*(11), 1102. doi:10.1038/s41419-018-1101-0

Wu, Y., Zeng, J., Zhang, F., Zhu, Z., Qi, T., Zheng, Z., … Yang, J. (2018). Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nature Communications*, *9*(1), 918. doi:10.1038/s41467-018-03371-0

Yang, Jiajia, Yao, Y., Wang, L., Yang, C., Wang, F., Guo, J., … Ming, D. (2017). Gastrin-releasing peptide facilitates glutamatergic transmission in the hippocampus and effectively prevents vascular dementia induced cognitive and synaptic plasticity deficits. *Experimental Neurology*, *287*(Pt 1), 75–83. doi:10.1016/j.expneurol.2016.08.008

Yang, Jian, Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*(1), 76–82. doi:10.1016/j.ajhg.2010.11.011

Yanpallewar, S. U., Barrick, C. A., Palko, M. E., Fulgenzi, G., & Tessarollo, L. (2012). Tamalin is a critical mediator of electroconvulsive shock-induced adult neuroplasticity. *The Journal of Neuroscience*, *32*(7), 2252–2262. doi:10.1523/JNEUROSCI.5493-11.2012

Yazdani, A., Yazdani, A., Méndez Giráldez, R., Aguilar, D., & Sartore, L. (2019). A Multi-Trait Approach Identified Genetic Variants Including a Rare Mutation in *RGS3* with Impact on Abnormalities of Cardiac Structure/Function. *Scientific Reports*, *9*(1), 5845. doi:10.1038/s41598-019-41362-3

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, *11*(2), R14. doi:10.1186/gb-2010-11-2-r14

Yu, L., Chibnik, L. B., Srivastava, G. P., Pochet, N., Yang, J., Xu, J., … Bennett, D. A. (2015). Association of Brain DNA methylation in *SORL1*, *ABCA7*, *HLA-DRB5*, *SLC24A4*, and *BIN1* with pathological diagnosis of Alzheimer disease. *JAMA neurology*, *72*(1), 15–24. doi:10.1001/jamaneurol.2014.3049

Zhang, L., Guo, X. Q., Chu, J. F., Zhang, X., Yan, Z. R., & Li, Y. Z. (2015). Potential hippocampal genes and pathways involved in Alzheimer's disease: a bioinformatic analysis. *Genetics and Molecular Research*, *14*(2), 7218–7232. doi:10.4238/2015.June.29.15

Zhang, Q., Sidorenko, J., Couvy-Duchesne, B., Marioni, R. E., Wright, M. J., Goate, A. M., … Visscher, P. M. (2020). Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. *Nature Communications*, *11*(1), 4799. doi:10.1038/s41467-020-18534-1

Zhang, Q., Vallerga, C. L., Walker, R. M., Lin, T., Henders, A. K., Montgomery, G. W., … Visscher, P. M. (2019). Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Medicine*, *11*(1), 54. doi:10.1186/s13073-019-0667-1

Zhao, T., Hu, Y., Zang, T., & Wang, Y. (2019). Integrate GWAS, eQTL, and mQTL Data to Identify Alzheimer's Disease-Related Genes. *Frontiers in genetics*, *10*, 1021. doi:10.3389/fgene.2019.01021

Zheng, X., Demirci, F. Y., Barmada, M. M., Richardson, G. A., Lopez, O. L., Sweet, R. A., … Feingold, E. (2015). Genome-wide copy-number variation study of psychosis in Alzheimer's disease. *Translational psychiatry*, *5*, e574. doi:10.1038/tp.2015.64

Zhou, Z.-D., Saw, W.-T., & Tan, E.-K. (2017). Mitochondrial CHCHD-Containing Proteins: Physiologic Functions and Link with Neurodegenerative Diseases. *Molecular Neurobiology*, *54*(7), 5534–5546. doi:10.1007/s12035-016-0099-5

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., … Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487. doi:10.1038/ng.3538

Ziller, M. J., Hansen, K. D., Meissner, A., & Aryee, M. J. (2015). Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nature Methods*, *12*(3), 230–2, 1 p following 232. doi:10.1038/nmeth.3152

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.